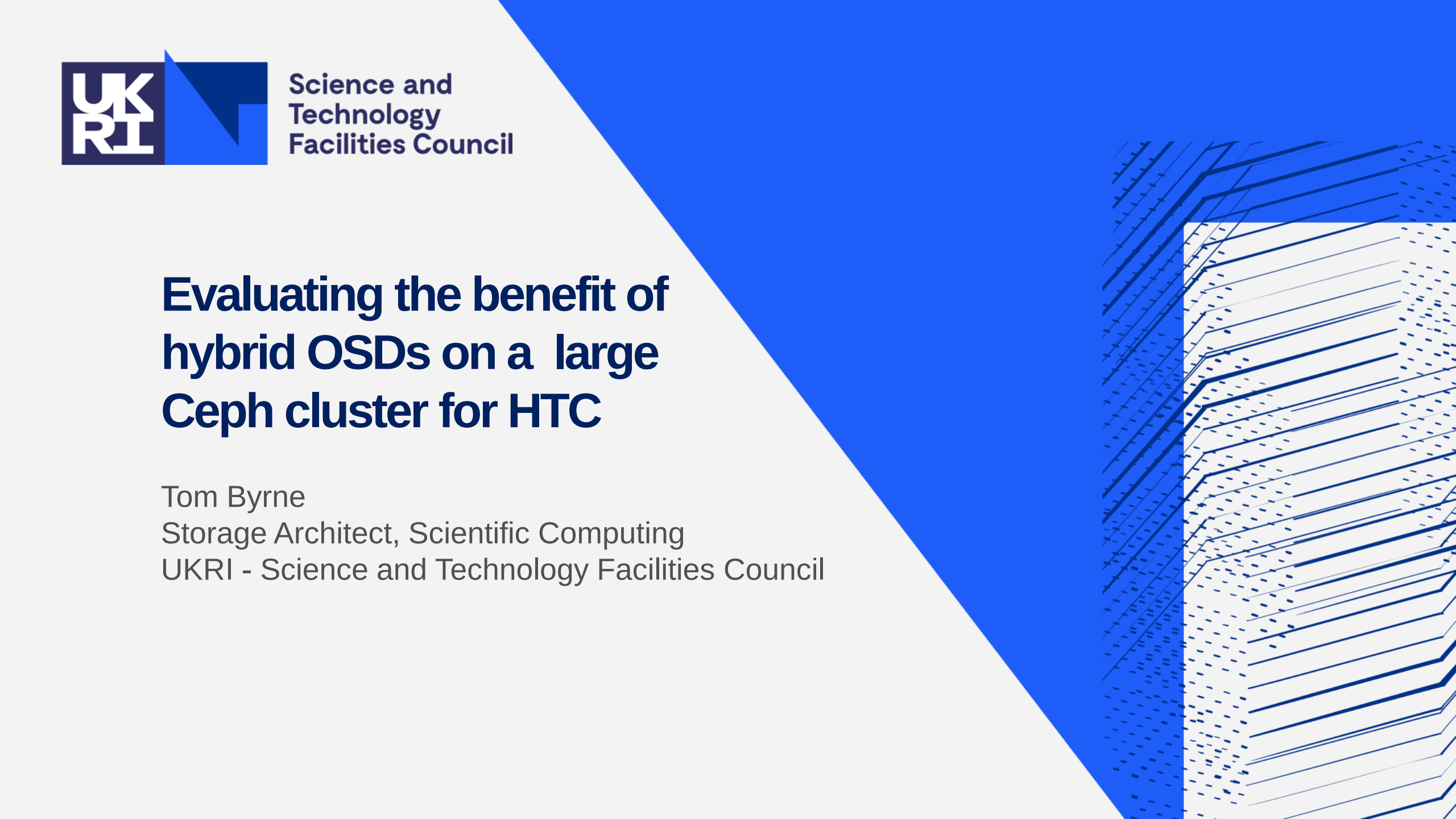
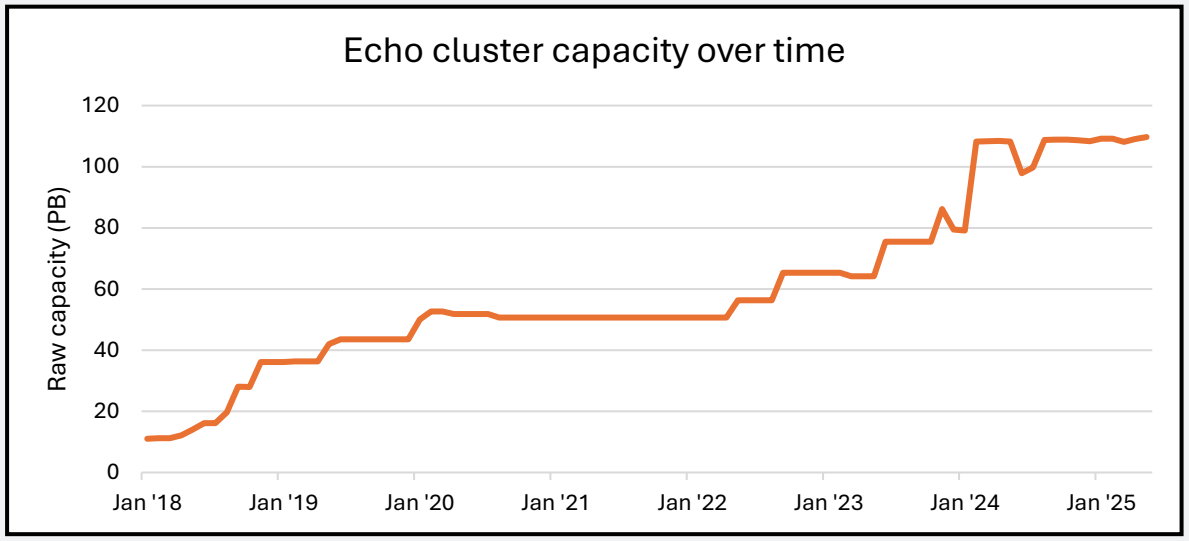


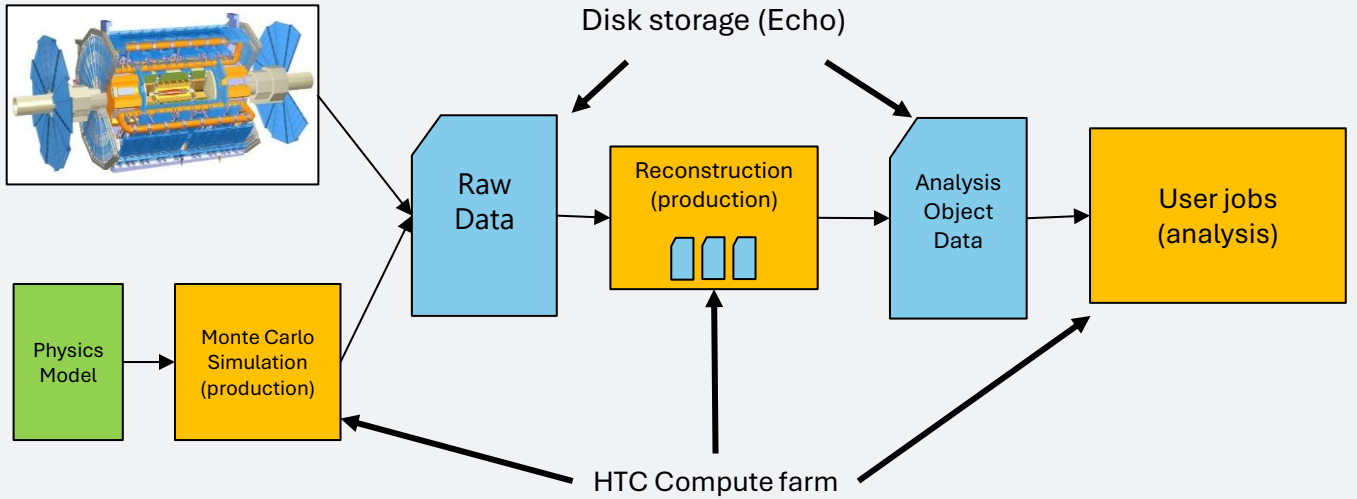
Evaluating the benefit of hybrid OSDs on a large Ceph cluster for HTC

Tom Byrne
Storage Architect, Scientific Computing
UKRI - Science and Technology Facilities Council

‘Echo’ – Ceph for LHC computing grid storage

- 300+ nodes, 6000+ OSDs, 110PB raw
  - Coming up to 9 years old, currently Quincy
- Data pools are all 8+3 EC
  - >70PB stored data
  - >20GB/s sustained transfer rates
- It provides the majority of the UK’s disk storage capacity for the Large Hadron Collider (LHC) experiments
  - located close to 50k cores of compute, used by the LHC experiments for bulk processing operations and user analysis

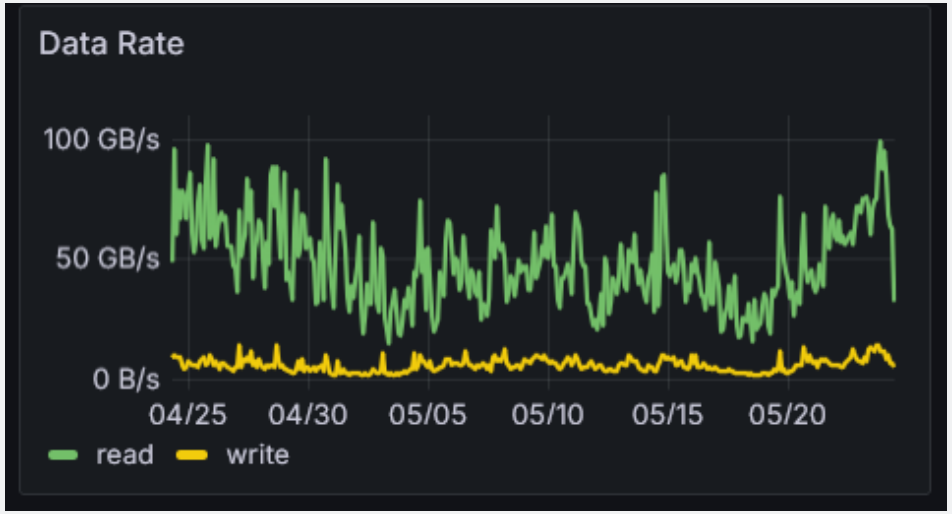


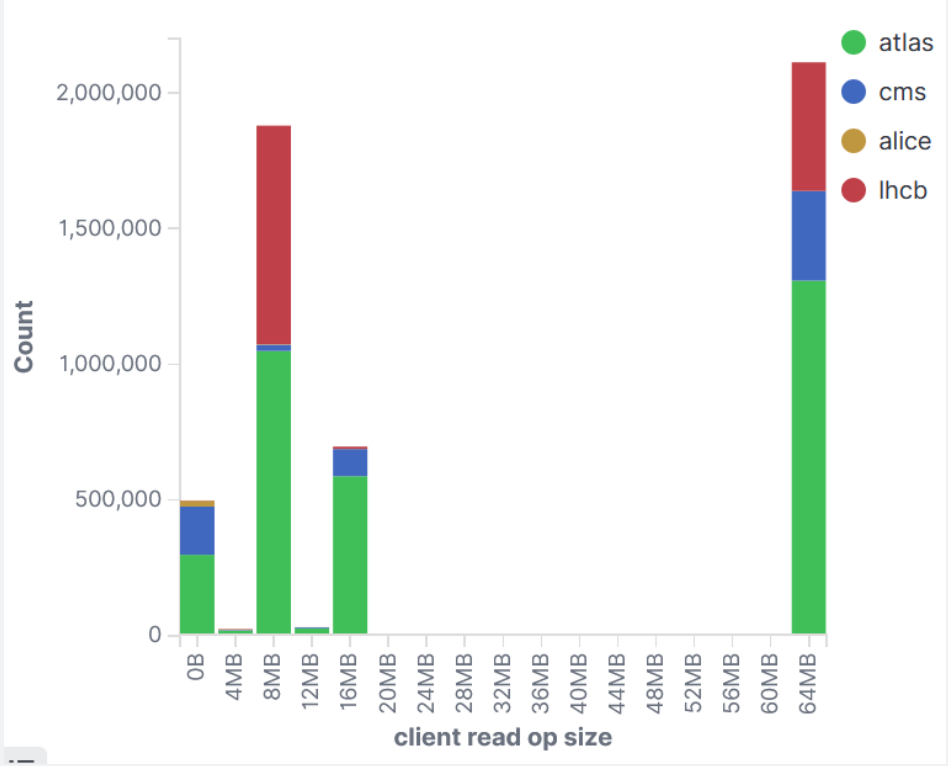


Science and Technology Facilities Council

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





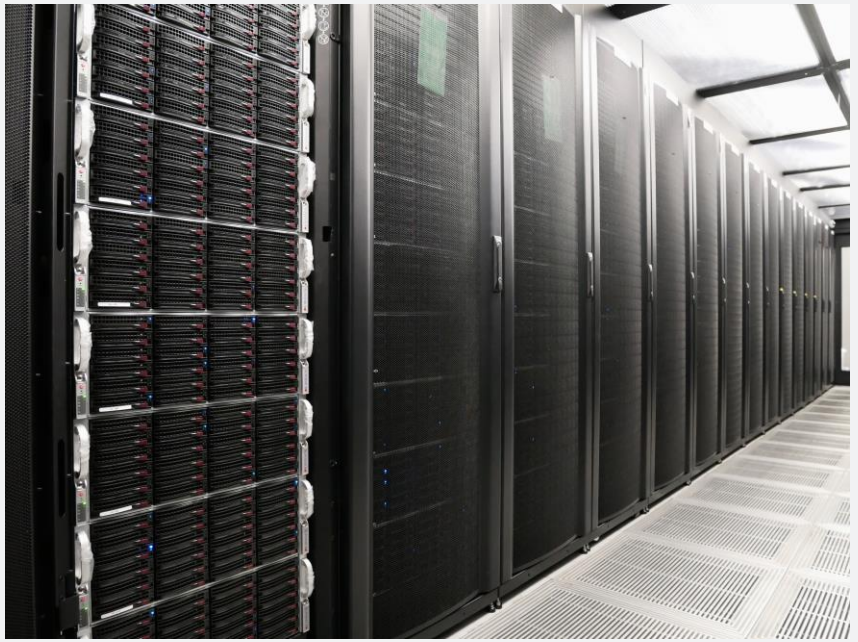
Cluster IO rates and sampled client read sizes, last 30 days

Science and Technology Facilities Council

XRootD

Echo hardware specification

- “Cheap, simple capacity”
  - Mostly 2U servers with ~24 ‘big’ HDDs - 8TB in 2015, rising to 24TB this year
  - ~20PB of storage bought every year – open tender exercise to ensure best value for money (within our constraints)
- No budget for flash for OSD WAL/DBs in the early days
  - Generally, this hasn’t been an issue for performance
  - Now 9 generations later with 5 generations of hardware in production, no desire to change *without data to back it up*
- A generation of Echo hardware was bought with a small amount of flash for R+D purposes
  - Deployed as a separate SSD pool, but…



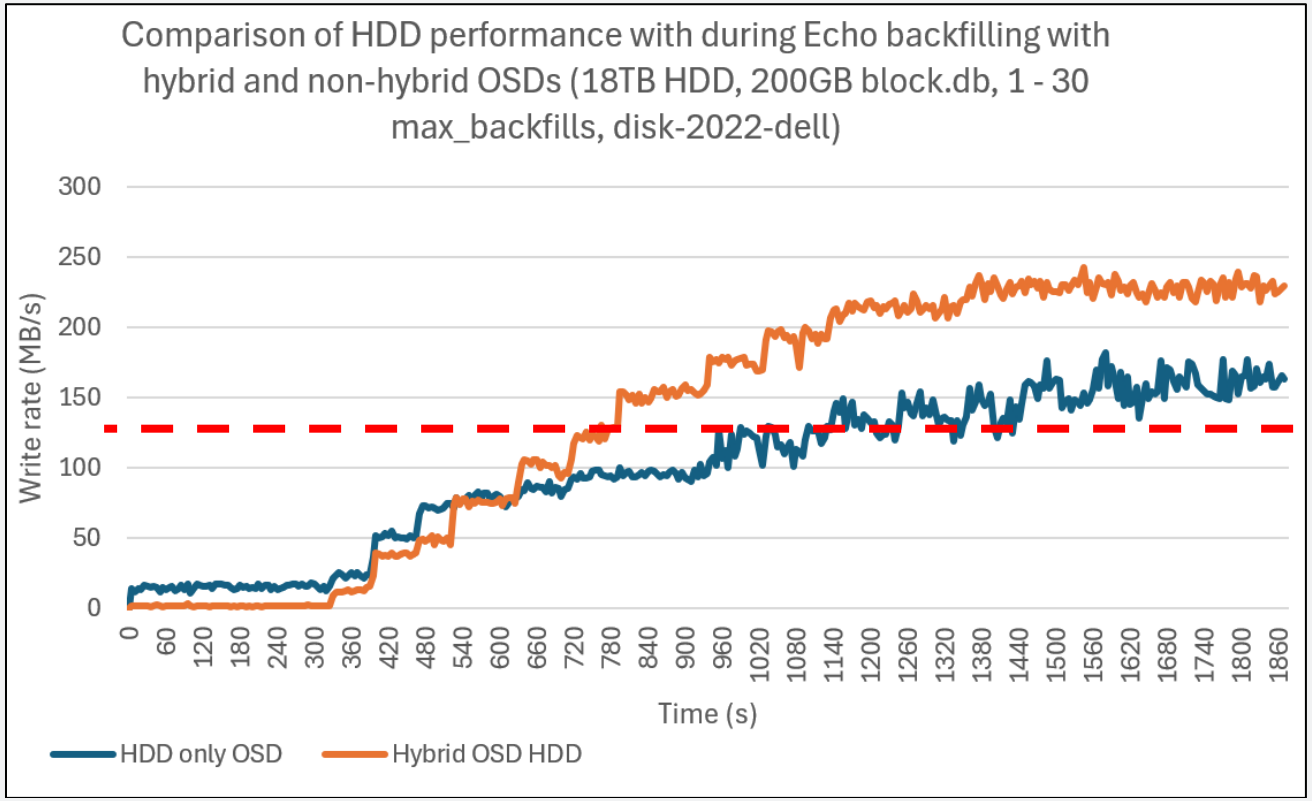
Would this cluster benefit from hybrid OSDs?

I (briefly) evaluated the need for flash for OSD DB/WAL before last years procurement.

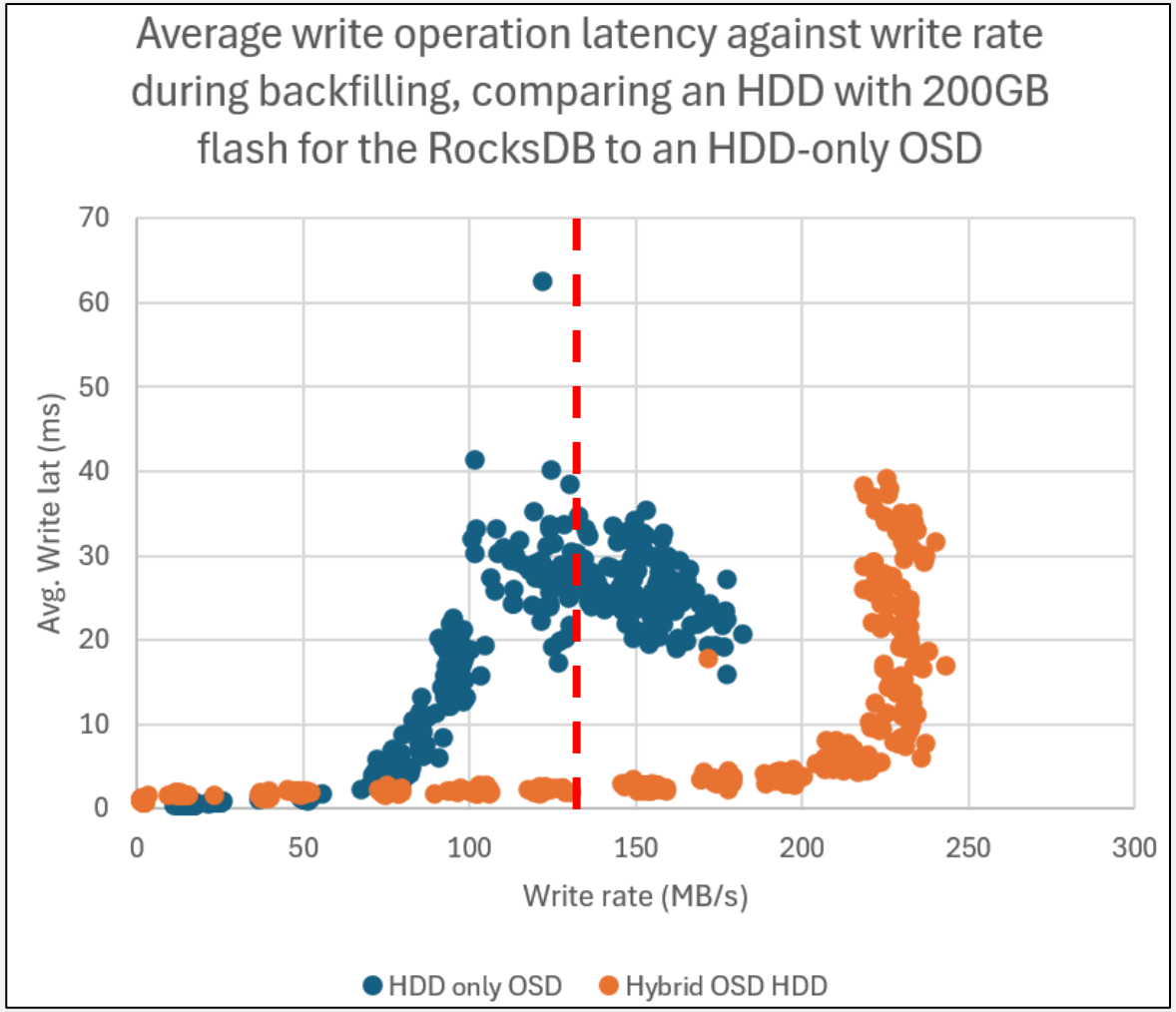
Method:

1. Redeploy selected production hosts with hybrid OSDs
2. Observe changes in load on the HDDs from real world Echo workloads
3. Quantify increase in performance, or potential performance headroom of the cluster

Backfill performance

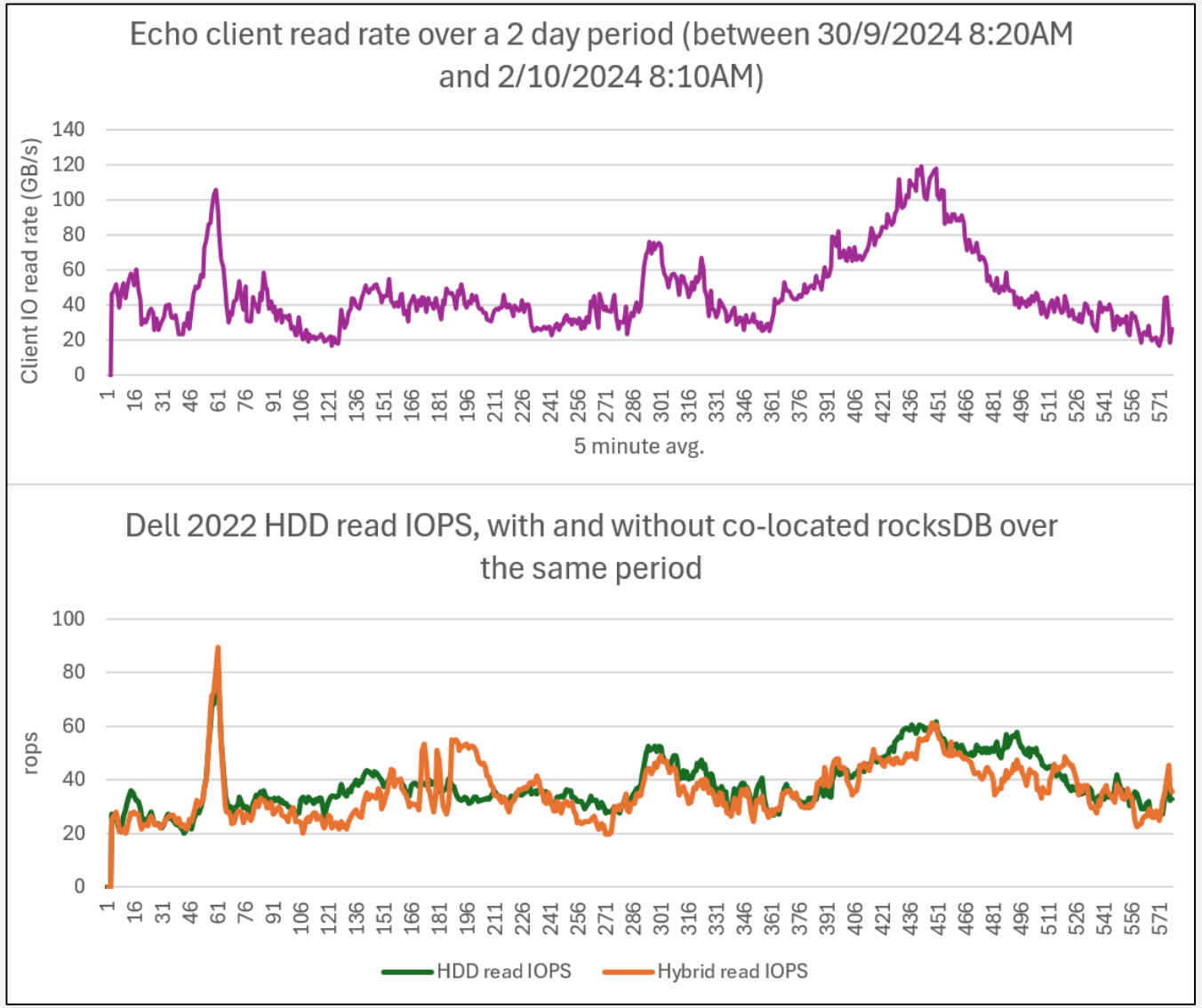


- Redeploying individual hybrid and co-located OSDs, ramping up backfill rates and observing HDD load
- Noticeable improvement in backfill performance for individual OSDs
  - Getting close to the advertised write rate of the drives
- However, 25Gig networking means only ~130MB/s available per OSD when backfilling a whole host



Normal cluster load comparison

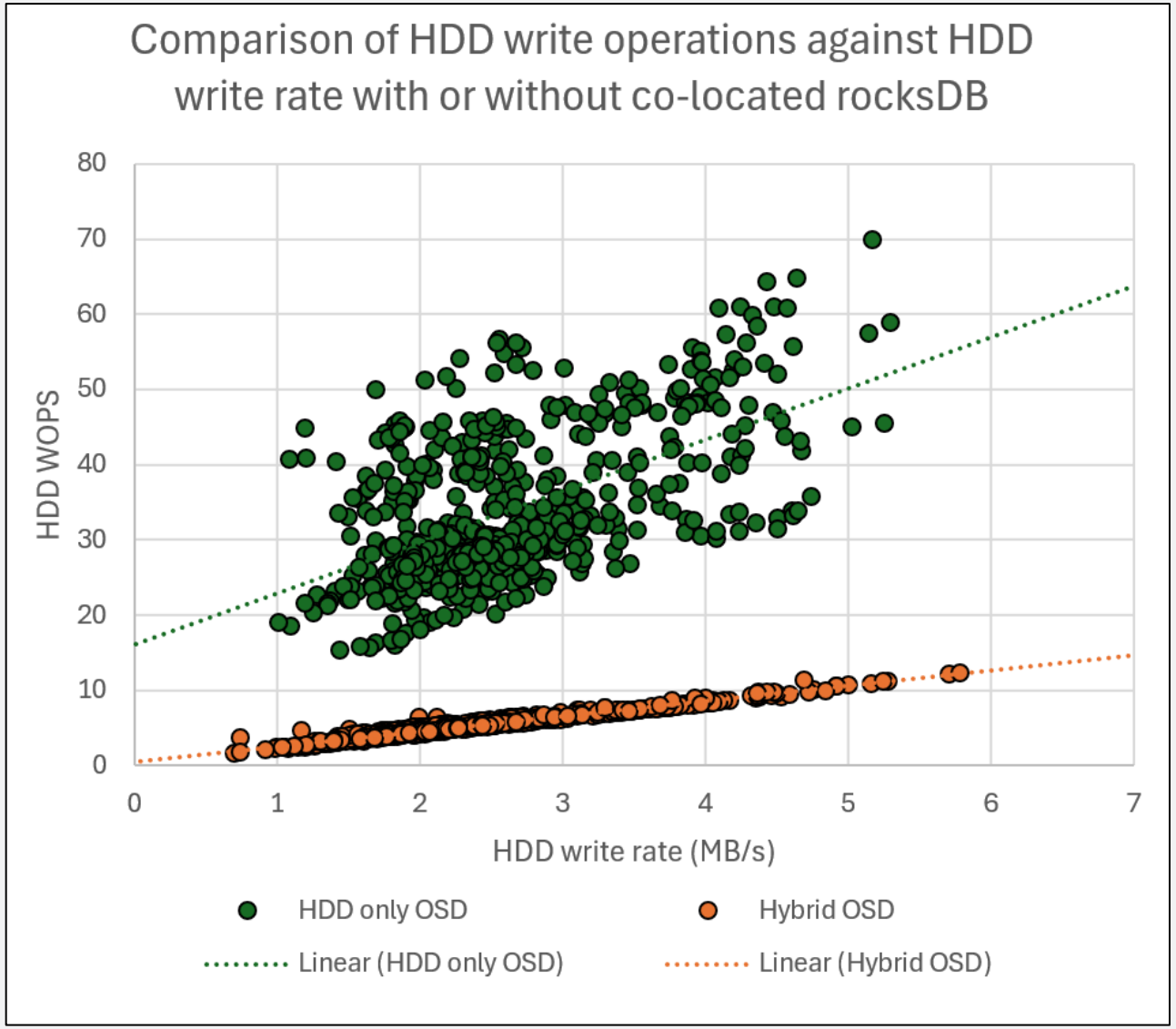
- Redeploying whole ‘dell-2022’ storage nodes with hybrid OSDs
- Evaluate load on HDDs compared to the rest of the generation (with co-located OSD)
- Comparing ‘device IO’ to ‘cluster IO’ to predict cluster limits in both OSD deployment scenarios
- For brevity the analysis focused on averaged data – both by time (5m means) and device metrics
  - We are probably missing hotspots, both at the device level and temporal hotspots

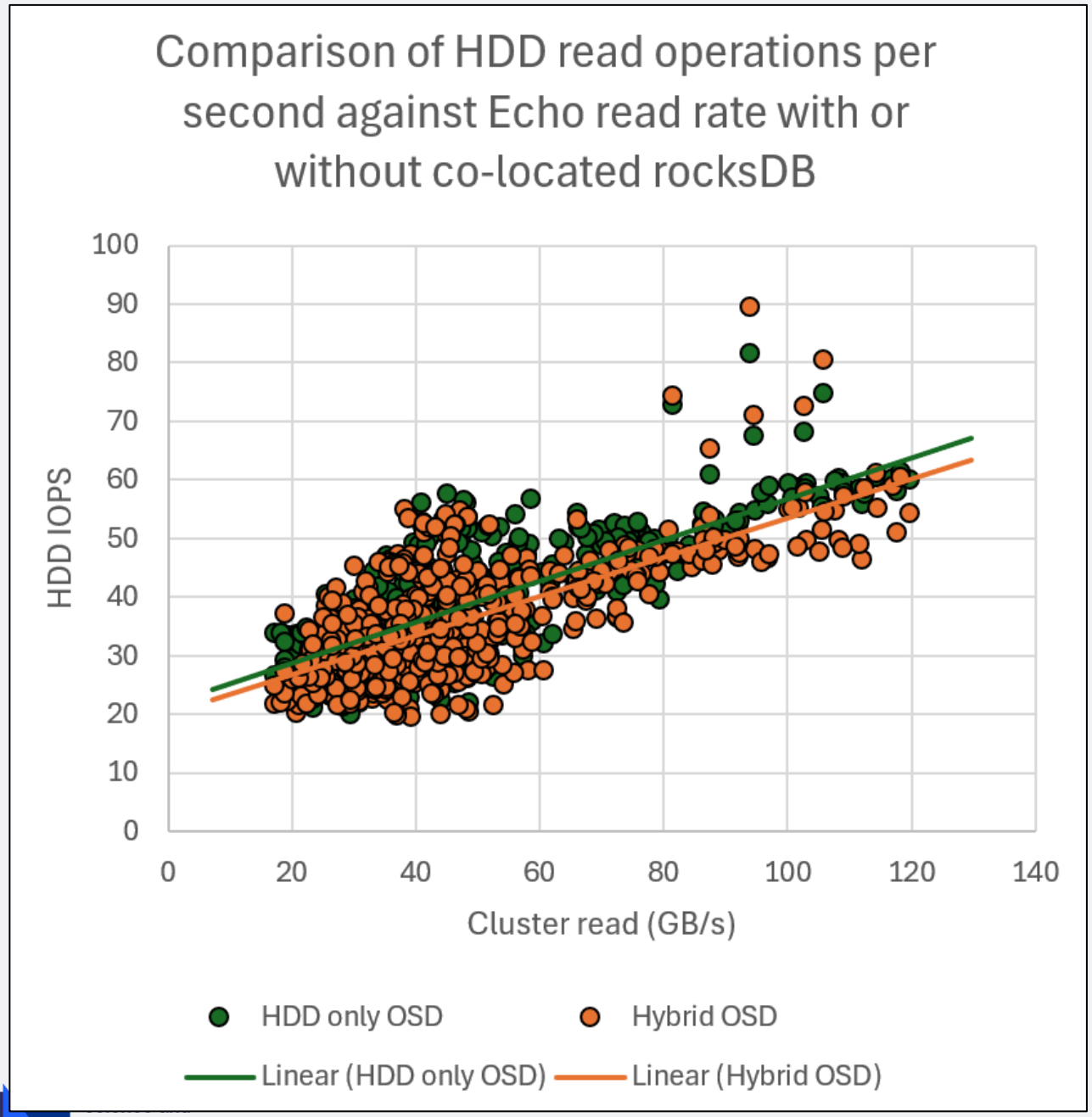


Science and Technology Facilities Council

Write operations

- Clear difference seen in amount (and volatility) of write operations hitting the HDDs.
- Most DC HDDs have some persistent cache for writes, and are surprisingly capable at small random write operations
  - 550 write ops advertised (for these drives)
- Lots of headroom under observed conditions, but co-located OSD would appear to saturate write operations @80MB/s, below the networking limit of the nodes.





Read operations

- No meaningful difference observed in HDD read ops when removing rocksDB from HDD
- Read operation latencies also basically identical in both scenarios
- This indicates very little RocksDB reads hitting the underlying block device.
  - Is the metadata required for this workload so light that everything fits in the OSD memory cache?

Conclusions and thoughts

- Co-located rocksDBs on HDDs appear to provide adequate performance for the cluster IO patterns observed *during this analysis*
  - Separating the block.db had no meaningful impact on reducing read operations on HDDs, which are our main limiting factor
- Continuing to focus flash usage on application-level caching and/or separate flash pools for small IO appears to be the best strategy here
- HDD random read operations are a precious commodity, and one that is becoming scarcer.
  - Less than 10 rop/s per raw TB, significantly less with EC pools
  - We may have to revisit this strategy if we keep going for larger hard drives