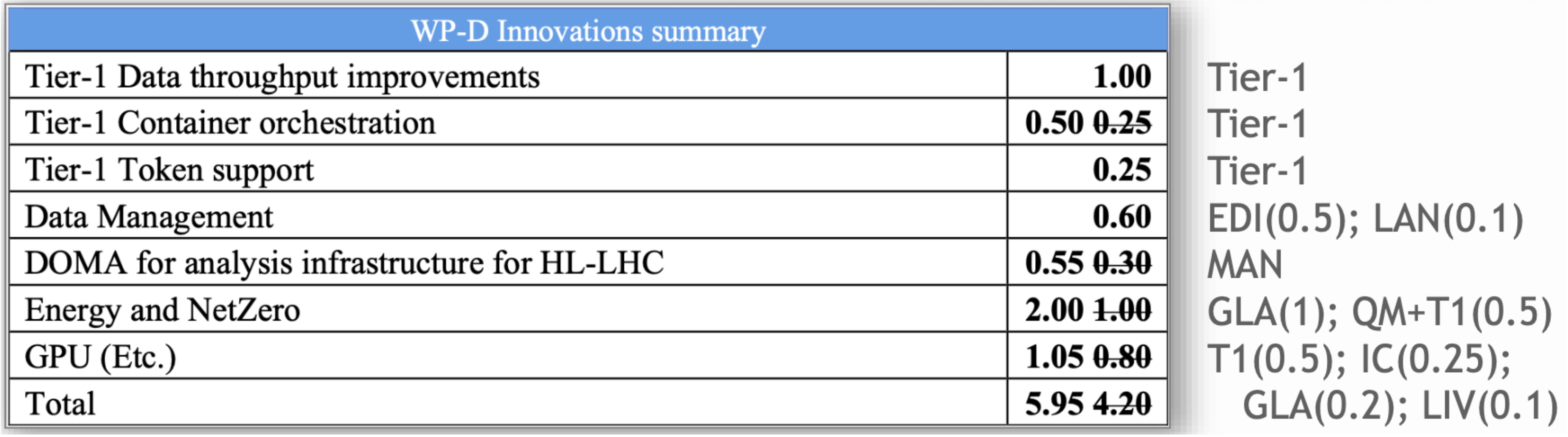
UKRI Science and Technology Facilities Council

Tier-1 Update

Alastair Dewhurst

Introduction

- High Level Tier-1 Status
  - Tier-1 Operational highlights will be covered by Liaisons.
- Tier-1 Plans for the year
- WP-D storage side

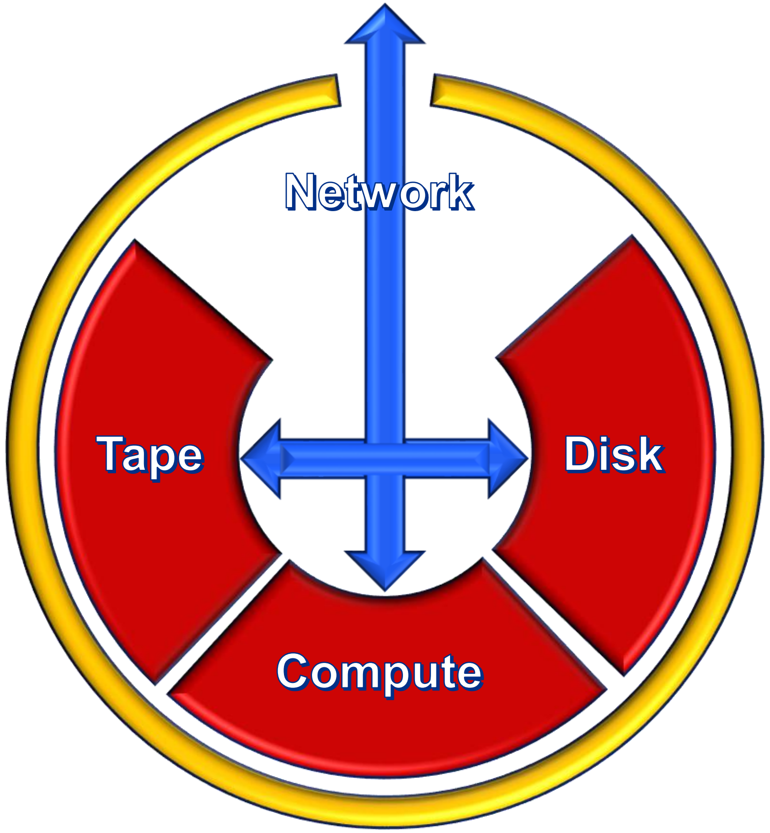


WP-D Innovations summary			
Tier-1 Data throughput improvements	1.00	Tier-1	Alastair
Tier-1 Container orchestration	0.50 ~~0.25~~	Tier-1	Alastair & Tom
Tier-1 Token support	0.25	Tier-1	Tom
Data Management	0.60	EDI(0.5); LAN(0.1)	
DOMA for analysis infrastructure for HL-LHC	0.55 ~~0.30~~	MAN	
Energy and NetZero	2.00 ~~1.00~~	GLA(1); QM+T1(0.5)	Alastair & Tom
GPU (Etc.)	1.05 ~~0.80~~	T1(0.5); IC(0.25); GLA(0.2); LIV(0.1)	Jyoti
Total	5.95 ~~4.20~~		

GridPP Tier-1

The GridPP Tier-1 is a world leading data intensive processing and archival platform.

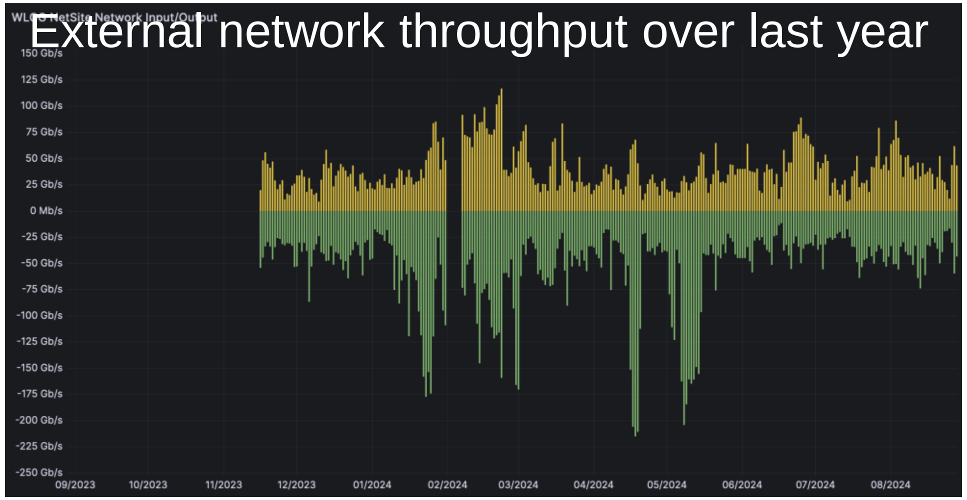
We use Cloud Native Technologies to build efficient scalable infrastructure



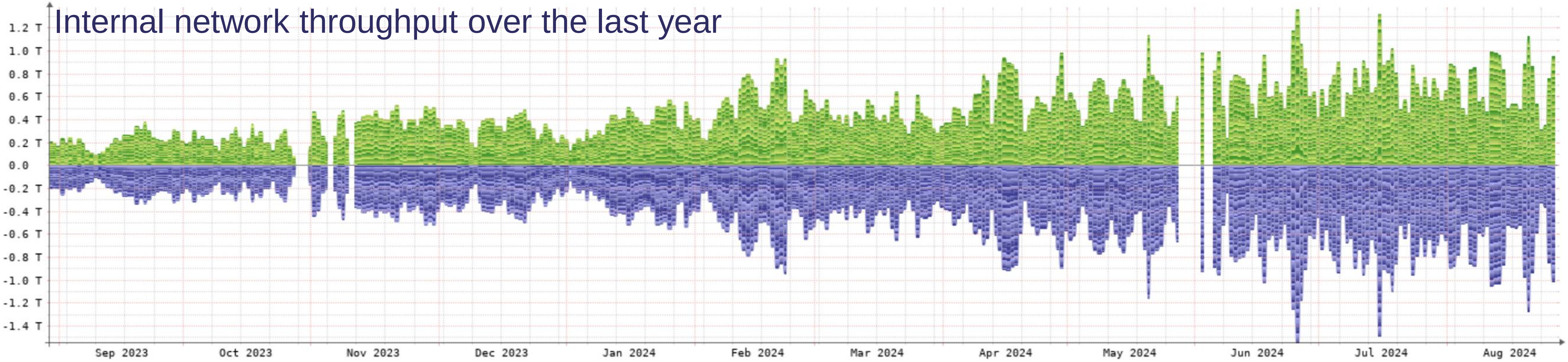
State of the art security infrastructure provides access to thousands of researchers.

Data Intensive Networking

- 200Gb/s link to CERN
- 400Gb/s link to JANET
- Leaf / Spine internal network provides non-blocking connectivity.



External network throughput over last year



Internal network throughput over the last year

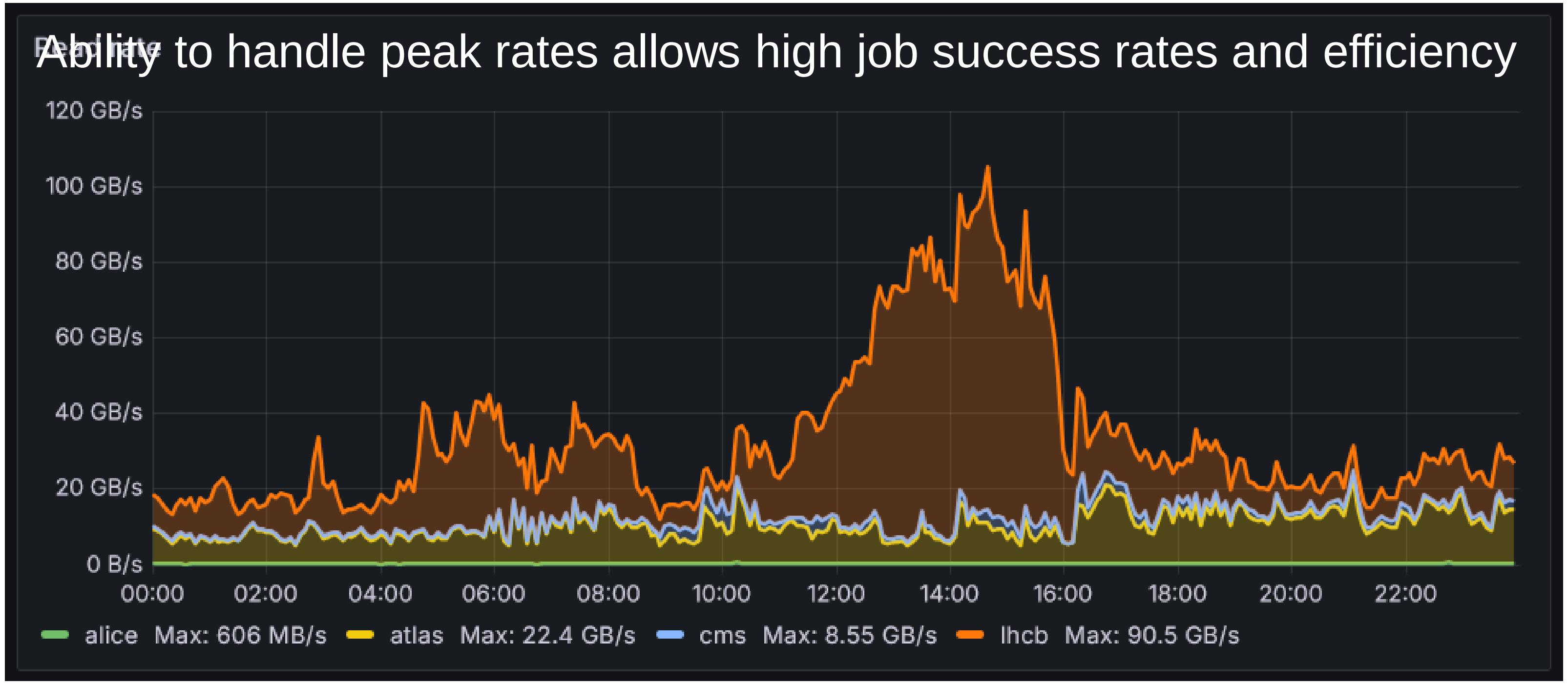
Data Intensive Processing

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

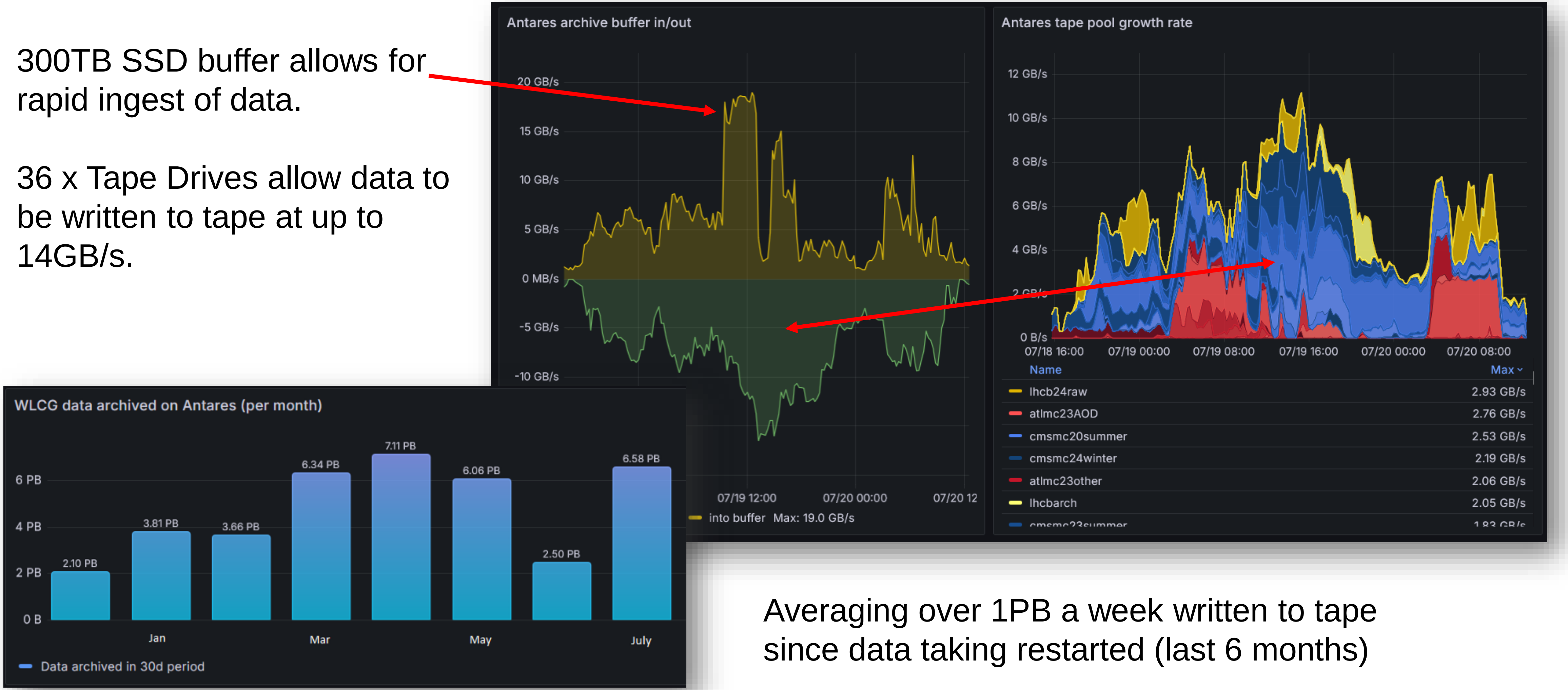


Ability to handle peak rates allows high job success rates and efficiency

Data Intensive Archival

300TB SSD buffer allows for rapid ingest of data.

36 x Tape Drives allow data to be written to tape at up to 14GB/s.



Name	Max
lhcb24raw	2.93 GB/s
atlmc23AOD	2.76 GB/s
cmsmc20summer	2.53 GB/s
cmsmc24winter	2.19 GB/s
atlmc23other	2.06 GB/s
lhcbarch	2.05 GB/s
cmsmc23summer	1.83 GB/s

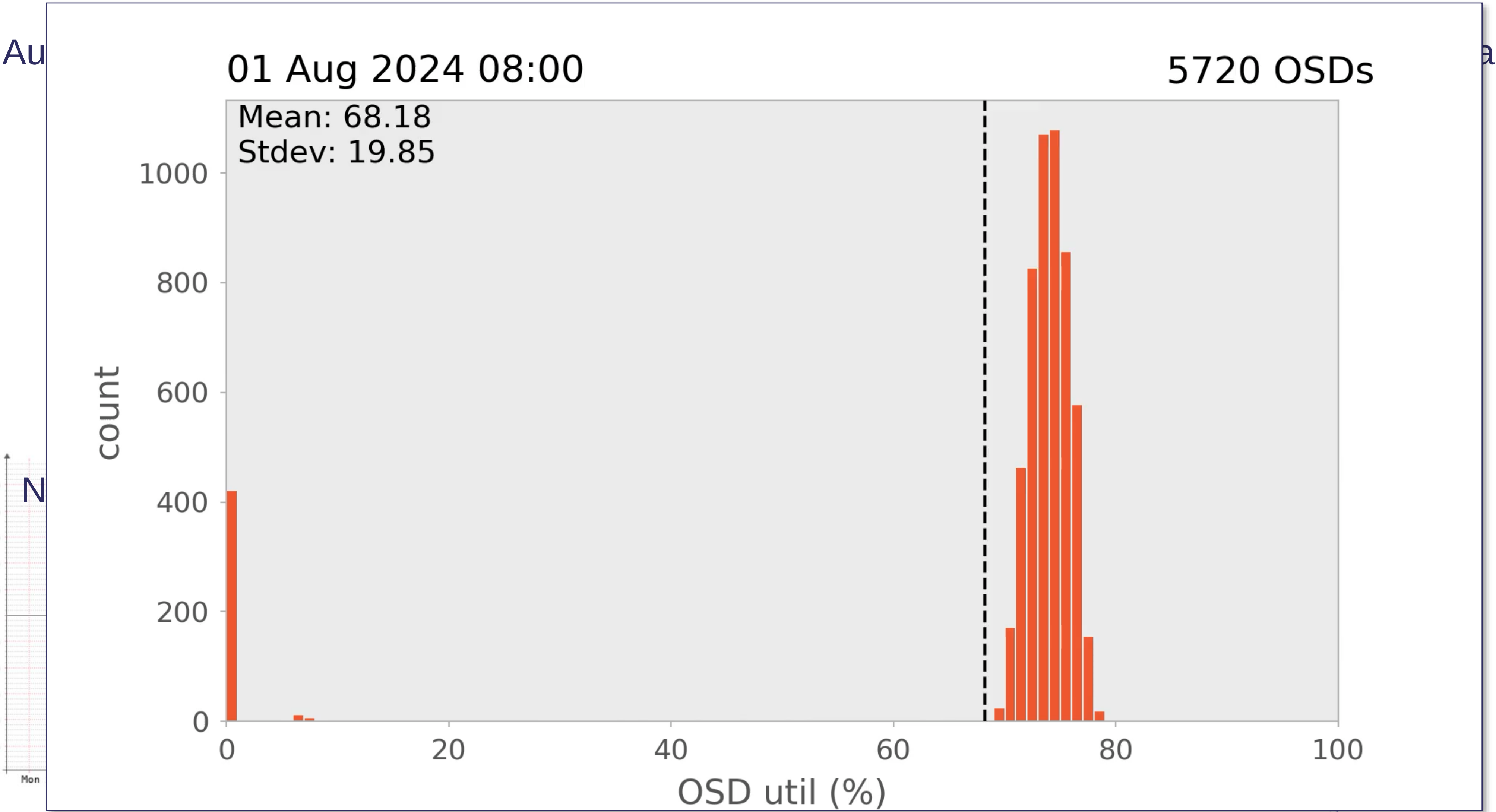
Averaging over 1PB a week written to tape since data taking restarted (last 6 months)

Rest of Talk Outline

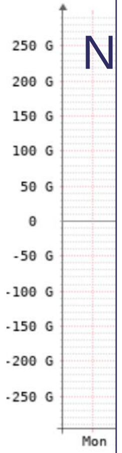
- Echo hardware addition
- Ceph upgrades
- Ceph Cluster management
- Infrastructure upgrades
- Antares EOS upgrade
- Data Transfer improvements
- Hardware management
- CVMFS upgrade
- Net-Zero
  - New Data Centre Room

Hardware Addition

In Au acity



N



Ceph upgrades

- Do 4 major version upgrades of Echo.
  - Double upgrade soon (once we are back from GridPP52).
  - Double upgrade before start of next years data taking.
- Concern: Ceph have dropped support for “Pacific” on EL8 (our intermediate state).
- In the first half of August Deneb was upgraded to Pacific.
  - Problem with MDS (needed for CephFS) but Echo doesn’t use this.
- We need to understand when we will upgrade to Rocky 9.
- We will move to using Cephadm on all clusters except Echo.
  - As we gain experience we want to move Echo to Cephadm.

Cephadm

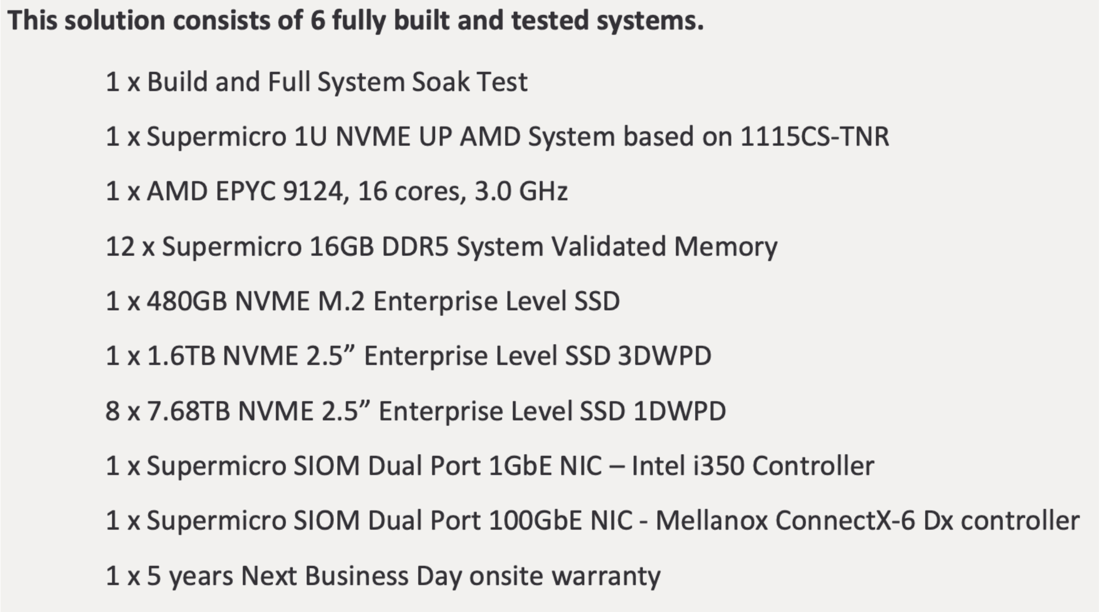
- A simple set of service definitions control daemon placement
  - "Deploy OSDs on all available storage devices"
  - "Deploy monitors on hosts with tag *monitor*"
- The *manager* is continuously scraping hosts for device info and enforcing placement rules
- All services deployed in a standard container runtime and controlled via *system*
- Having a tightly coupled orchestration layer allows for complex operations to be automated
  - Rolling upgrades can be fully automated, including specific upgrade idiosyncrasies
  - `ceph orch host maintenance enter <host>` - put any host into a safe state for intervention, hiding the complexities of different procedures for daemon types
- Supports easily deploying 'extra' services alongside the core Ceph cluster
  - Gateways - *HTTP loadbalancers, highly available NFS gateways, iSCSI and NVMe-oF targets*
  - Monitoring and crash reporting stacks - *deployed by default for new clusters*
  - Trivial to extend - *deploy custom services via cephadm (e.g. XRootD gateways?)*

Infrastructure improvements

- While we have moved the majority of our hardware to the new network the legacy network is still causing problems.
  - Move of final services like the VMWare infrastructure.
  - James Adams making simplifications to remove dependencies.
- While it hasn’t been formally agreed it is looking likely that the Tier-1 (along with other SCD platforms) will move from VMWare to Proxmox.
- We are creating a new network leaf which will be in the UPS room:
  - CVMFS Stratum-1
  - XRootD development gateway.
- We are creating a new network lead for the EOS upgrade.

Antares EOS upgrade

- We plan to upgrade Antares EOS instance this year.
  - Approaching 5 years old.
  - We need to move it to the new network.
- Hardware will be ordered soon.
- New network leaf is being created.
- Hope to have it ready for migration shortly after data taking ends.



This solution consists of 6 fully built and tested systems.

1 x Build and Full System Soak Test

1 x Supermicro 1U NVME UP AMD System based on 1115CS-TNR

1 x AMD EPYC 9124, 16 cores, 3.0 GHz

12 x Supermicro 16GB DDR5 System Validated Memory

1 x 480GB NVME M.2 Enterprise Level SSD

1 x 1.6TB NVME 2.5” Enterprise Level SSD 3DWPD

8 x 7.68TB NVME 2.5” Enterprise Level SSD 1DWPD

1 x Supermicro SIOM Dual Port 1GbE NIC – Intel i350 Controller

1 x Supermicro SIOM Dual Port 100GbE NIC - Mellanox ConnectX-6 Dx controller

1 x 5 years Next Business Day onsite warranty

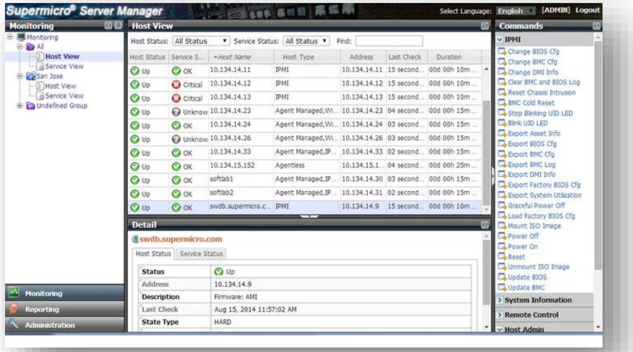
Data Transfer improvements

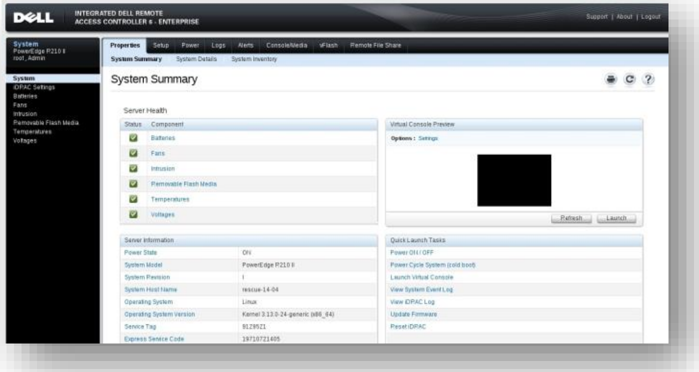
- 100Gb/s gateways is waiting for deployment.
  - We don’t really know where to expect the bottlenecks to appear when we try scaling up.
- XrootD development:
  - Deletions – can we scale or do we need async?
  - Writable WN gateways
  - Containerized XRootD
  - Improving buffer layer in XrdCeph

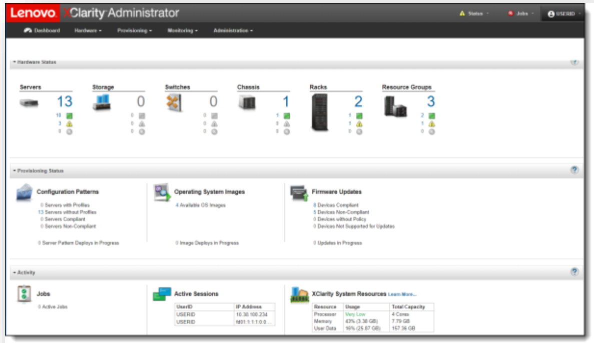
Hardware management

- Tier-1 Hardware currently from 3 Vendors.
- When purchasing hardware investing in better BMC software is worthwhile.
- Investigate setting up a common interface to all.
- Will allow us to understand the overhead of supporting multiple vendors.

DMTF Redfish?





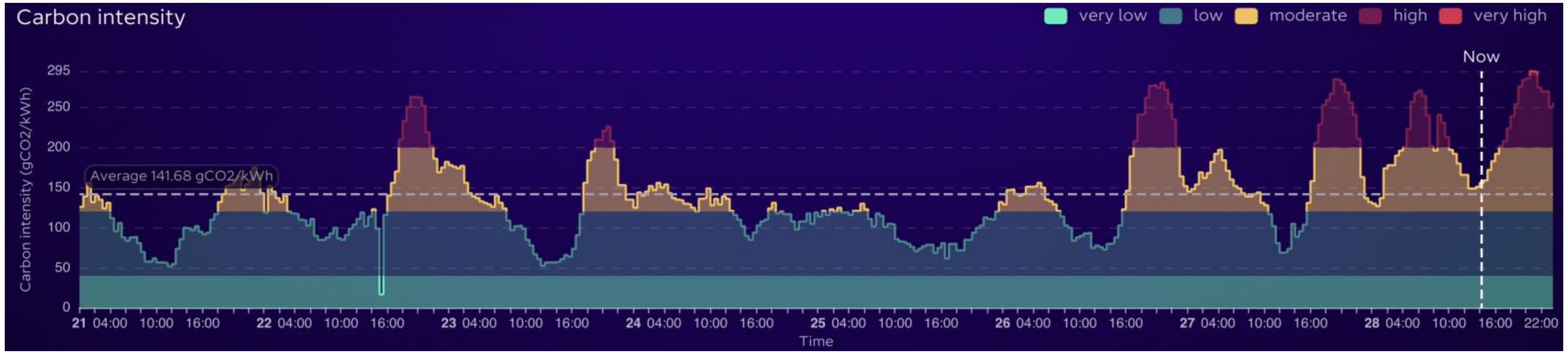


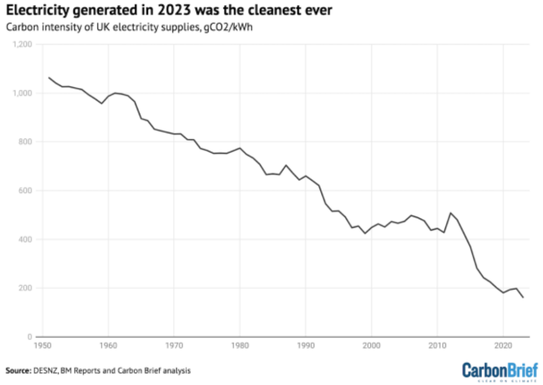
CVMFS Upgrade

- In October 2023 there was a serious incident when the physical hardware behind the Stratum-1 failed.
- In June 2024 there was a serious incident when the virtual infrastructure the Stratum-0/1 was built on wasn’t designed to cope with the workloads.
- Long downtime of Stratum-1 was declared and sites were moved to other Stratum-1s.
- We have ordered two new servers (should arrive this week).
  - It was identical to what Dave Dykstra ordered for FermiLab.
  - We are going to use ZFS to sync between the servers.
- Jose has tested the deployment on existing hardware.
- Aim to be back in production by the end of September.

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.

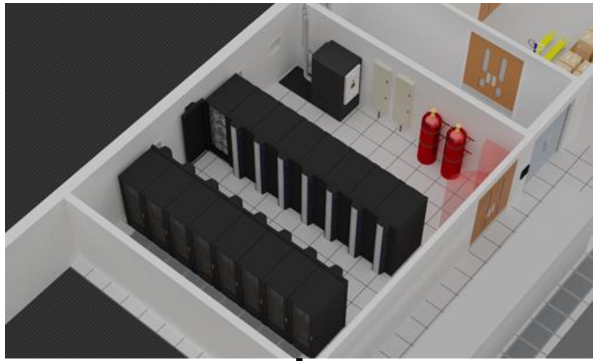




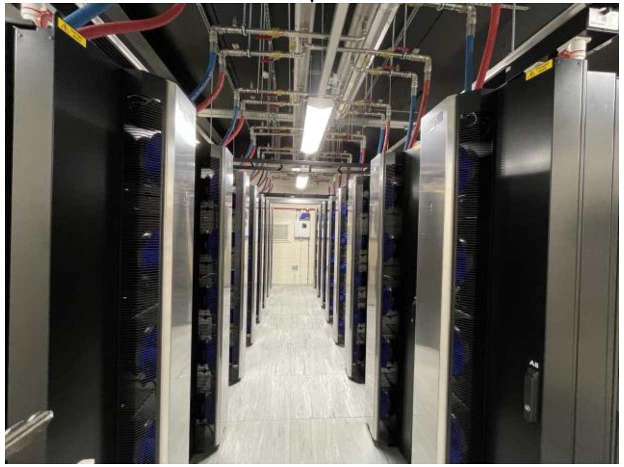
Source: DESNZ, BM Reports and Carbon Brief analysis

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

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

Science and
Technology
Facilities Council

Questions?