Science and Technology Facilities Council

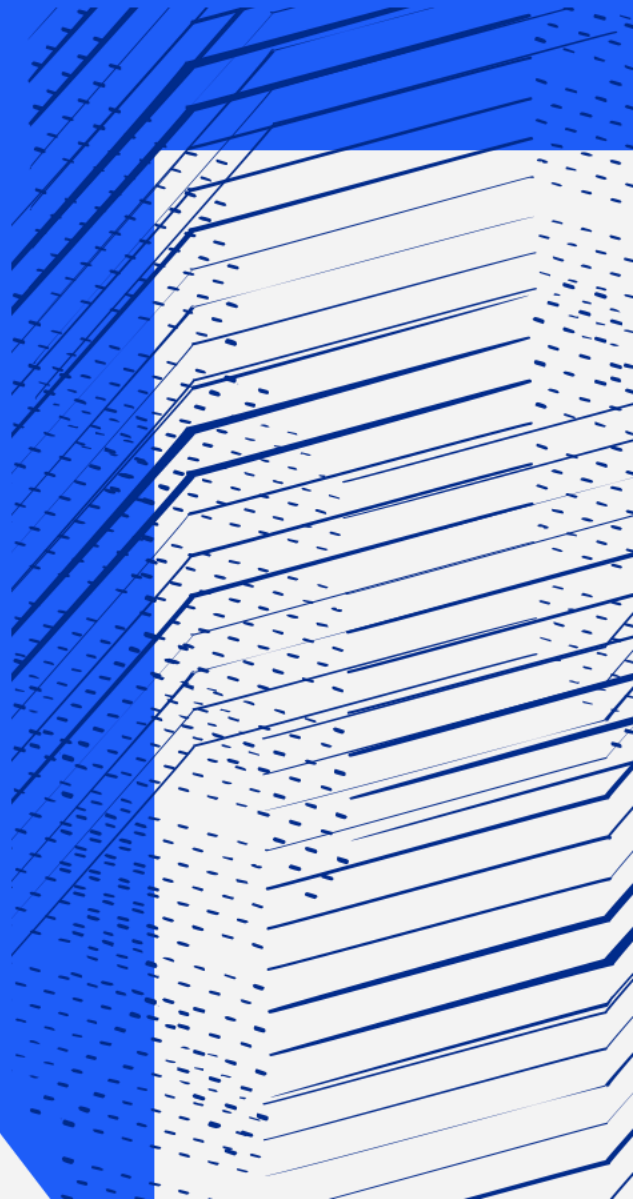
Scientific Computing

RAL Site Report

HEPiX Fall 2024, Oklahoma

Martin Bly et al.

November 2024



Outline

- DC stuff
- OS update
- Services
- Hardware
- Tape

Thanks to colleagues for their input

Summary of Tier1

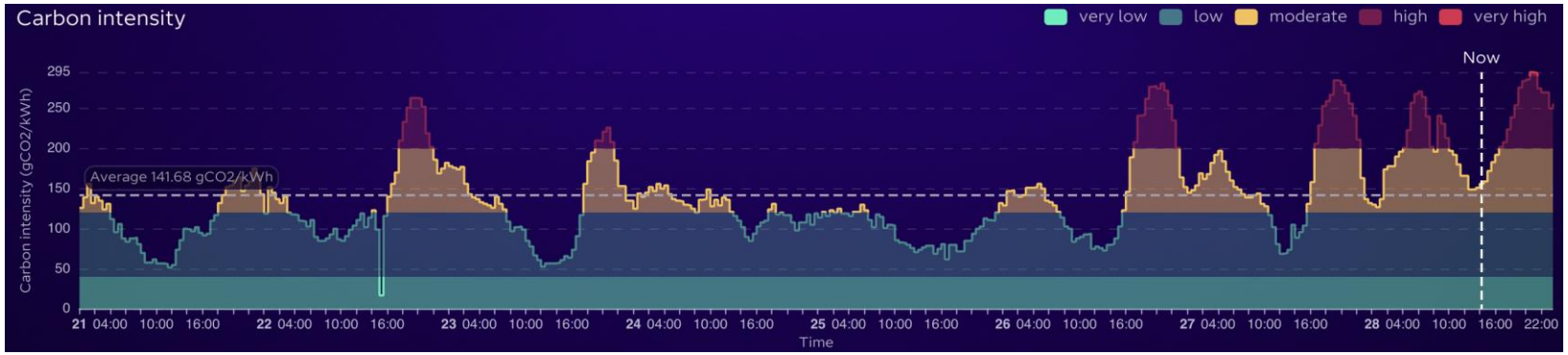
- Funded by GridPP project to provide UK Tier1 facility for WLCG and to support other VOs and projects as required
  - Storage
  - Compute
  - Tape service
  - Archive
- Supports Alice, ATLAS, CMS, LHCb + many others
- Datacentre also hosts JASMIN data facility, STFC's cloud, SCARF, DAFNI, … (and SKA in future)

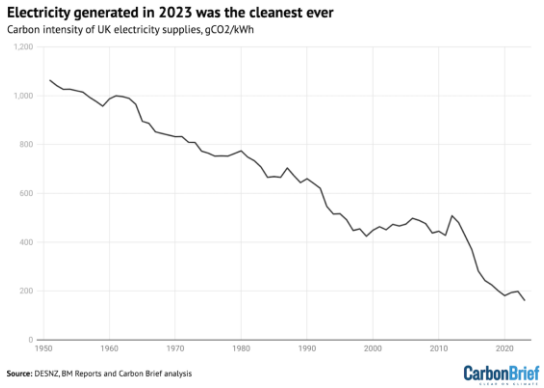
New DC Update

- Converting one of the operations areas in the existing DC
  - Commissioning during November
    - 400kW of heaters to help tune the cooling
- Summary:
  - 16 x 750mm wide racks, water cooled doors
  - 600kW N+1 chillers (~37kW rack)
    - Space for third chiller
  - Services, power etc., - top feed
  - No UPS, compute only
- RCC project is delayed due to finance and wider review of DC provision for STFC and UKRI
  - Subject to UK government spending review
  - Landscape for DC need and provision will have changed by the time any funding would actually be available.

Net-Zero Goals

- What are we trying to achieve with Net-Zero?
  - While we would like to minimize our carbon usage we also have SLA to meet and a finite amount of effort and capital.
- UK government aims for our energy generation to be Net-Zero by 2035.
  - Aims for 95% Net-Zero energy generation by 2030.
- Reduce power usage if it leads to minimal performance loss.
- Keep hardware running longer.
- Temporarily reduce power usage when carbon intensity is high.





https://www.carbonbrief.org/analysis-uk-electricity-from-fossil-fuels-drops-to-lowest-level-since-1957/

OS migration: SL7 -> ?

- Have been running SL7 for a long time, some hosts on OL7
- Most now moved to:
  - Rocky 8, with some Rocky 9 where limited options for HW or SW support
  - Oracle DBs migrating to OL8 (support for OL7 until end of 2024)
  - Very small number of services remain on RHEL7 on extended license
- Used as opportune moment to refactor/upgrade/decommission services
- ~3300 Hosts/VMs migrated
- 250 Service component configurations required migration
- Migration took for majority to be completed, some went early, a few stragglers
- Challenges
  - Python2 -> Python3 for some services
  - Incorporation into our config management system (NetworkManager in Rocky 9)

Batch service / WNs

- Configuration and enabling of IPv6 connectivity for ARC-CE's and WNs
- Successful implantation of tokens for job submission in the pre-production environment
- Optimisation of fairshare logic to remove multicore tranche across Batch Farm
- Rollout of vector-read fixes for local xrootd instances on Batch Farm WNs. Now included in upstream xrootd project (5.7+)
- Support LSST as a new VO for the Tier1.

Batch service / WNs – next few months

- Upgrade to HTCondor 24 (or 23 depending on testing)
- Start planning for local xrootd-gateway rearchitecting
- Rollout token support for job submission into production Batch Farm
- Create and test new job queues for GPU resource requirements
- Support Moedal and Comet experiments

Storage (Echo)

- Large Erasure Coded Ceph object store
- 100PB storage (raw)
  - Everything to el8 (Rocky 8)
  - Moving to High-Level Failure domains has been promoted from "Next year" to "Next 6 months"
    - Not necessarily rack-level any more
  - Finish upgrade to Pacific
  - Upgrade to Reef in planning, yet to be scheduled

Data Intensive Processing

- Echo is built on Ceph which provides 73PB of usable storage across 268 servers and more than 6000 HDD.

In the last 90 days:

77.64PB
of data transferred

144,560,889
total transfers

Data Transfer improvements

- 100Gb/s gateway is waiting for deployment
  - We don’t really know where to expect the bottlenecks to appear when we try scaling up
- XrootD development:
  - Deletions – can we scale or do we need async?
  - Writable WN gateways
  - Containerized XRootD
  - Improving buffer layer in XrdCeph

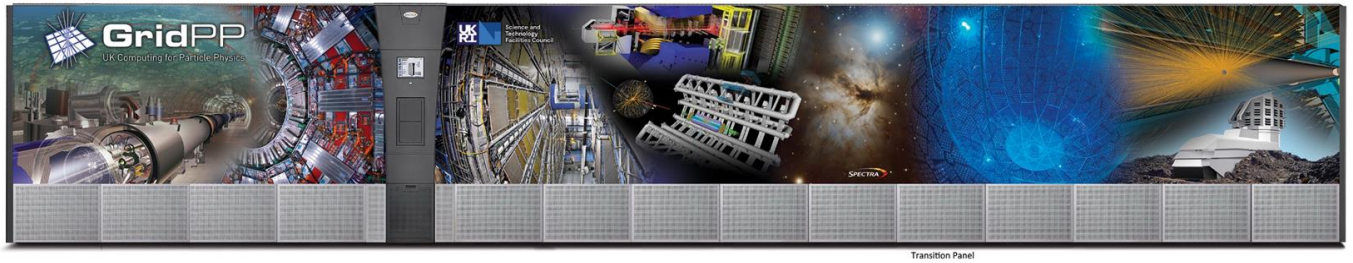
CVMFS Stratum services

- Service architecture using a CephFS mountpoint to store all relevant data for both the Stratum-0 and the Stratum-1 did not work
  - operations were too slow, and it did not scale
  - after migration to Rocky 8, the CVMFS Stratum-0 server was not even capable of writing output to CephFS
    - Most probably some incompatibility issue between the kernel version of the host and the [old] version of CephFS.
- Took service offline, redirecting clients to other CVMFS instances
- New architecture: bespoke servers using FermiLab spec
- For the Stratum-1, instead of CephFS, we store the data on ZFS
  - Each CVMFS replica uses its one ZFS pool
  - Currently, there is no synchronisation between the pair of Stratum-1 hosts.
    - In future the plan is to let them update each other using ZFS snapshots.
- Expecting new service on-line shortly, possibly this week.

Hardware

- 2023/24
  - Storage and Worker procurements in service
- 2024/25
  - New CVMFS service nodes, entering service soon
- Procurement
  - Storage
    - Tender for replacement of older generation of Ceph storage nodes
  - Compute
    - None expected

Tape



- 2 x Spectra Logic tFinity libraries
  - “Asterix” 15 frames (3 drive frames) 1400 slots licensed
  - ~140PB GridPP data stored
  - 20xTS1160 drives, 5076 TS1160 media
  - 16xLTO9 drives, 3980 LTO9 media
- “Obelix” 13 frame (3 drive frames) 1180 slots licensed
  - ~150PB facilities data stored
  - 18xTS1170 drives, 935 TS1170 media
  - 24xTS1160 drives, 6131 TS1160 media
  - 6xLTO9 drives, 1386 LTO9 media in the library, 305 stored offsite
  - 17xLTO8 drives, 728 LTO8/LTOM8 media in library, 2185 stored offsite

Tape II

- Data in both libraries managed via CTA
- Small DMF system also using Obelix and two TS1160 drives
  - ~150TB on disk cache, 1.5PB on tape (3 copies)
  - Chiefly science archive data
- Drives use FC to connect to FC switches and to the tape servers
  - We have a mix of 2 or 4 drives/server
- Each library has 2 RIM units for the control paths
  - We have had issues with these in the past, now have a better understanding of the causes, so we can mitigate the problems
  - Now have license key for the ADI interface where the control path goes via the tape drives
    - Will be evaluating this over the next few months

Tape III

- One old StorageTek SL8500 library
  - 6 T10KD drives
  - Only one service using library:
    - Old (~30+ year) backup service
    - Minimal in-house knowledge of the system now
    - Project to migrate service to a NetBackup/Veritas based service
    - Only has maintenance till end of March 2025
  - To be decommissioned…
  - Space taken up by library likely to be required for other hardware

Network

- Mostly it's "done". The plans for the next few months are mostly to do with properly demoting the external routes to the old network.
- We're also now tagging BGP communities on LHCONE as per the MultiONE initiative

Science and Technology Facilities Council

Scientific Computing

Questions?