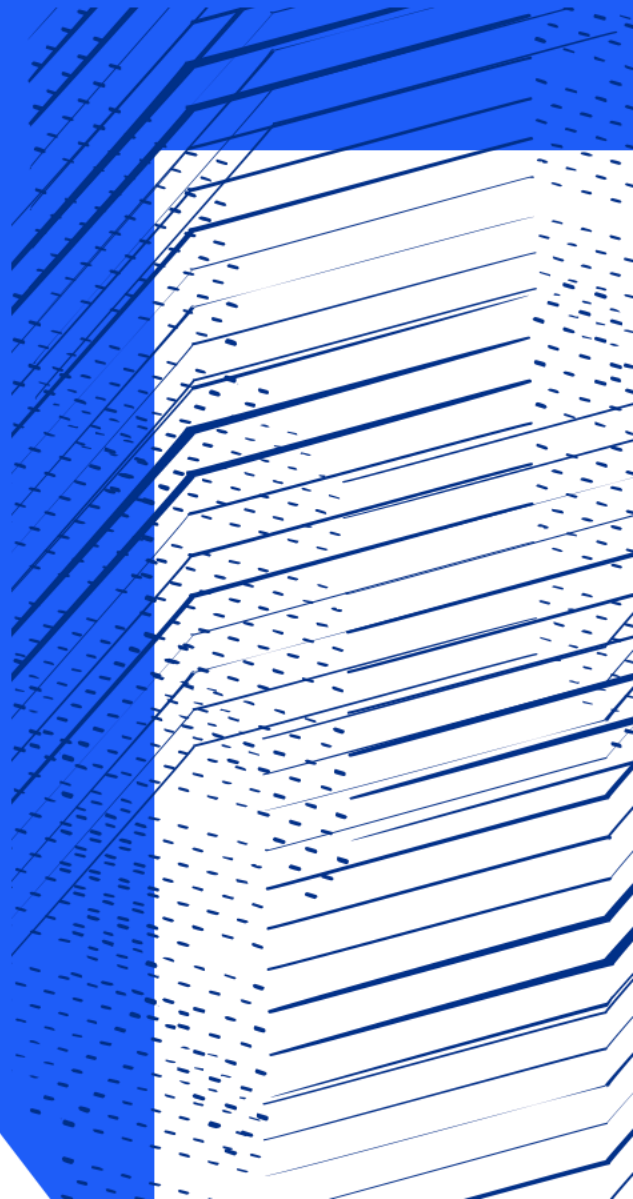
Science and Technology Facilities Council

Tier-1 Update

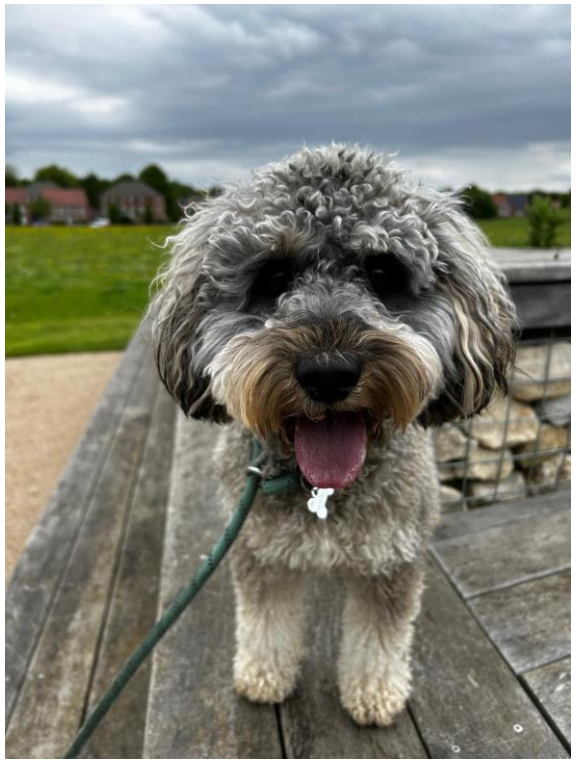
Alastair Dewhurst



Agenda

- Financial Status
- Platforms and Services
- Tier-1 from 2018 – 2025
- Plans for this FY
- Tier-2 helping Tier-1 meet pledge in FY26/27

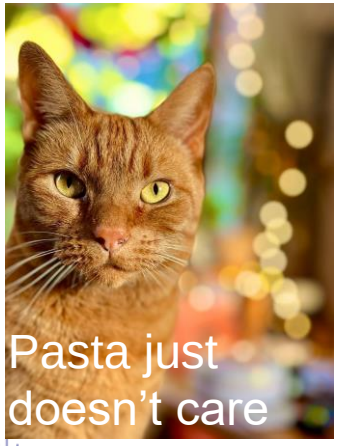
Despite STFC cuts Buddy is living his best life!



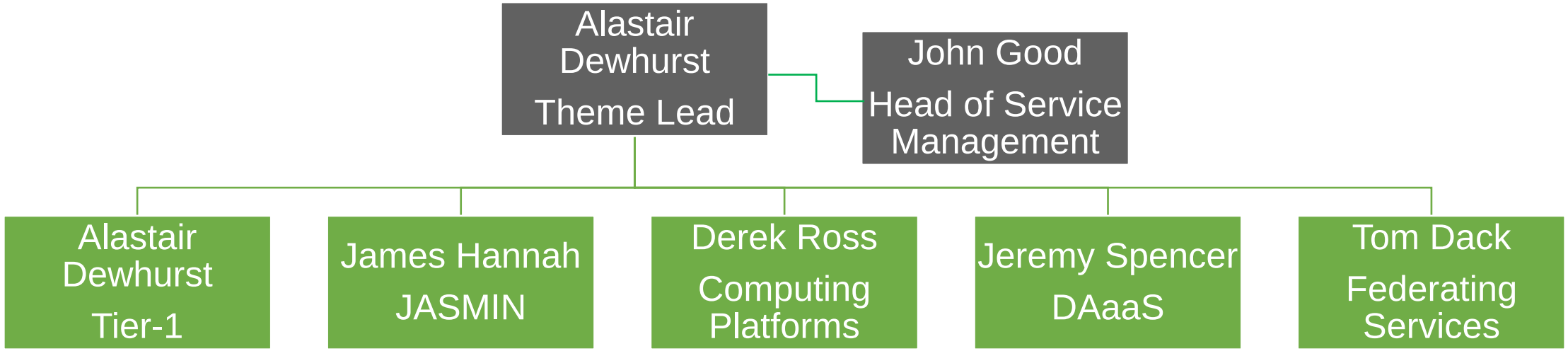
Introduction

- There is an ongoing financial crisis in STFC.
- Large impact on FY24/25 procurements.
  - Disk and Tape only just delivered, late to meet this year's pledge.
- Tier-1 expects no Capital this year.
  - We will not be able to deliver our pledge
  - Must prepare for opportunistic capital just in case.
- Swift-HEP funding ending.
- Recruitment freeze means that we have been unable to recruit replacement staff.
  - Rob Johnson and Oliver (1FTE) from Fabric team
  - Jose Caballero from Grid Services
  - Katy Ellis is reducing to 50% on GridPP
  - Alastair Dewhurst becoming Platforms and Services Theme lead

Toastie thinks everything will be OK



Platforms and Services



- Platforms and Services provides cutting-edge scientific computing platforms to enable world leading research.
- The Theme will provide leadership across SCD on IT Service Management.
- Theme is split into groups that focus on the community they support.

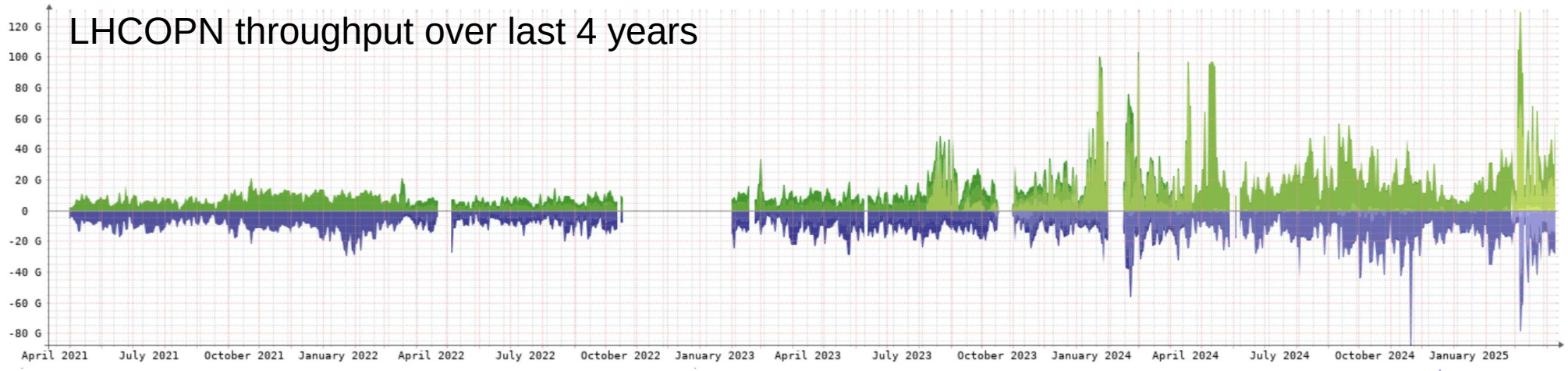
Tier-1 Group Leader

Tier-1 Group Lead:

- This person would be in SCD and report to me (in my new role). They would:
  - Line Manage ~8 people in the Tier-1 Group.
  - They would be responsible for the operational delivery of the Tier-1 MoU pledges (Tape, Disk, CPU).
  - Shared responsibility for development and projects at the Tier-1.
  - They would take a leading role in security operations / policy.
- In my new role I will:
  - Pursue all available options for replacing staff who have left and keeping existing staff.
  - Pursue all opportunities for additional funding (both Capital and Resource).
  - Look for potential efficiency savings resulting from SC25 changes (e.g. more common services, joint procurements).

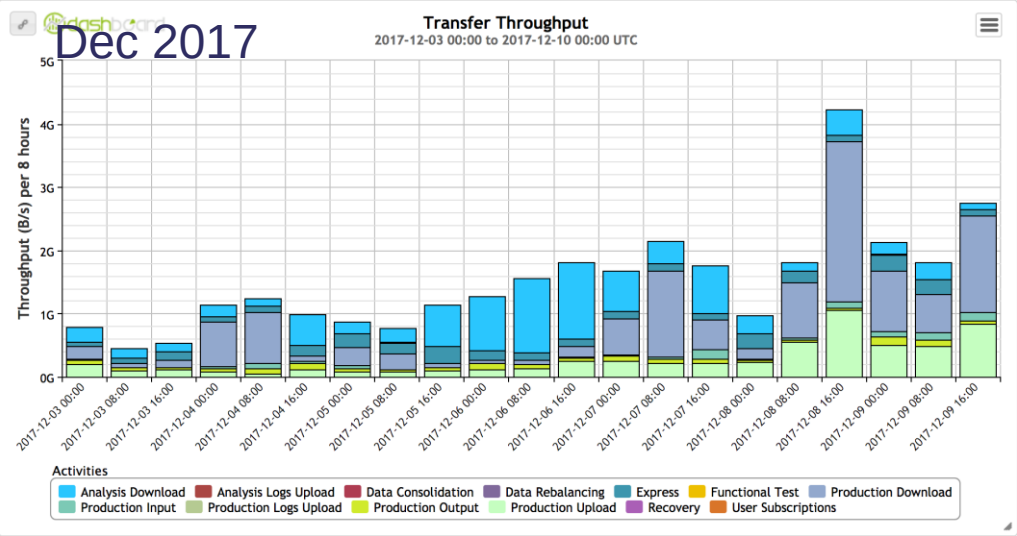
Network 2018 - 2025

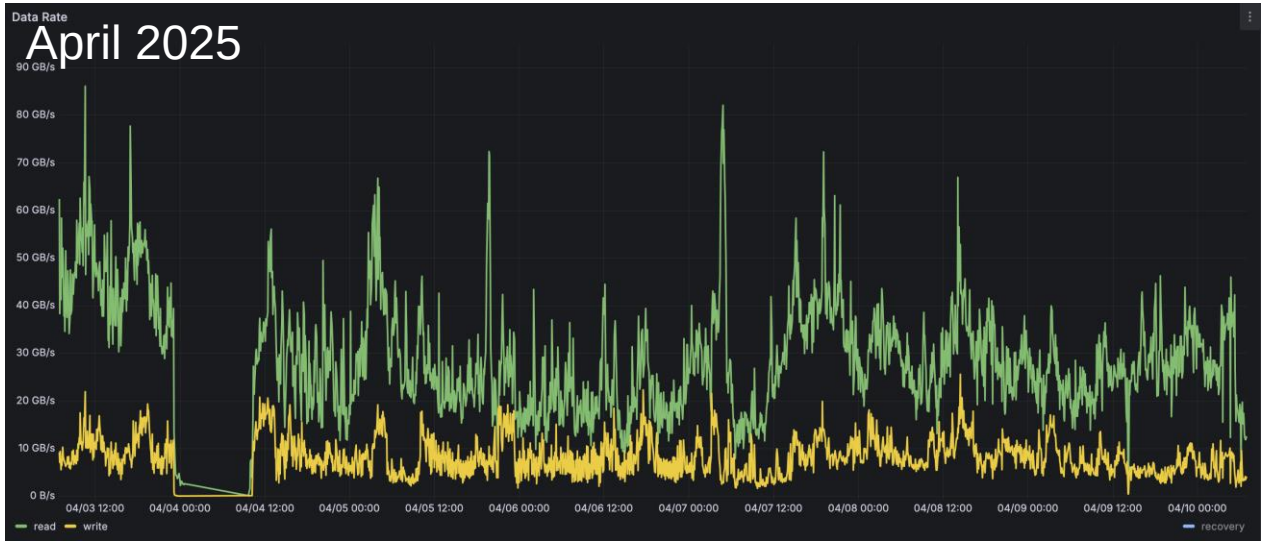
- One of my first decisions as Tier-1 Manager was to upgrade the LHCOPN link from 20Gb/s to 30Gb/s.
  - Increased to 100Gb/s in July 2020
  - Increased to 200Gb/s in January 2023
- Network has been completely re-designed.
- Joined LHCONE (properly) in March 2023.



Echo 2018 - 2025

- In 2018 only ATLAS had been (partially) migrated to Echo.
  - Only GridFTP plugin working
- Capacity has grown:
  - 2018: 193 storage nodes providing 26PB of allocatable capacity.
  - 2025: 332 storage nodes providing 98PB of allocatable capacity.
- External Gateways have grown from 3 to 20+
  - Added in XRootD and Webdav support
  - Redirectors for load balancing, WN gateways and many other optimisations added.
- Throughput has increased factor ~10



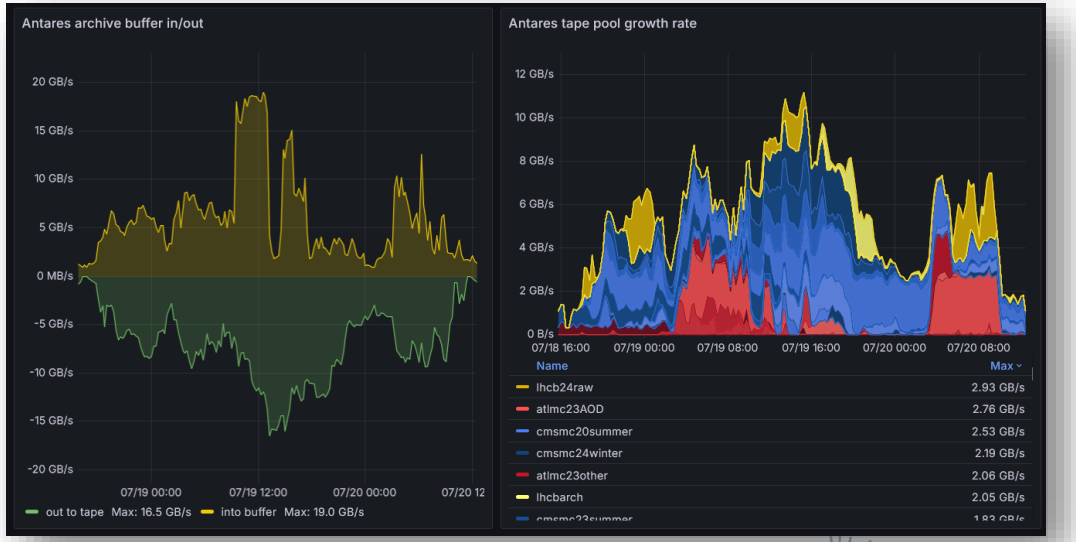


Tape 2018 - 2025

- In 2018 we had Castor and the Oracle Tape Library and media.
- Decided to migrate to CTA in 2019.
  - Antares entered production on 4th March 2022

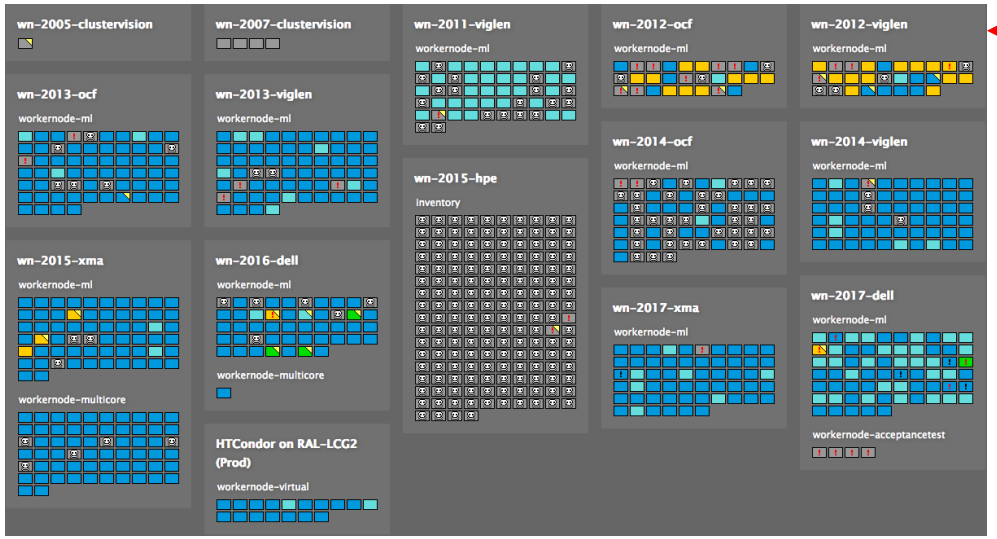
Purchased new Tape Robot in February 2020

  - Installation complete just before Covid!
  - Finished migration from old Robot in May 2021
- Capacity has grown from 43PB in 2018 to 256PB today.
  - Throughput has only ~doubled.

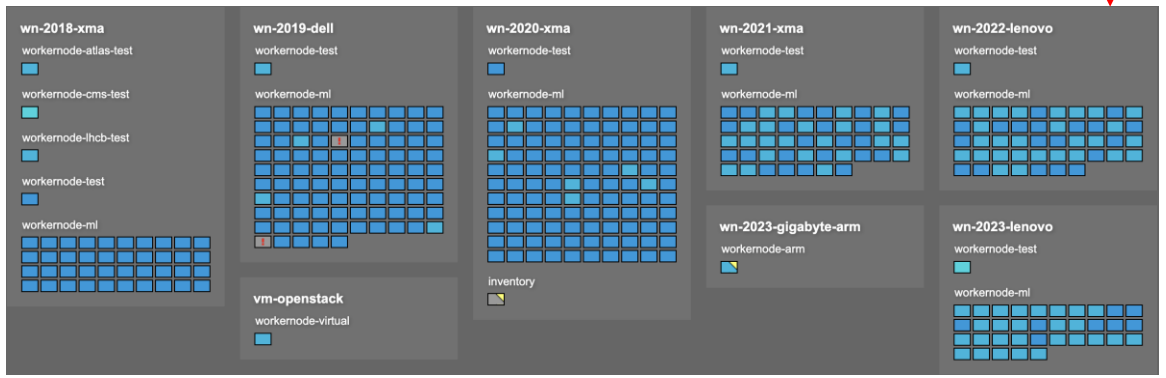


Batch Farm 2018 - 2025

- Batch farm has remained on HTCondor + ARC CE since well before 2018.
  - Many version upgrades.
- Hardware is evolved rapidly
  - We have purchased AMD since 2019 and core counts have gone from 64 - 384
  - 2018 – HDD, 2019 – SSD, 2021 – Separate OS SSD, 2022 – NVMe(s)



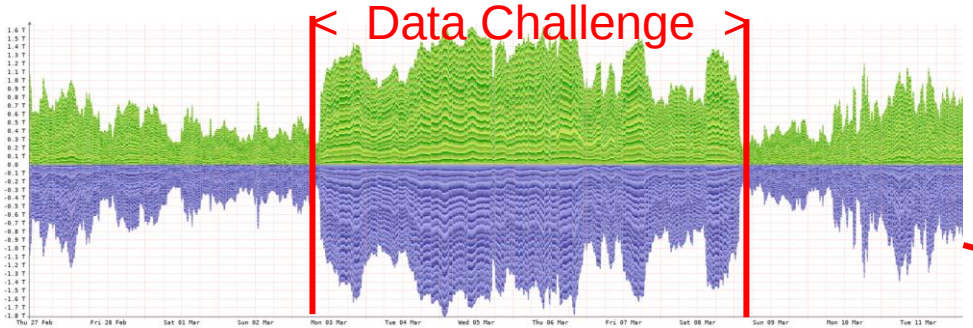
In 2018 we had 824 WN providing 29,664 slots.
In 2025 we have 384 WN providing 67,840 slots.
WN out of production have dropped significantly!



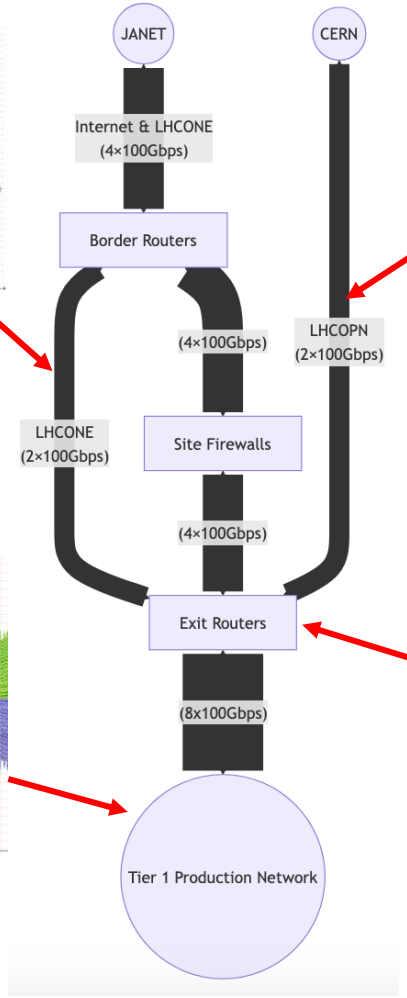
Tier-1 Data Challenge 2025

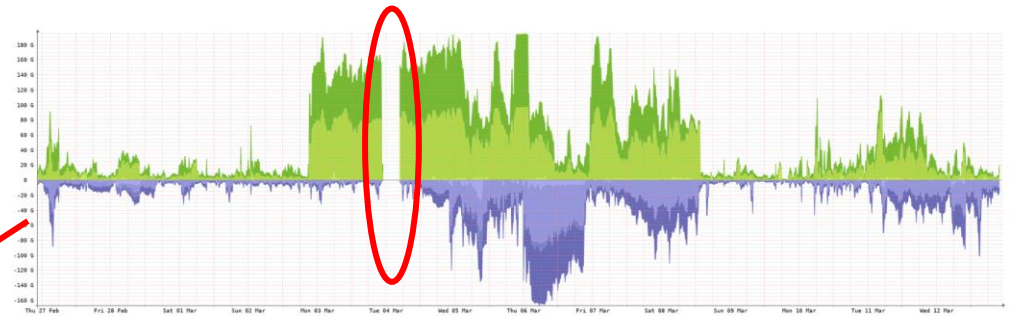
The Tier-1 provides computing resources for the LHC experiments. From the 3rd to 7th March we ran a data challenge (in addition to production work) to test our external network links and disk storage (Echo).

The LHCONE is a VPN which allows us to bypass the site firewall when transferring data to other sites also on the LHCONE.

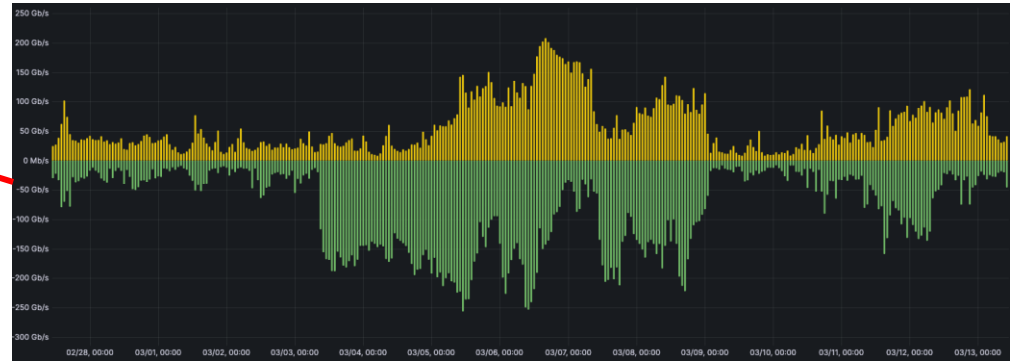


The internal Tier-1 network sustained over 1Tb/s during the challenge.





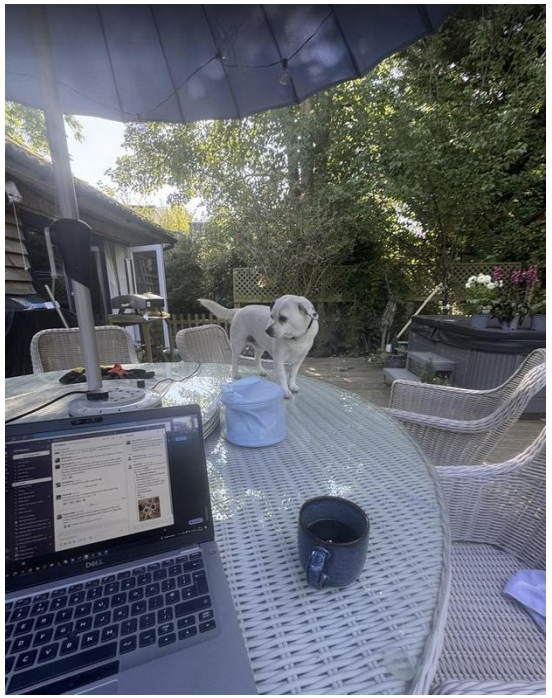
The LHCOPN is a dedicated 2 x 100Gb/s link to CERN. We were able to saturate this link!



The total peak throughput from external sources was 256Gb/s ingress and 208Gb/s egress.

Plan for the year

- Deployment of FY24/25 procurements into production:
  - Disk
  - Tape
- Development opportunities
  - Opportunistic CPU resources
  - XRootD throughput improvements
  - SSD developments
  - New data centre room



When stressed Pico likes to touch Grass

Disk procurement

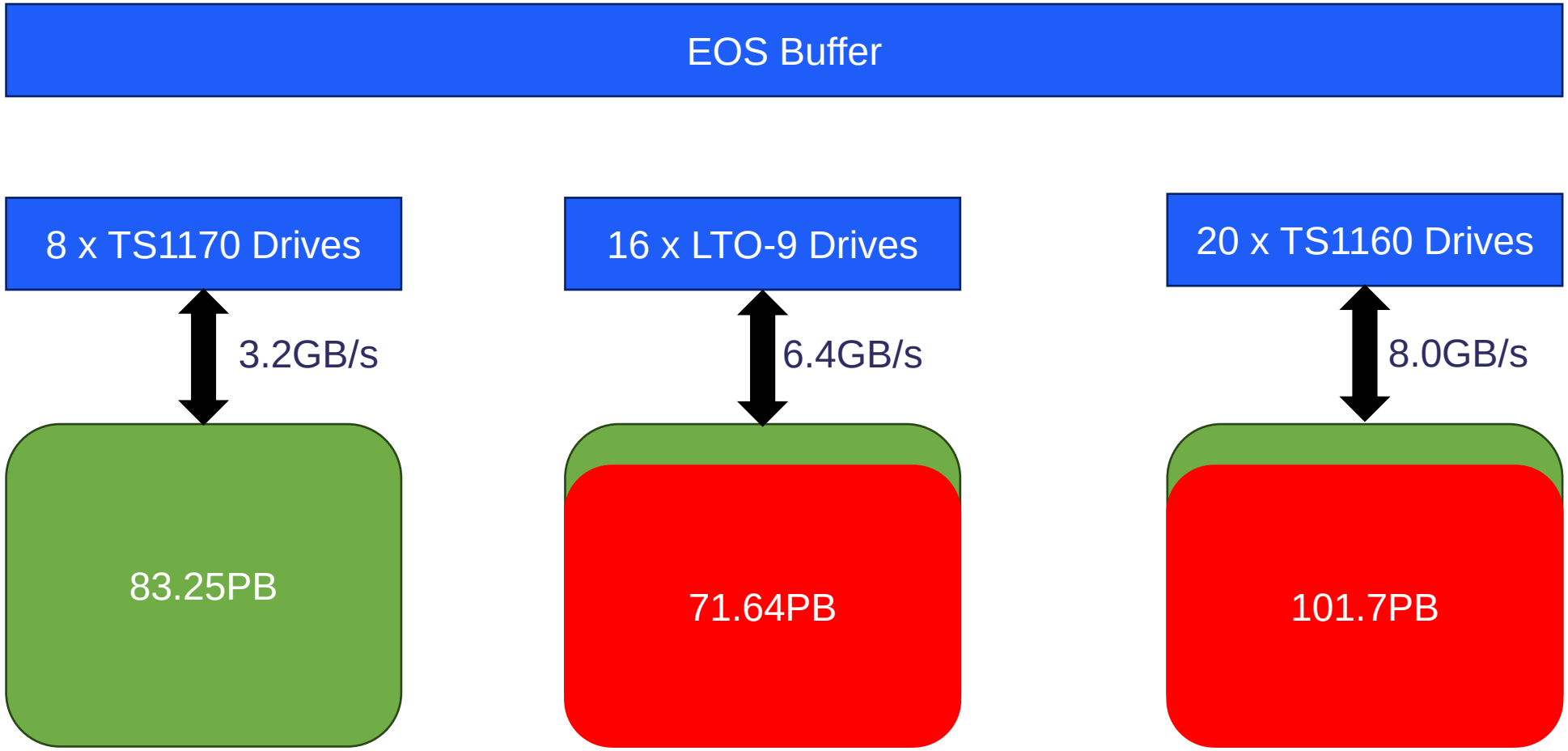
- 43PB raw capacity upgrade to Echo, changes from last year include:
  - Single socket CPU on storage nodes.
  - High Frequency CPU on Monitor
  - 100GB NIC on Gateways

	Storage	Ceph Monitor	XRootD Gateway
Number	64	2	4
Chassis	Dell PowerEdge R760xd2	Dell PowerEdge R6615	Dell PowerEdge R6615
CPU	1 x Intel® Xeon® Silver 4514Y (16C/32T)	1 x AMD EPYC 9374F (32C/64T)	1 x AMD EPYC 9354P (32C/64T)
Memory	8 x 32GB DIMMs	12 x 16GB DIMMs	12 x 16GB DIMMs
OS Disk	1 x 480GB SSD	2 x 480GB (RAID 1) SSD	2 x 480GB (RAID 1) SSD
Capacity	28 x 24TB HDD	1 x 3.2TB Mixed use NVMe – for LevelDB	1 x 3.84TB Read Intensive NVMe – for logging
Network	Nvidia 25GB/s NIC	Nvidia 25GB/s NIC	Nvidia 100GB/s NIC

Tape Upgrades

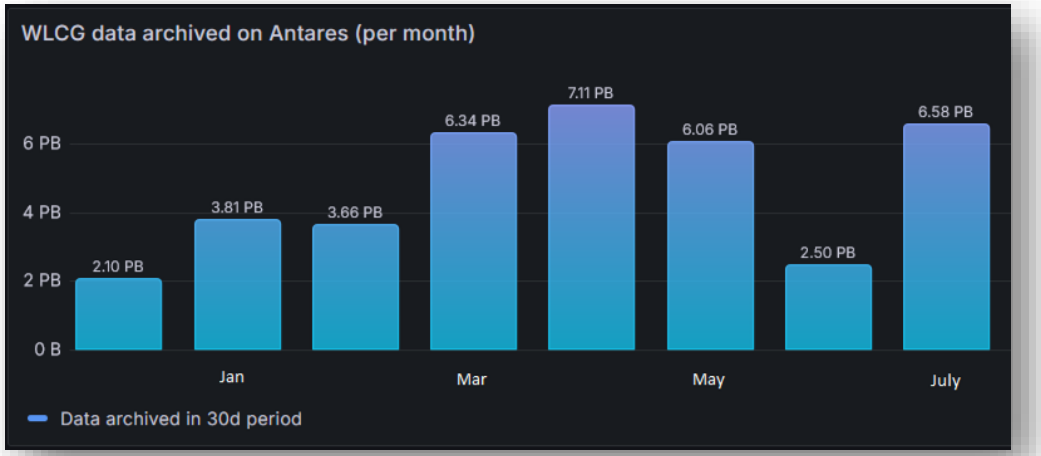
- Majority of Tape service hardware was purchased in 2019 / 2020.
  - Ongoing upgrade and replacement project.
- 83,250 TB capacity upgrade using 8 x IBM TS1170 drives and media.
  - We will continue to use our 16 x LTO-9 and 20 x TS1160 drives for majority of writes for new data.
  - TS1170 will be repacking old data continuously in background.
- 7 x new EOS nodes for tape (pipeline) buffer:
  - 16C/32T CPU, 192GB memory, 480GB NVMe OS disk, 8 x 7.68 NVMe (1 DWPD), 1 x 1.6TB NVMe (3 DWPD), 100Gb Nvidia NIC.
  - New nodes will be dual stack and accessible via LHCOPN and LHCONE.

Repack Problem



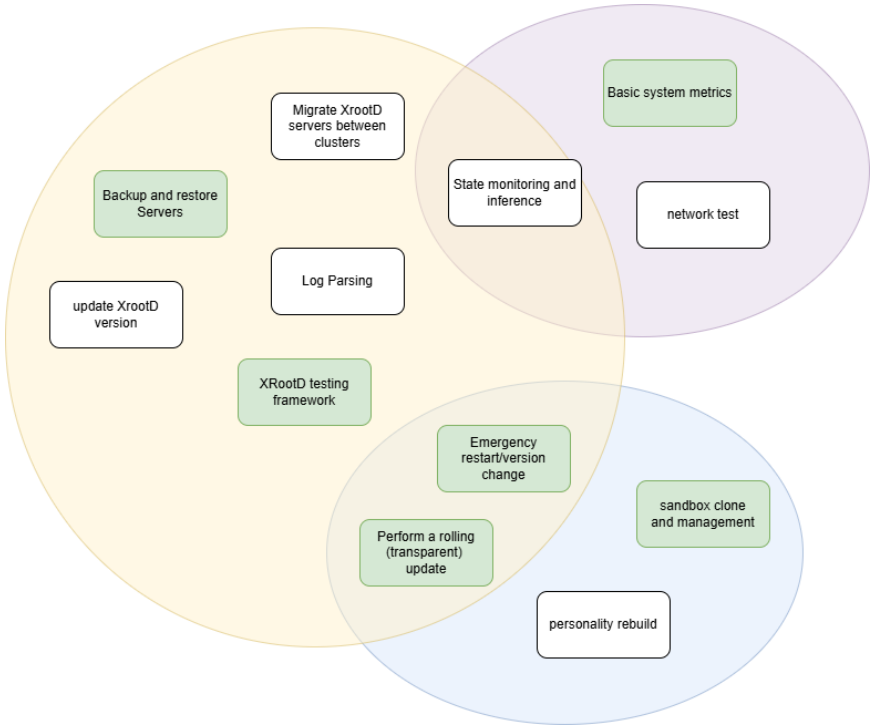
Tape Upgrades

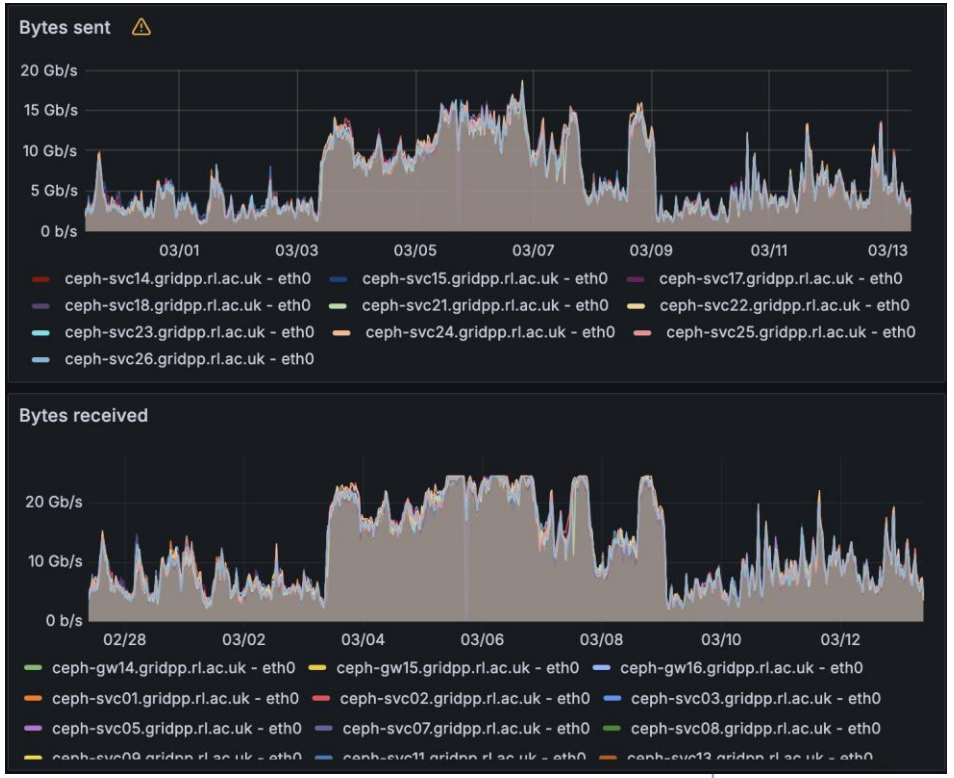
- Tape writes rates can peak at >10GB/s.
- We can repack at 1.5PB a week.
- Last year during data taking there were sustained periods where we exceeded that.
- Plan was to purchase Tape upgrade early so we could do several months of repacking…



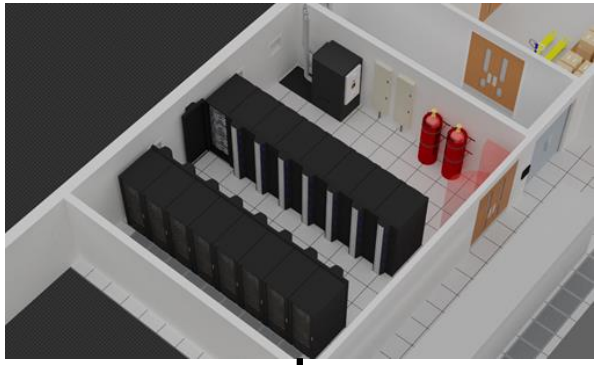
Data Transfer improvements

- 100Gb/s gateways now available for deployment.
  - We don’t really know where to expect the bottlenecks to appear when we try scaling up.

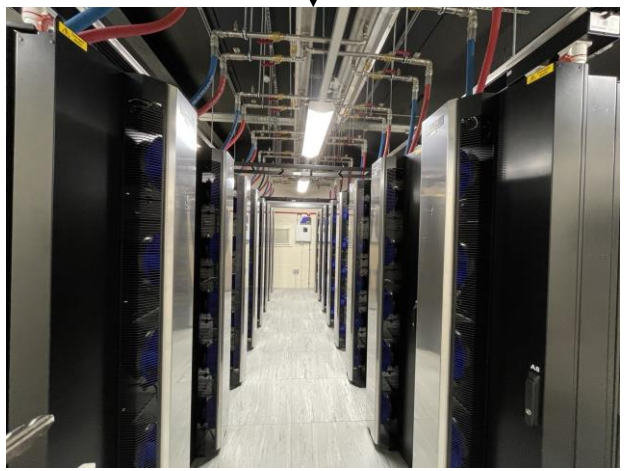




Net Zero & new data centre



- New “Ultra High Power” Data Centre Room has been built at RAL.
  - 600kW capacity.
  - PUE of 1.2 expected.
- Proposal: Move existing Tier-1 batch hardware to this room.



ΔPUE = 0.1
£ / kWh = 23p
gCO_2 / kWh = 143
Batch Farm ~200kW

Potential saving:
£40k / year
25,000 $kgCO_2$ / year

APART FROM TAPE
WHAT HAS THE TIER-1 EVER DONE FOR US?
imgflip.com

Where can Tier-2s help?

- Tier-1 pledges are dominated by ATLAS and LHCb.
  - We want help with the hard stuff, not the MC simulation
- Networking is very important!
  - Join LHCONE?
- I am always worried about the backend storage.
  - The Data Intensive work really stresses it.
  - Make sure you have spare capacity if something fails.
- It’s always your fault, even when its not you still have to fix it.

Science and Technology Facilities Council

Questions?