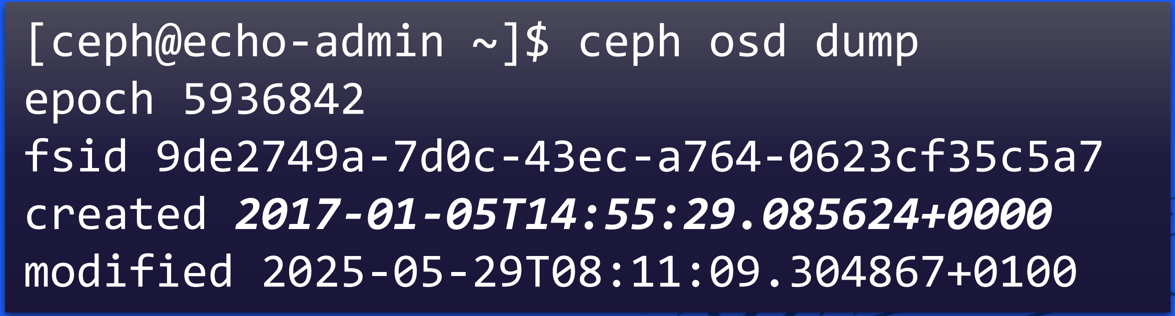
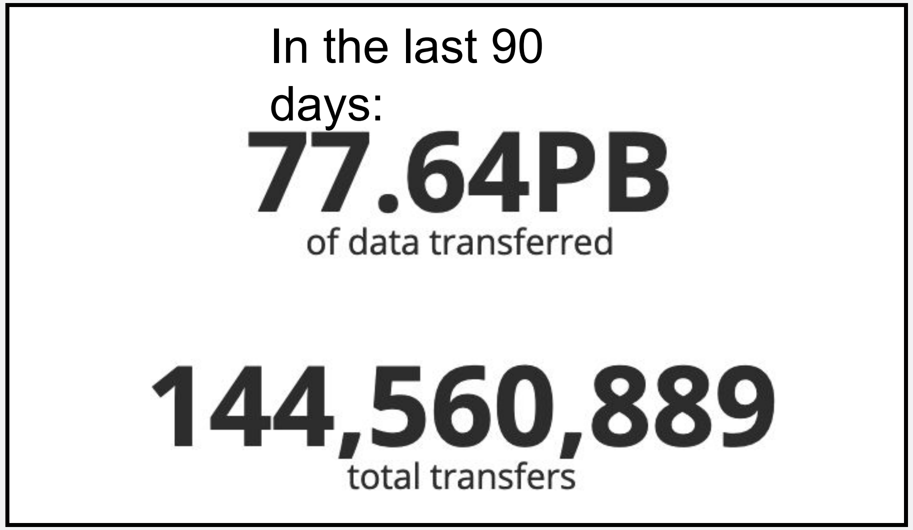


```
[ceph@echo-admin ~]$ ceph osd dump
epoch 5936842
fsid 9de2749a-7d0c-43ec-a764-0623cf35c5a7
created 2017-01-05T14:55:29.085624+0000
modified 2025-05-29T08:11:09.304867+0100
```

# Learning opportunities from a 100PB, 8-year-old Ceph cluster

Tom Byrne
Storage architect, Scientific Computing
UKRI – Science and Technology Facilities Council

# Echo – LHC computing grid storage



In the last 90 days:

**77.64PB**

of data transferred

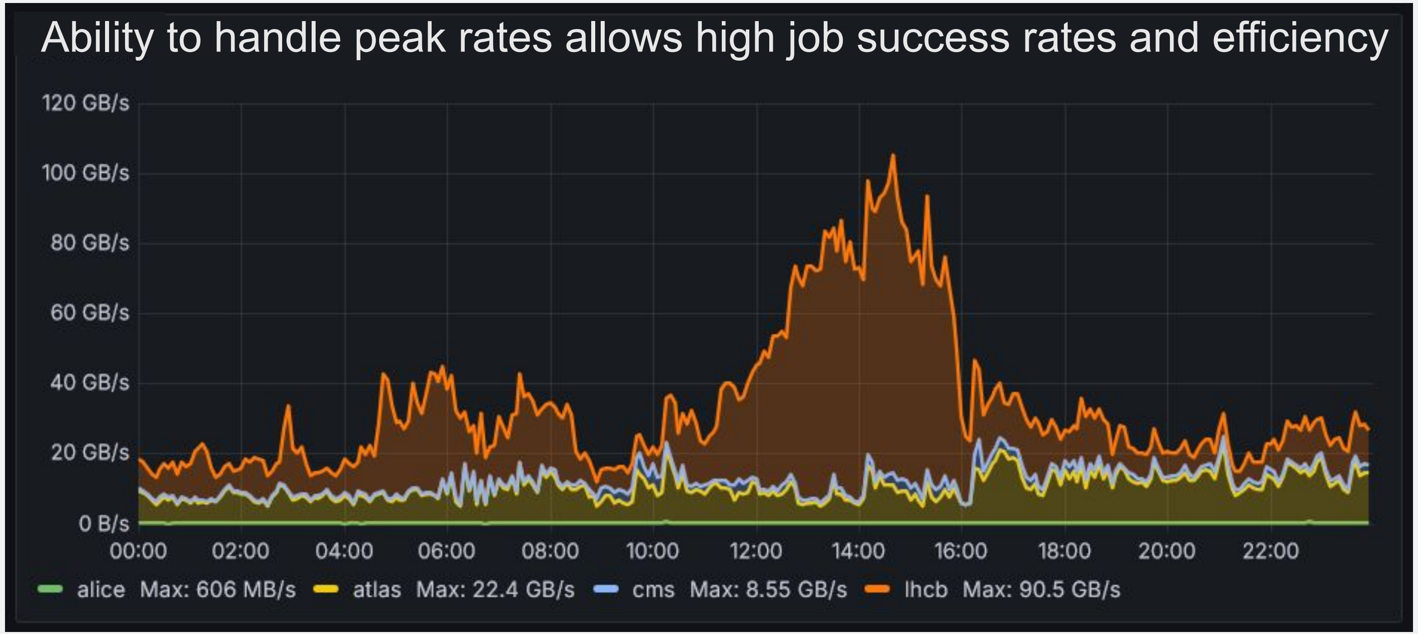
**144,560,889**

total transfers

Echo provides most of the UK's disk storage for the Large Hadron Collider experiments at CERN

Co-located with a 50k core HTC cluster, together they are used for collision simulation, event reconstruction and user analysis

- 300+ nodes, 6000+ OSDs, 110PB raw
- Originally Jewel, now Quincy
    - 5 major Ceph version upgrades!
- Data pools 8+3 EC
    - 70PB stored data, >20GB/s sustained transfer rates

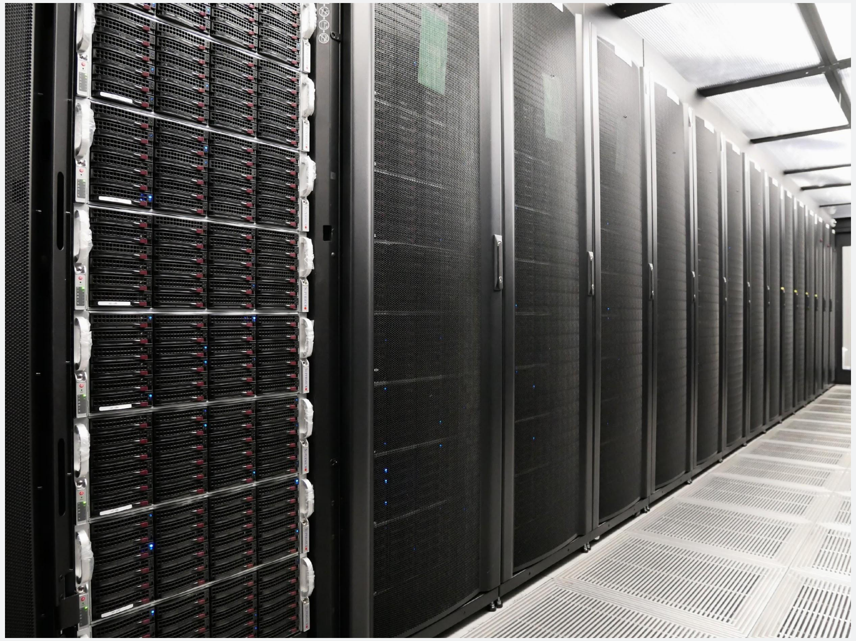


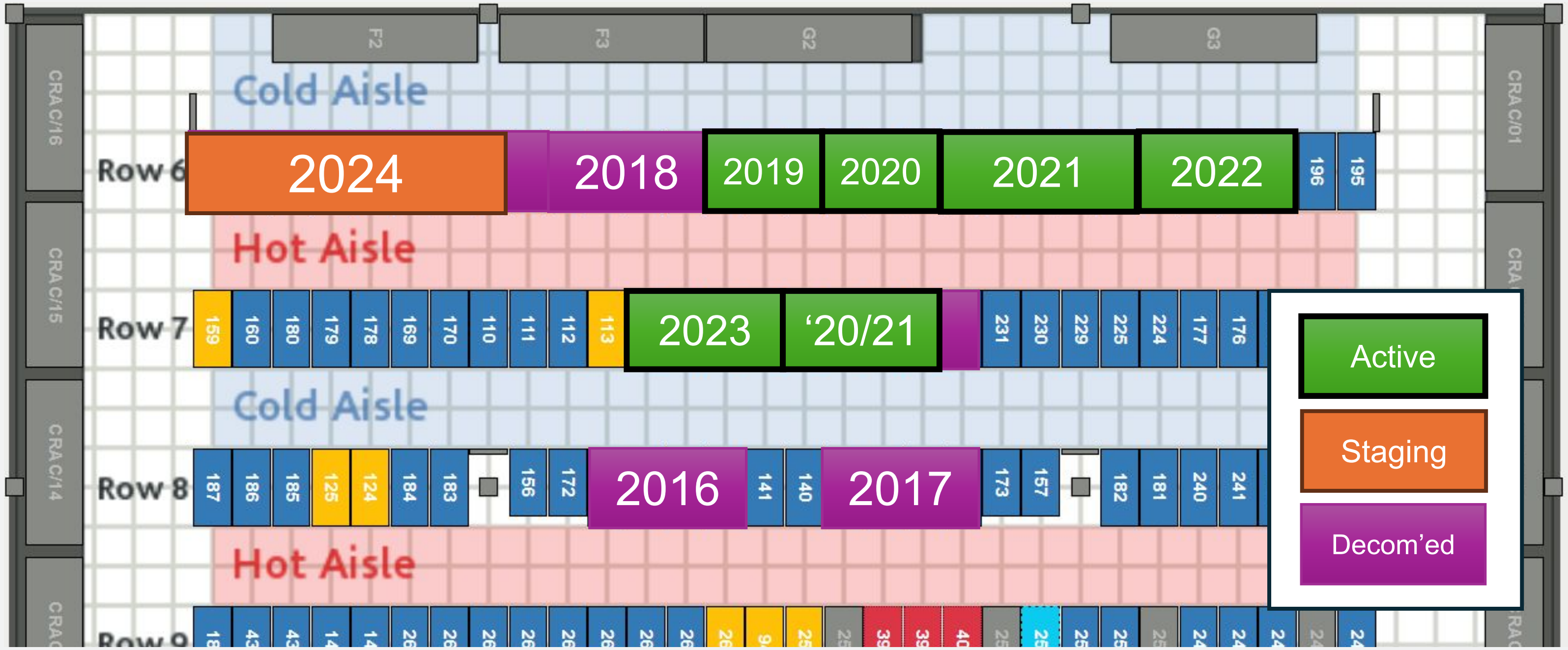
Ability to handle peak rates allows high job success rates and efficiency

Science and
Technology
Facilities Council

# Echo hardware

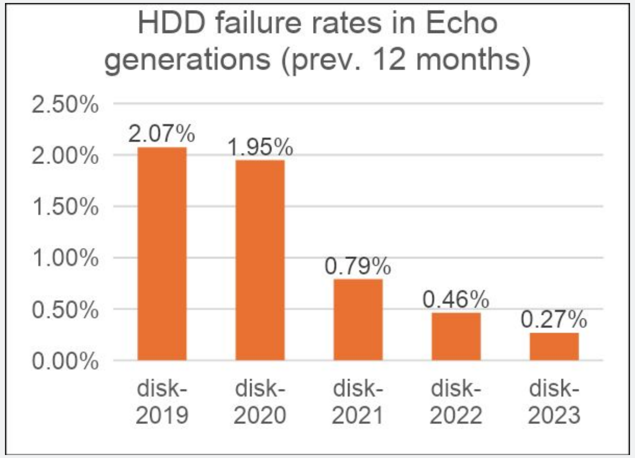
- **‘Simple, cheap, commodity’**
  - Performance is a by-product of capacity
- ~20PB of storage bought yearly
  - open tender exercise to ensure best value for money (within our constraints)





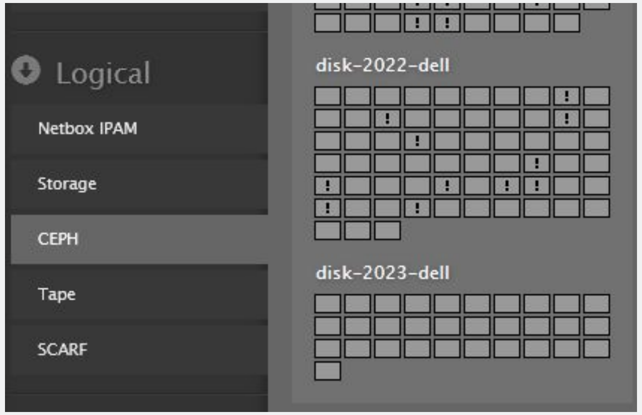
- 13 unique hardware generations
  - 5 generations in production
- Mostly 2U servers full of ‘big’ HDDs
  - 8TB in 2015, latest generation has 24TB HDDs

# General guidance for scale



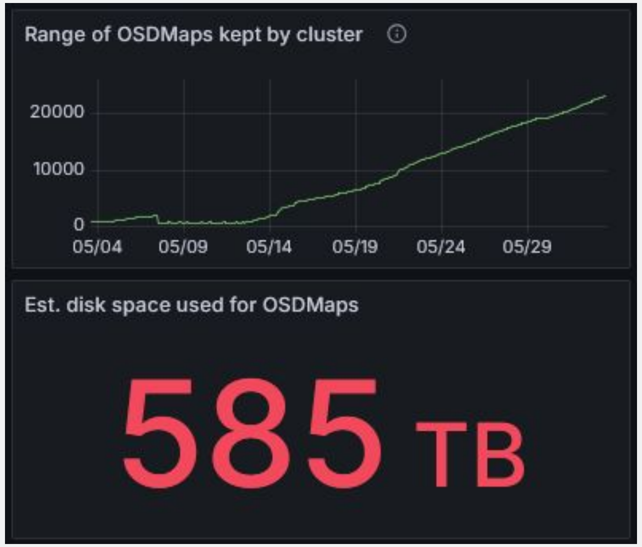
**1.** Disk hygiene is always important

- Stay on top of crashing OSDs: redeploy, replace, remove
- Don’t ignore oddities and transient warning states



**2.** Understand the expected cluster state and be able to easily identify inconsistencies

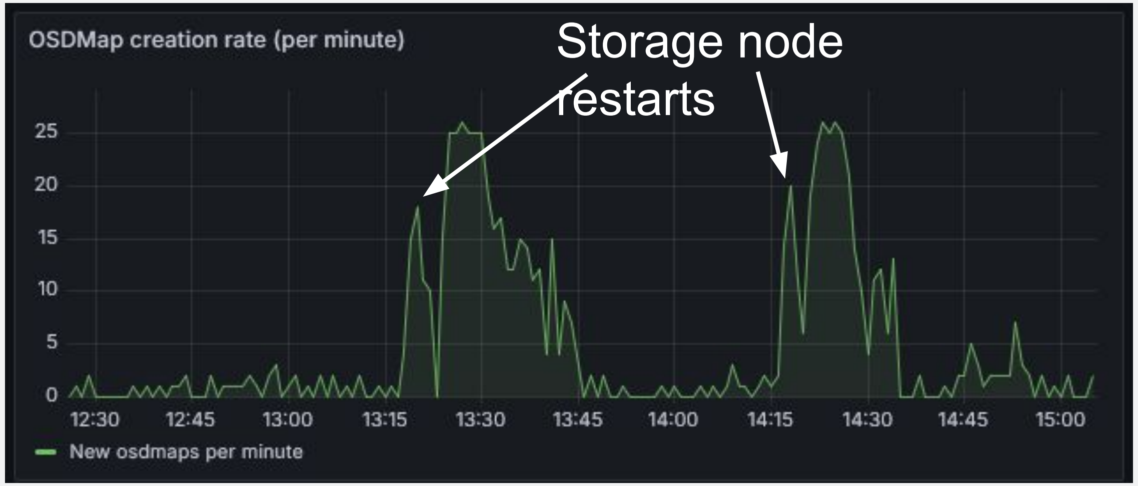
- How many OSDs should a host have? Which storage nodes should be in production?

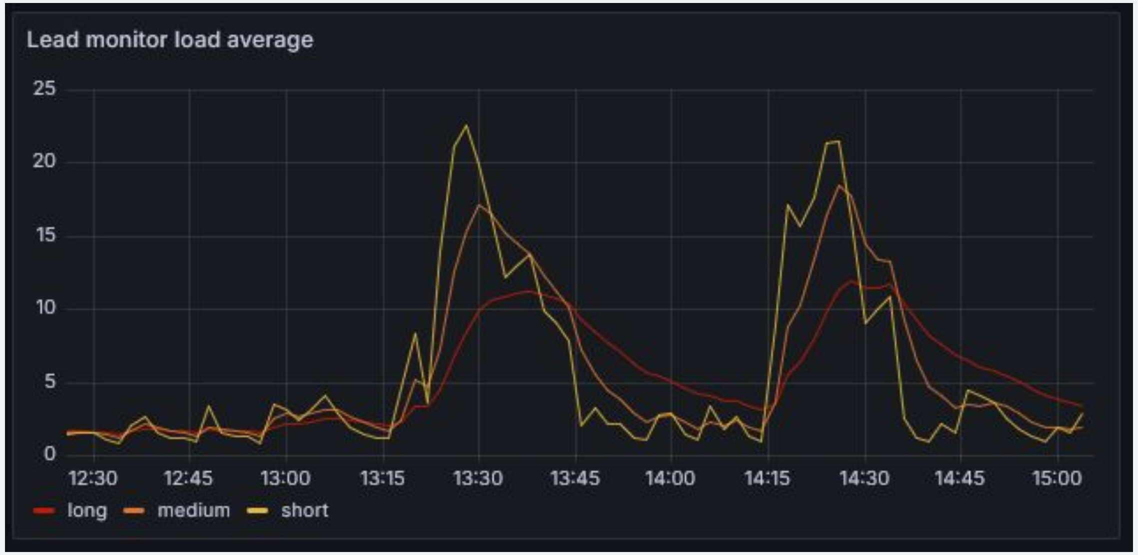

Range of OSDMaps kept by cluster
20000
10000
0
05/04
05/09
05/14
05/19
05/24
05/29
Est. disk space used for OSDMaps
585 TB

**3.** Monitor OSDmap churn rate and minimise where possible

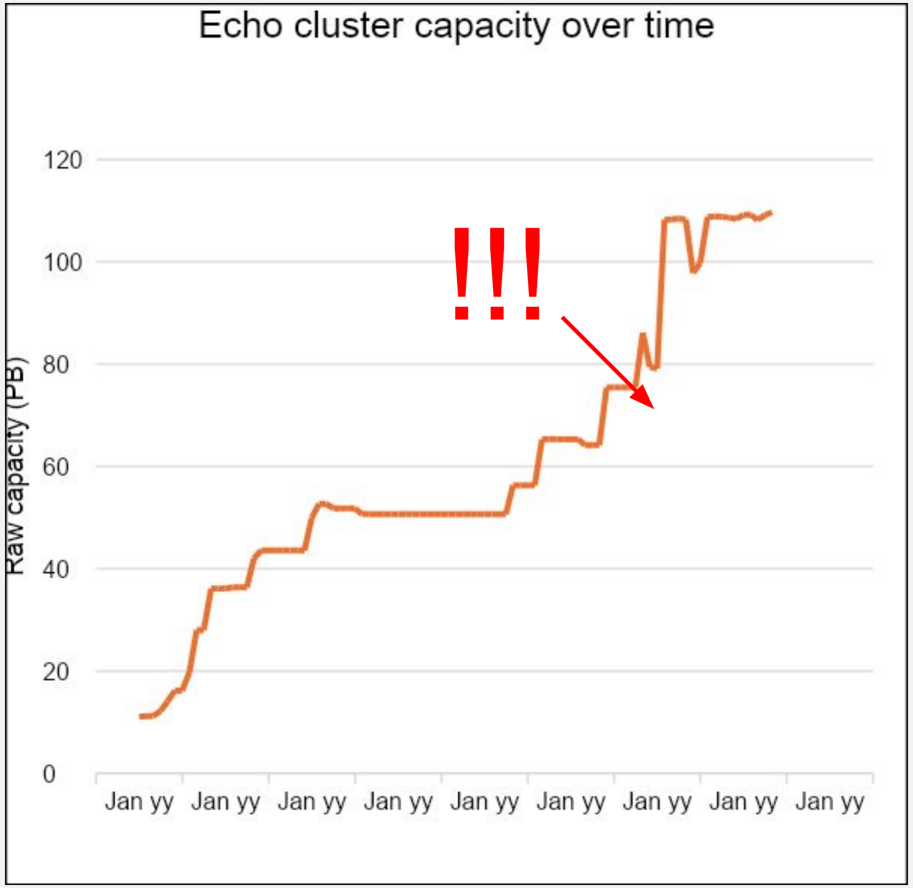
# Monitor load

- OSDmap creation places significant load on the lead monitor during periods of OSD state change
  - the >4MB OSDmap takes over 2 seconds to create (with `mon_cpu_threads = 50`)
  - main cause of 'operational sluggishness' of this cluster
- The monitor quorum duration (`mon_lease`) needs to be more than the time taken to create an OSDmap
  - OSDmap creation time varies by complexity (`pg_temps`, `pg_upmap_items`, etc), so make sure you leave headroom





# CRUSH



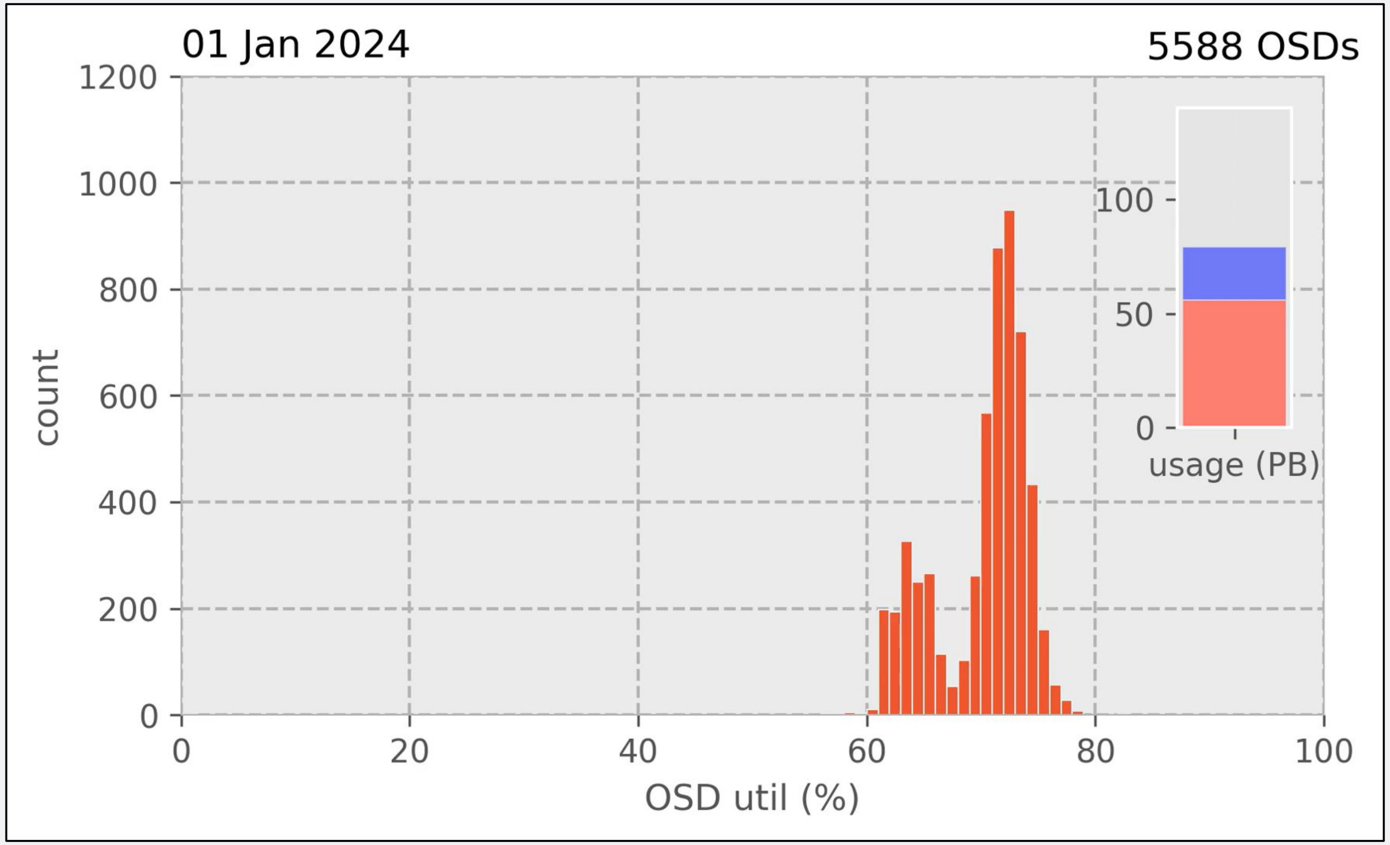
- The sum of weights in the crushmap can't exceed `max(uint16)` – 65535
  - At the default scaling of 1 = 1TiB, this is ~70PB of storage
  - Hitting this prevents any new OSDs being added: '`Numerical result out of range`'
- We rescaled Echo to 1 = 1 PiB
  - A straightforward (if slightly spooky) operation
  - Plenty of room for growth now ☺
- Note: things like `crush_update_on_start` assume the 1 = 1TiB scaling
  - We've resorted to `crush_initial_weight=0` to avoid excitement

Science and Technology Facilities Council

Nicely documented by retinadata: https://www.retinadata.com/blog/heaviness-of-large-ceph-clusters/

# Summary

- Ceph continues to provide a reliable and resilient storage layer to support LHC science in the UK
  - 8 years of largely continuous running
- Ceph generally scales well into the ~100PB range
  - Minimal tuning required
  - Standard cluster management practices continue to work as expected



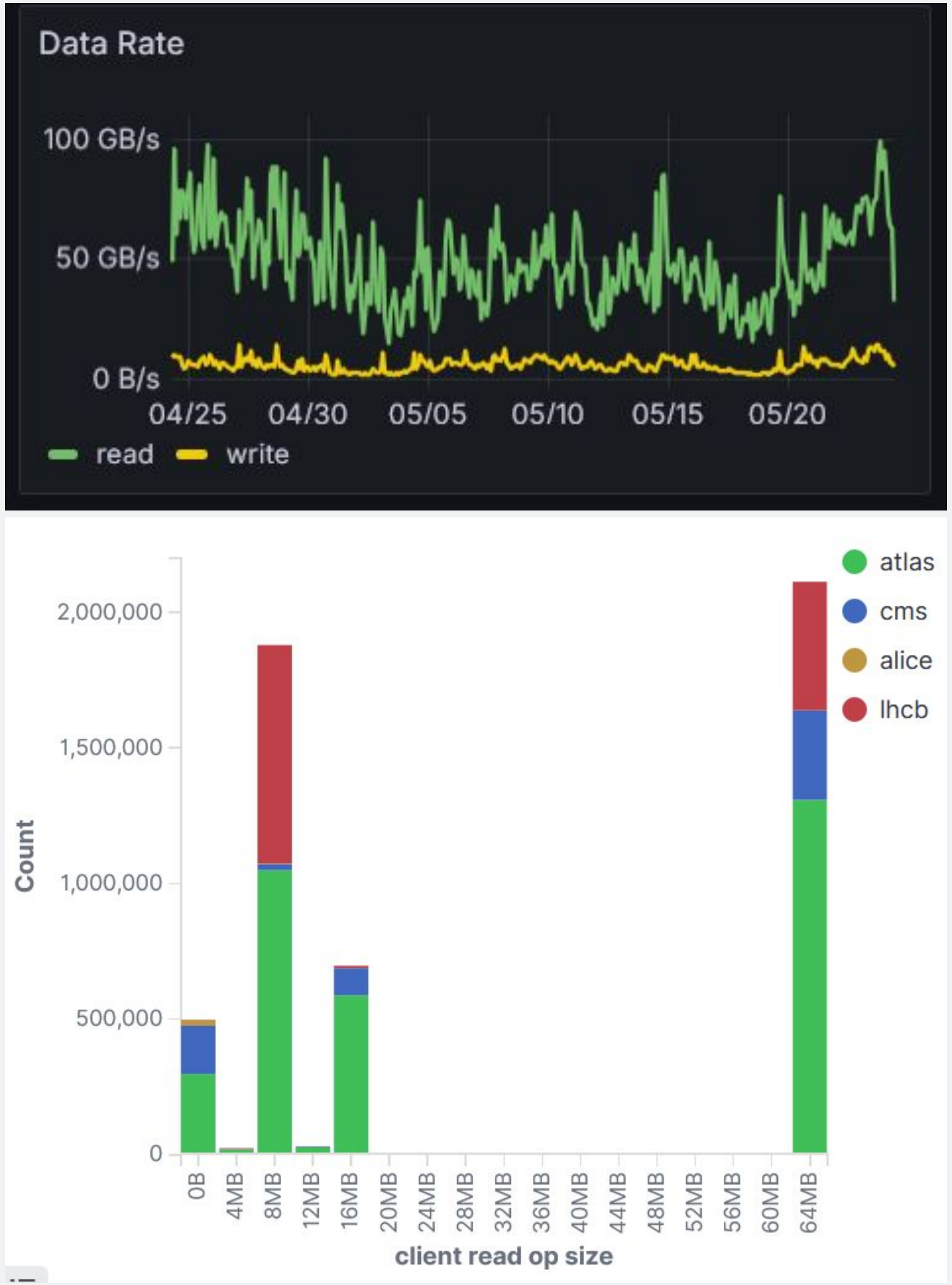
Animation of the 2022 generation addition, the 2018 generation removal and the 2023 generation addition

Science and
Technology
Facilities Council

# Questions?

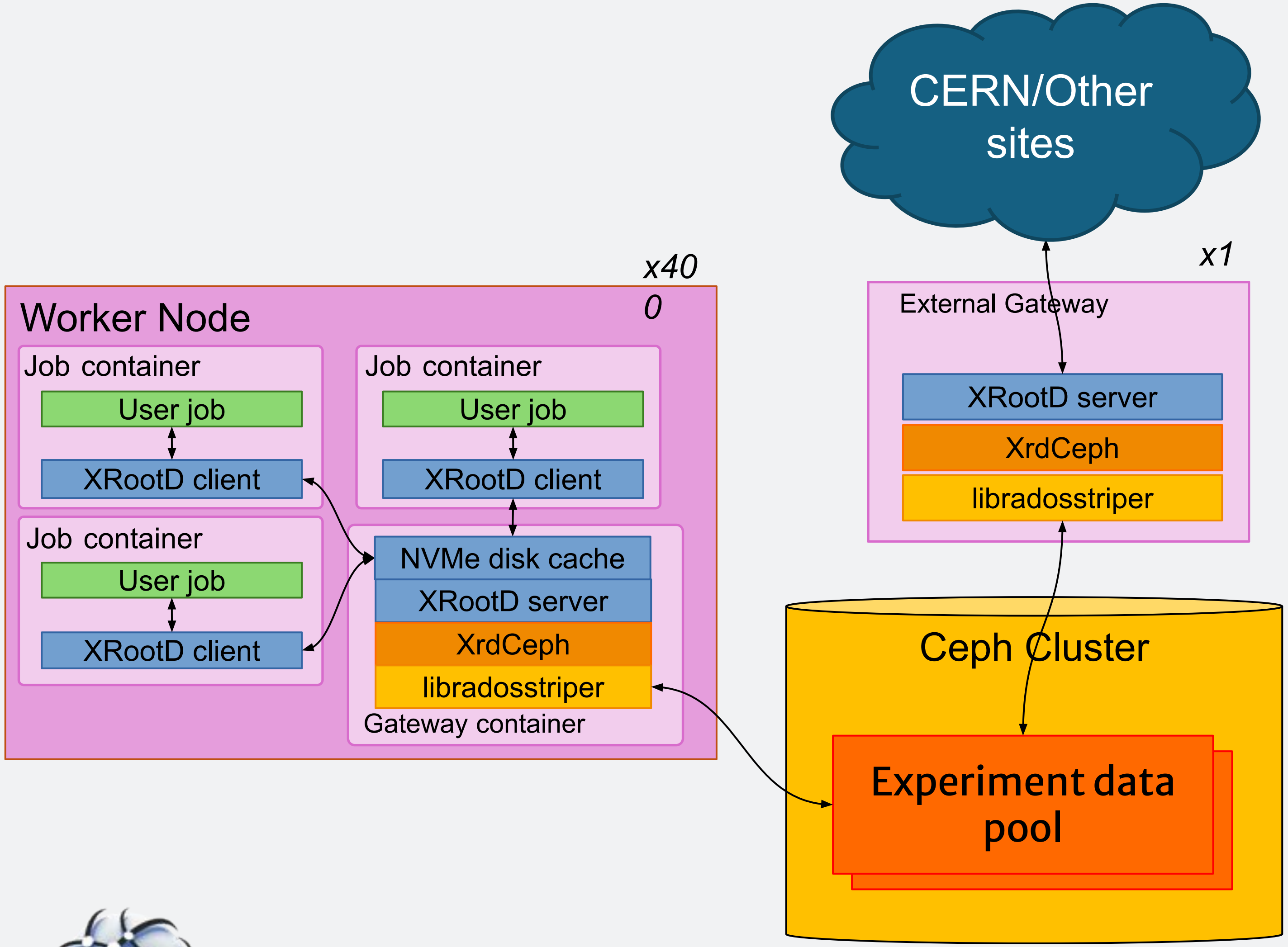
# Echo data access

- Data is accessed using **XRootD**, a data transfer framework developed for use by high energy physics experiments
- The “XrdCeph” plugin allows Ceph pools to act as the data storage backend for XRootD
  - XrdCeph uses **librados** (via **libradosstriper**) to read and write objects from the cluster
  - Filenames map directly to pool:object pairs, consciously limited FS operation support
- Distributed gateway stacks with NVMe disk caches
  - Control over read block sizes hitting the cluster via prefetching
- Almost no metadata load on the cluster



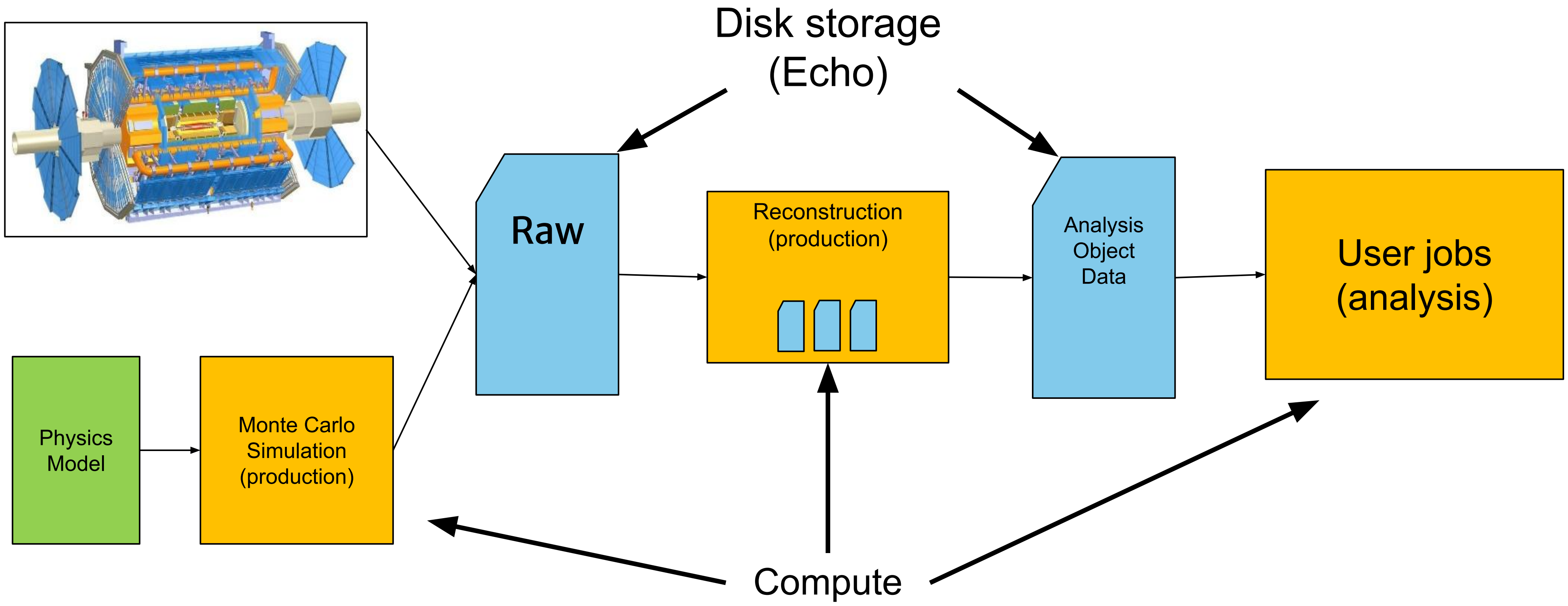
Cluster IO rates and sampled client read sizes, last 30 days

# Echo data access



XRootD

Science and Technology Facilities Council



Disk storage (Echo)

Raw

Reconstruction (production)

Analysis Object Data

User jobs (analysis)

Physics Model

Monte Carlo Simulation (production)

Compute