Science and Technology Facilities Council

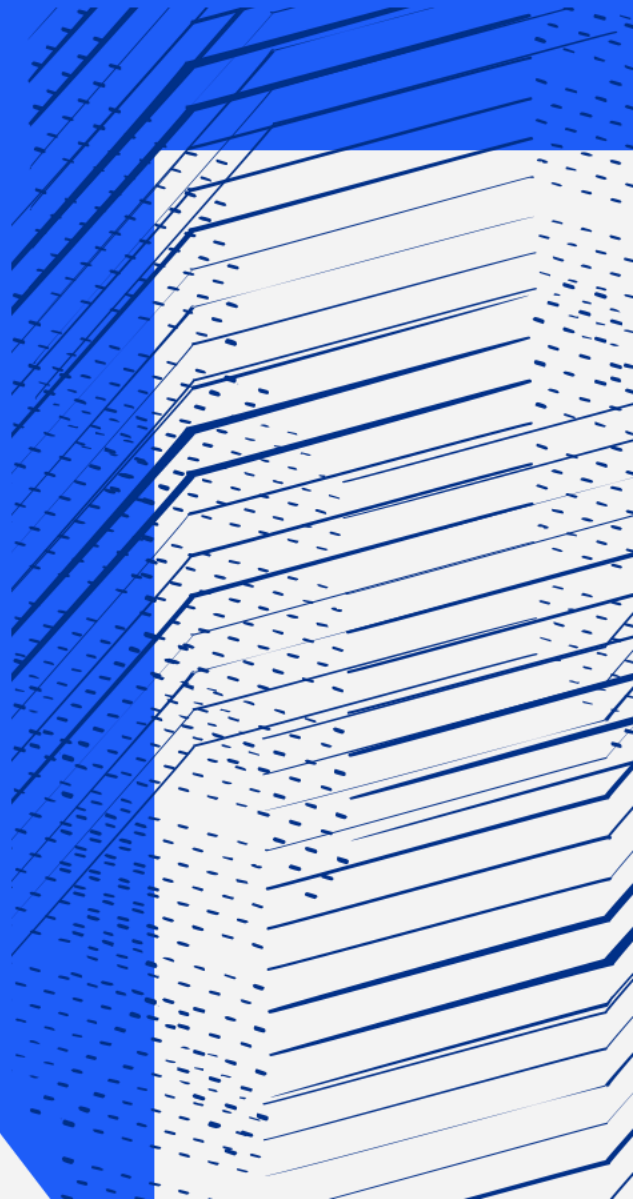
Scientific Computing

RAL Site Report

HEPiX Autumn 2023, Victoria BC

Martin Bly et al.

17th October 2023



Outline

- Tier1 + Plans
- Cloud
- Data Centres

Thanks to colleagues for their input

RAL Tier-1 Plans

- The next iteration of Tier-1 funding, known as GridPP7, covering April 2024 to March 2028 has been confirmed.
  - Funding increase (just about) covers inflation.
  - Strong guidance for more innovation particularly around NetZero.
- For storage, we will be looking at deploying an increasing amount of SSD storage and how that can be effectively used.
- For compute, we will be exploring non-x86 options.
  - We have installed a test Ampere Altra Max server and plan to buy 3000 – 5000 cores this FY.
  - We have access to different types of GPU via the STFC Cloud and will deploy them as necessary.

Tier-1 Procurement

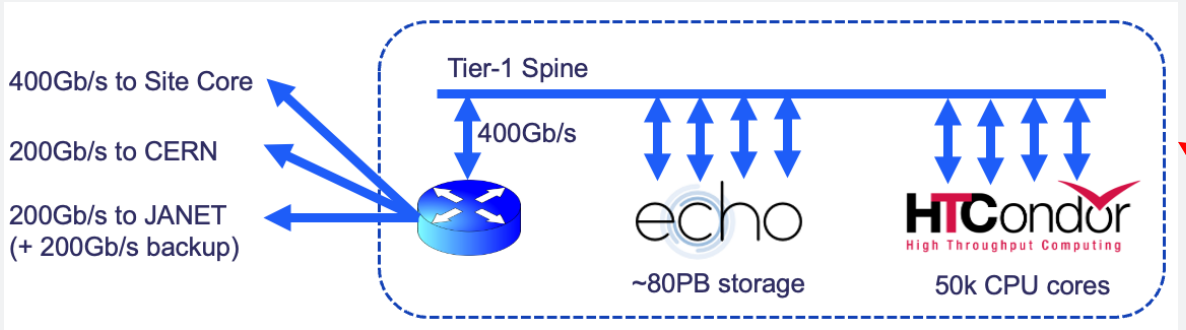
- This year we have 4 main procurements:
  - Storage ~20PB raw
  - CPU: ~ 100k HS23 x86 and ~ 50k HS23 ARM
  - VMWare Cluster replacement.

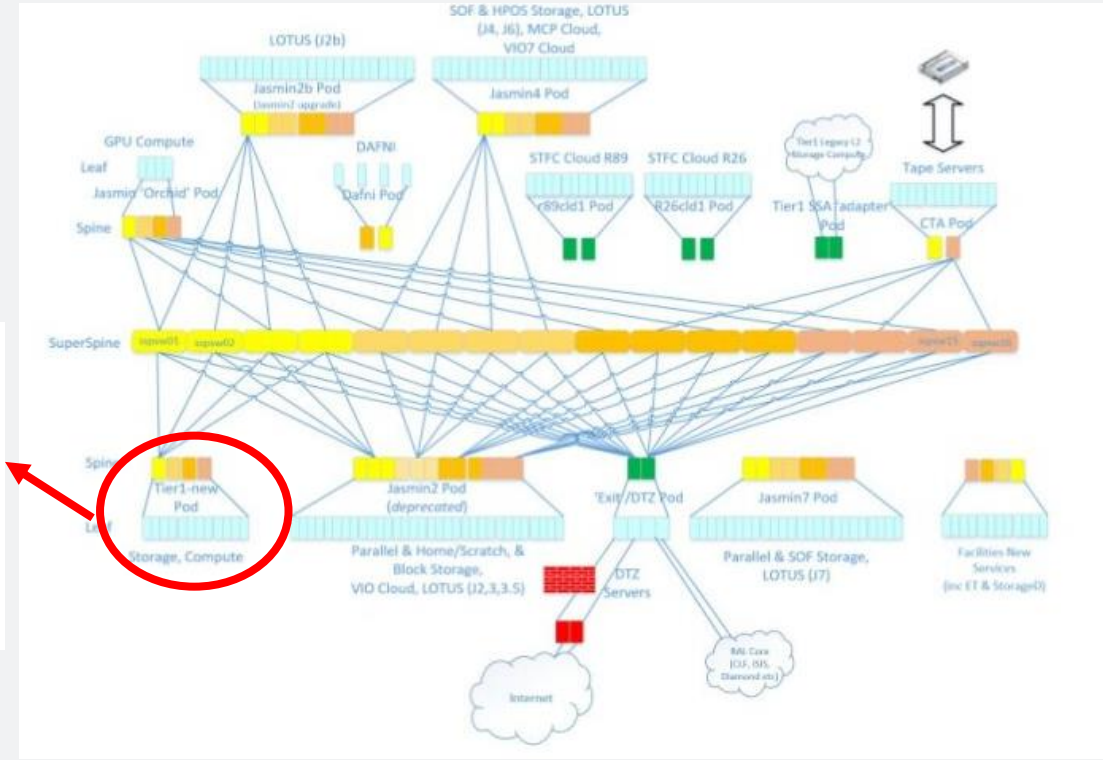
- No additional tape media required as rates have been lower than expected this year.
  - Particularly LHCb whose usage is 20PB lower than their request!

- The Storage and CPU procurements are tenders with orders expected to be placed before the end of the year.
  - Where possible we combine our procurement with other projects within Scientific Computing.

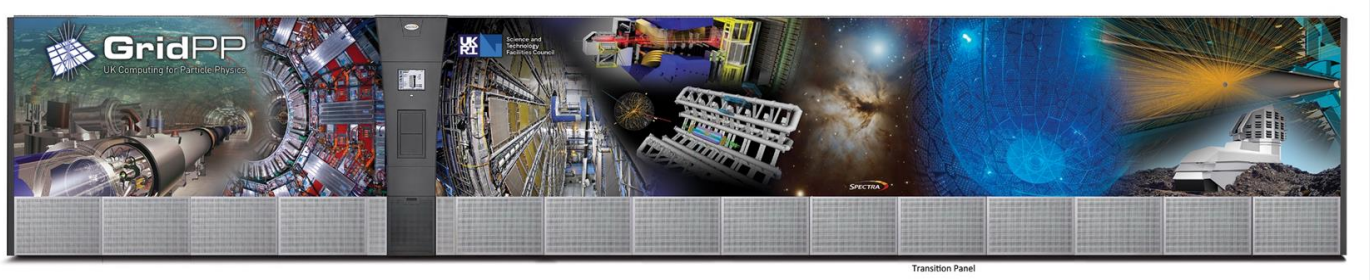
Tier-1 Network

- For the last 3 years there has been significant working within RAL to upgrade the network.
  - Completely replace the Tier-1 network
  - Build a network Super Spine to link the various HPC projects.
- RAL Tier-1 is on LHCONE
  - Recently announced 2001:630:54::/52
  - WN have been for several months (IPv4 only), storage (dual stack) is being migrated now.
- Routing preference:
  - LHCOPN
  - LHCONE
  - JANET link





Tape Libraries

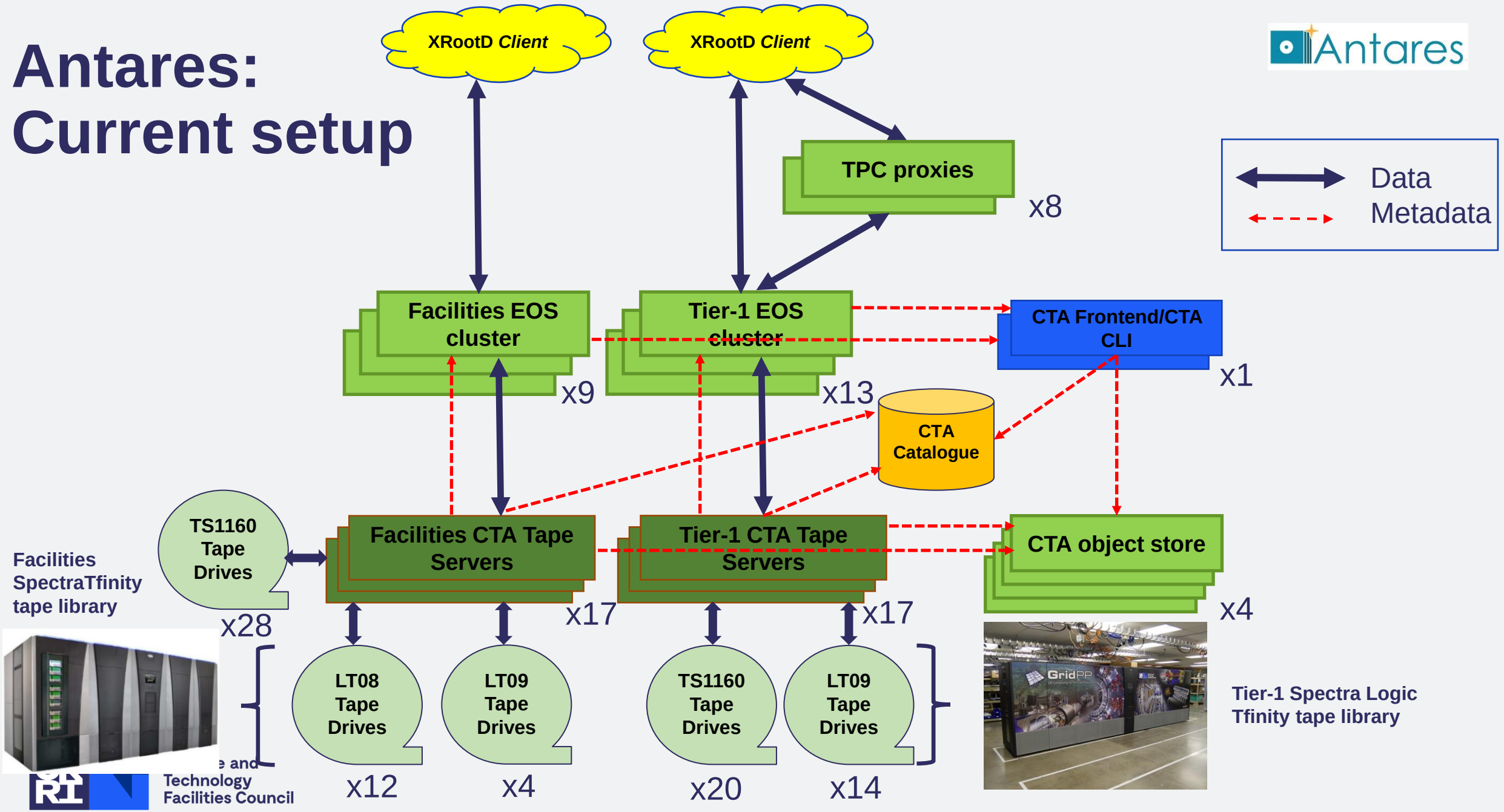


- 2 x SpectraLogic tFinity libraries
  - Asterix (WLCG) – 15 Frames
    - Potential for 240PB data, currently 95PB, lots of space for LHC Run 3 data
    - 20 x TS1160 from 10 tape servers, 5157 media
    - 16 x LTO9 from 8 tape servers, 1680 media
  - Obelix (Others) - 13 Frames
    - Potential for 210PB data, currently 120PB
    - 34 x TS1160 from 10 tape servers, 5888 media
    - 6 x LTO9 from 3 tape servers, 780 media
    - 16 x LTO8 from 4 tape servers, 1870 media
- 1 x Oracle SL8500
  - System backups and DMF service
  - Being phased out, DMF migration to tFinity library imminent

CTA Tape service (Antares)

- 2 EOS clusters, Tier-1 and Facilities, against a single CTA instance
- EOS: 4.8.105-1, CTA: 4.8.7-1
- Tier-1 EOS cluster: 13 x 1.5TB SSD nodes
  - Ingest space: 200TB, Retrieve space: 172TB
- Facilities EOS cluster: 2 x 3.5TB + 7 x 1.5TB SSD nodes
  - Ingest space: 190TB, Retrieve space: 265TB
- Migration of the last CASTOR instance completed in early June

Antares: Current setup



Facilities SpectraTfinity tape library

Tier-1 Spectra Logic Tfinity tape library

Ceph Storage

- Clusters
  - Echo – 80PB HDD erasure coded object storage for the WLCG
  - Deneb – 8PB HDD erasure coded CephFS for local facilities
  - Arided – 1.2PB SSD 3*replicated CephFS providing sharable storage for internal cloud
  - Sirius – 450TB 3*replicated RBD block storage for private cloud
- Ceph software upgrade project underway
  - Standardise all clusters on Rocky8/Ceph Pacific
- Major expansion of Echo's external gateway hosts expected shortly
  - IPv6 access over LHCONE for all remote sites now sorted ☺
- See talk by Rob Appleyard on Wednesday:
  - https://indico.cern.ch/event/1289243/contributions/5590766/

Tier1 Batch Service

- Running HTCondor
  - Condor 10 beginning of November
    - This will give us continued LTS support for Condor
- Working on:
  - Continued testing of Tokens through ARC-CE, most notably the testing of ARC-CE7 when available.
  - xrootd container optimisation and architectural “re-imagining” of how the Docker xrootd containers will interact.
  - Rocky 8 for the ARc-CE’s
    - Likely will take the longest, lots of continued testing

Cloud Services

- Expanded use of SCD storage.
  - Arided is now in production (CephFS) for the cloud.
  - Significantly expanded our object store offering and usage using Echo.
- Expanded GPU offering. ~500 GPUs in service. A100 and V100 compute GPUs, A4000 and RTX4000 Visualisation GPUs.
- Experimental deployment of AMD W6600 and Intel ARC A770 GPUs to explore market alternatives.
- Significantly expanded size of team to support more services/user communities. Cloud Group now consists of distinct Operations and DevOps Teams.

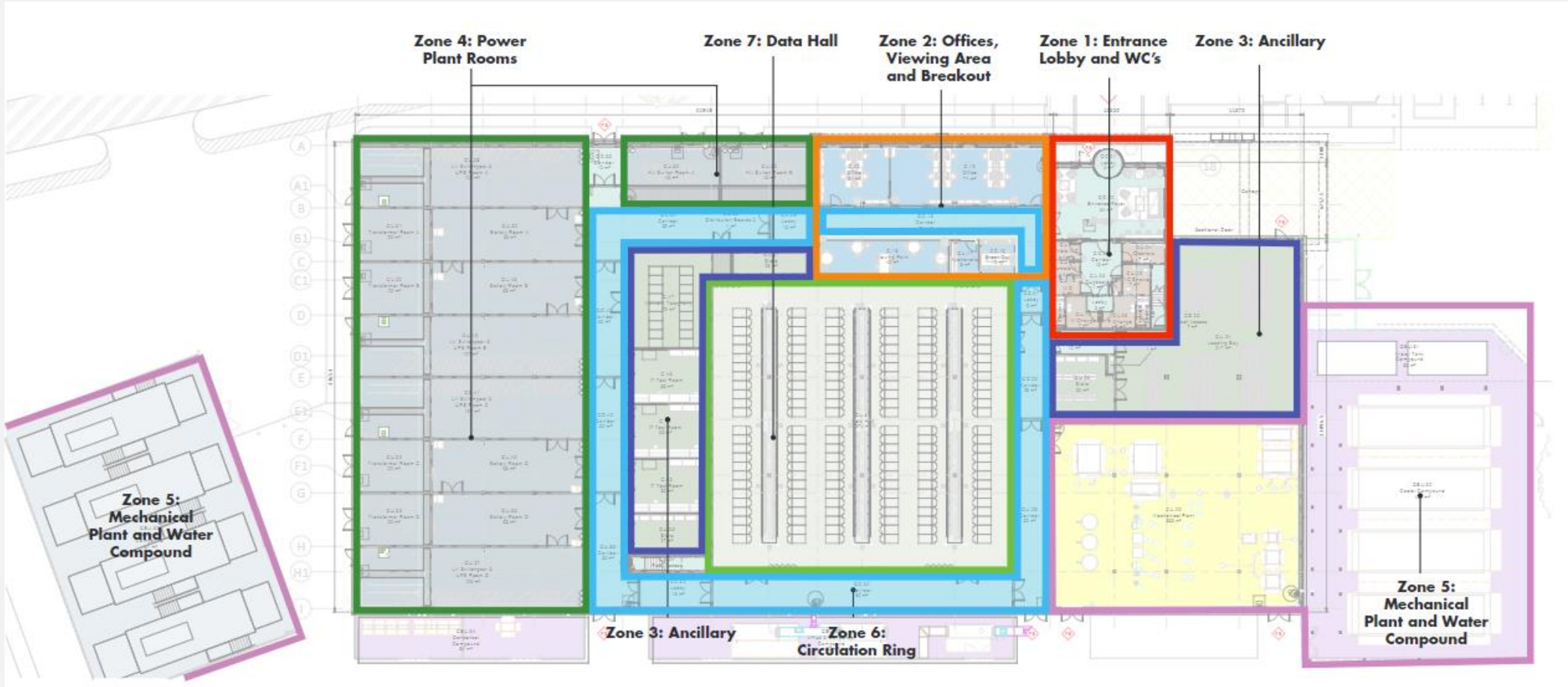
Cloud Plans

- Series of OpenStack version upgrades planned to move to a more recent version. Train -> Antelope/Bobcat.
- Reworking deployment method to make this easier in the future. From RPM deployment to Kolla Ansible and Kayobe
- Deployment of OpenStack Ironic planned to offer bare metal CPU and GPU services to cloud users

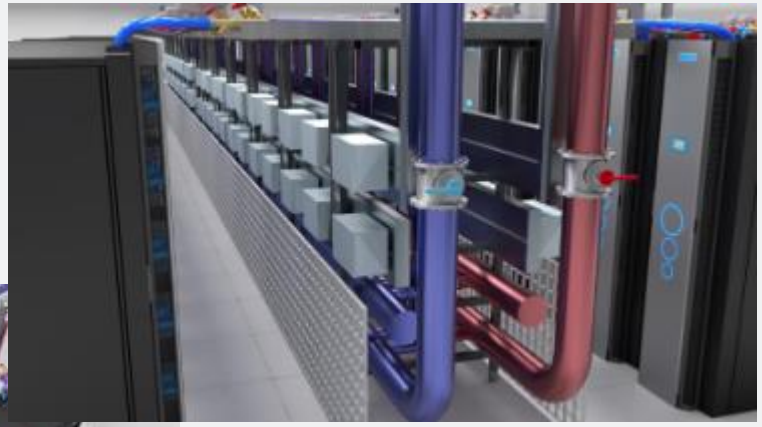
New Data Centre - RCC

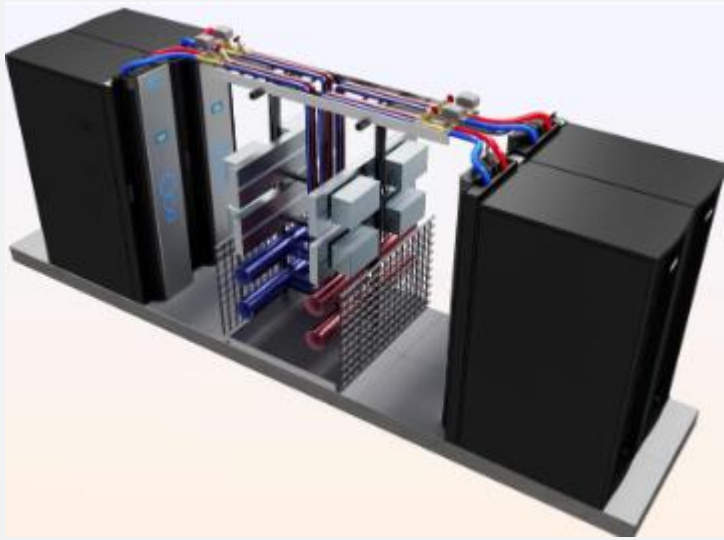
- What, another one?
  - Yes! – the Research Computing Centre
- Targeted for 6.5MW of IT load
- 150 racks in the data hall, average of ~43kW/rack
- Rear door cooling on each rack
  - optimum for 10-65kW, able to do up to ~95kW
  - => flexible rack loading
- 4 separate power rails each with UPS and generator
  - UPS for smoothing and bridging load while generators start
- Aim for annualised PUE of 1.1 at load
- Compute only, no storage services

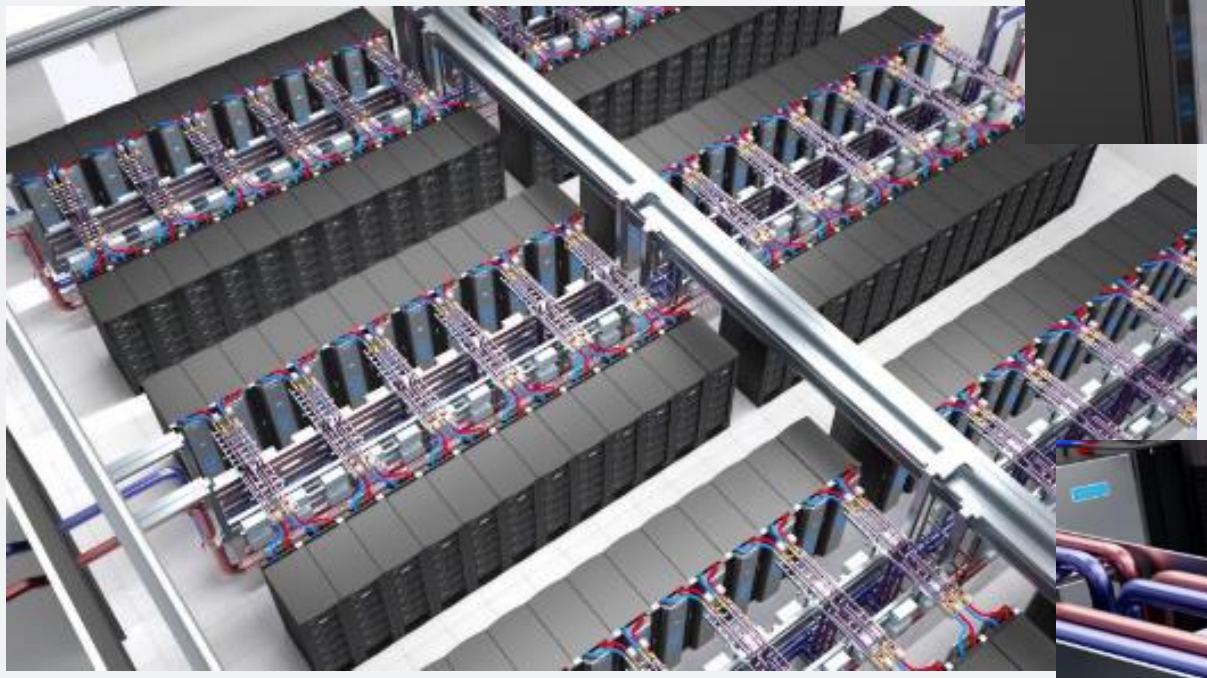
RCC – R130

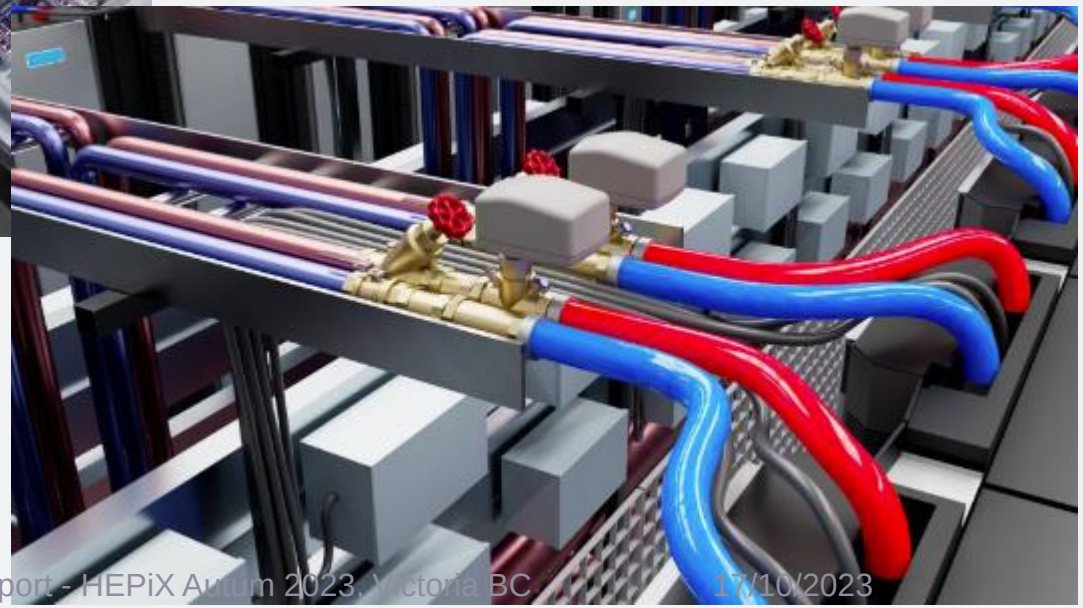

Zone 4: Power Plant Rooms
Zone 7: Data Hall
Zone 2: Offices, Viewing Area and Breakout
Zone 1: Entrance Lobby and WC's
Zone 3: Ancillary
Zone 5: Mechanical Plant and Water Compound
Zone 3: Ancillary
Zone 6: Circulation Ring
Zone 5: Mechanical Plant and Water Compound

RCC – R130









Science and Technology Facilities Council

Scientific Computing

While we wait… New DC part 2

- RCC will not be available for a while
  - Moving through the various RIBA stages for buildings
- In the meantime, looked at a modular data centre
  - Interim space for high power density facilities
    - Several iterations…
  - Budget is for this FY only!
  - Current plan: converts one of the operations areas in the existing DC
    - Makes use of some planned infrastructure changes and upgrades, with new coolers
    - 16 x 750mm wide racks, water cooled doors, rack power depends on size of coolers
    - Services, power etc., - top feed
    - No UPS, compute only

Science and Technology Facilities Council

Scientific Computing

Questions?

Science and Technology Facilities Council

Scientific Computing

Thank you

scd.stfc.ac.uk

@SciComp_STFC