Science and Technology Facilities Council

Cephalocon 2025 highlights

Tom Byrne
Scientific computing storage architect
Rutherford Appleton Laboratory, UK

What is it?

- 2-day, 3-track conference on all things Ceph
  - This year hosted in Vancouver, Canada
- User stories, development updates, BoF sessions
- Diverse user groups
- Lively evening discussions

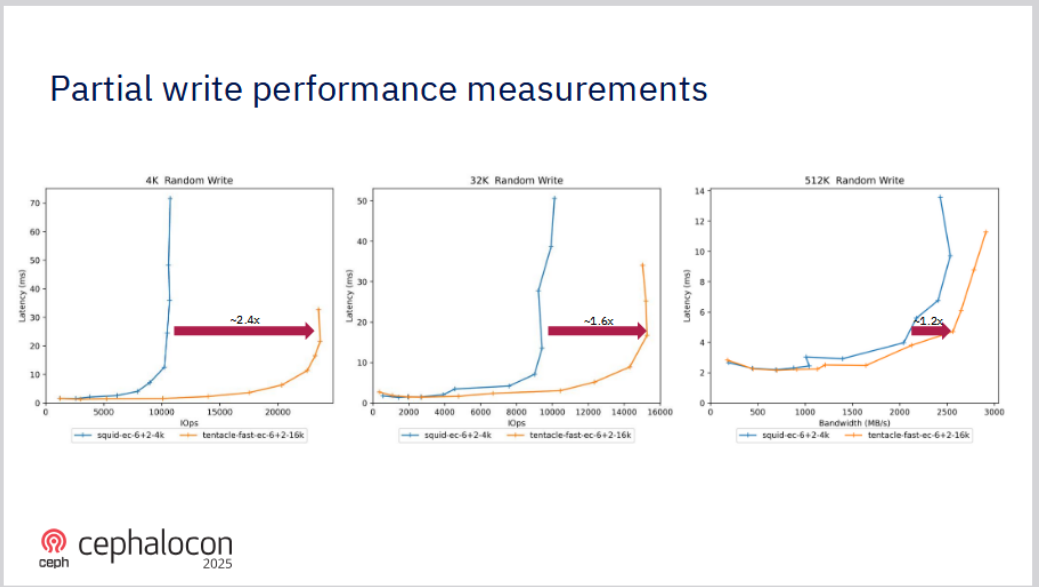
Science and Technology Facilities Council

Core Ceph talk and updates

Science and Technology Facilities Council

Fast EC

- Several EC improvements have landed in Tentacle
  - Partial read optimisations
    - Sub-stripe access
  - Better small object overheads
    - Store sub-stripe object as m+1 replicas
  - Efficient partial stripe overwrites
    - Parity delta calculated and applied rather than whole stripe calc
- Well tested but slow legacy behaviour still default
  - Enable on a per pool basis

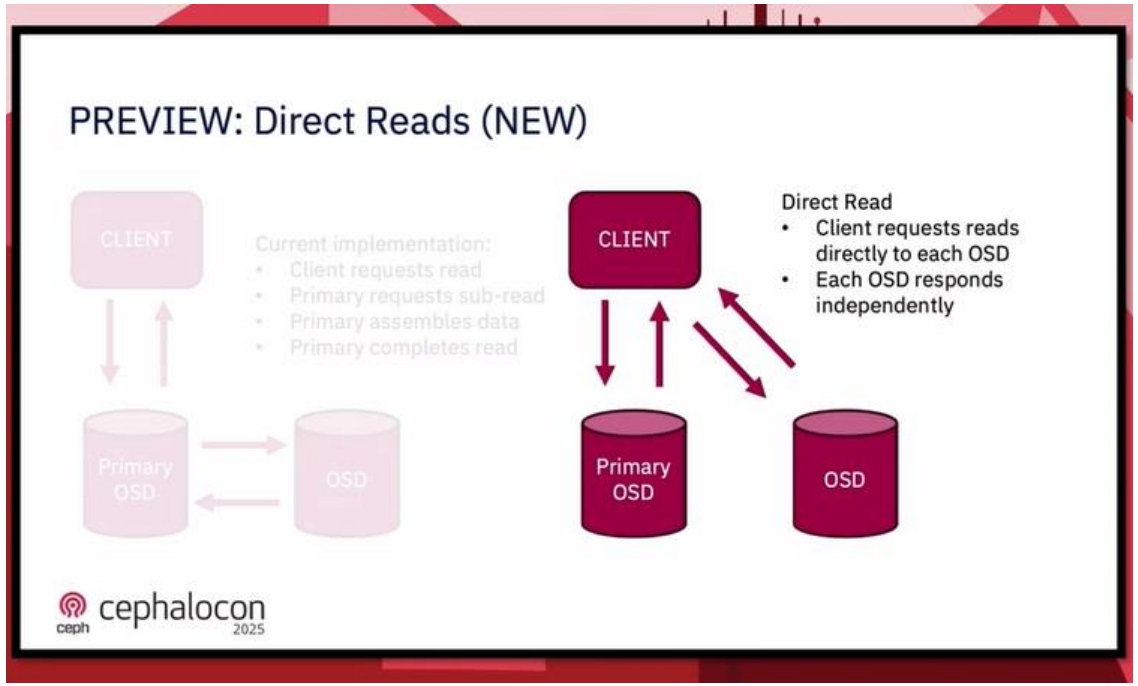


https://cephalocon2025.sched.com/event/27f2C/inking-out-inefficiencies-in-ceph-erasure-coding-alex-ainscow-bill-scales-ibm

Science and Technology Facilities Council

Fast EC

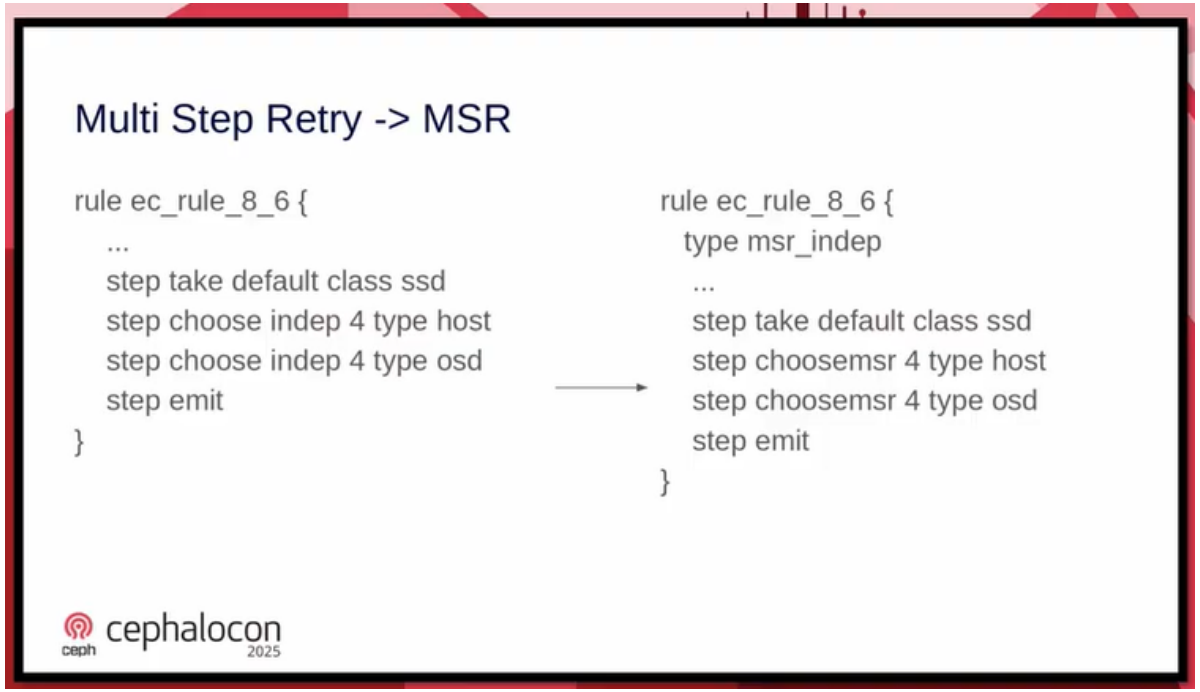
- Future optimisations under investigation include:
- Direct read support
  - Remove latency of extra hop and network overhead
- Small object packing
  - Pack small (smaller than stripe width) objects into a single EC object to improve storage efficiency and recovery rates



https://cephalocon2025.sched.com/event/27f2C/inking-out-inefficiencies-in-ceph-erasure-coding-alex-ainscow-bill-scales-ibm

Science and Technology Facilities Council

CRUSH Multi-Step Retry



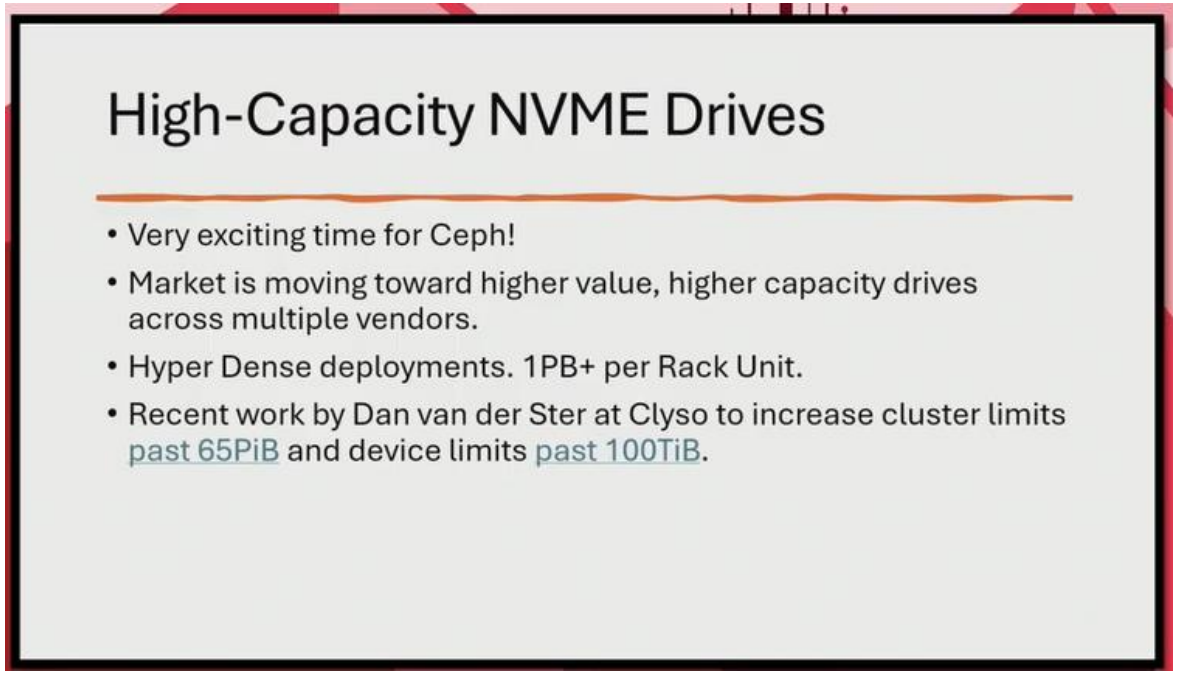
- A great talk about the new ‘depth-first’ CRUSH rule behaviour
  - Better support for wide EC codes over small clusters
  - e.g. 8+3 with <=2 PGs per host
- Generalises the ‘good’ chooseleaf behaviour for better data redistribution during OSD/Host/Rack unavailability

https://cephalocon2025.sched.com/event/27f2R/msrmutli-step-retry-an-generalization-of-crush-allowing-multiple-osds-per-failure-domain-sam-just-ibm

Science and Technology Facilities Council

Dense Flash and large cluster support

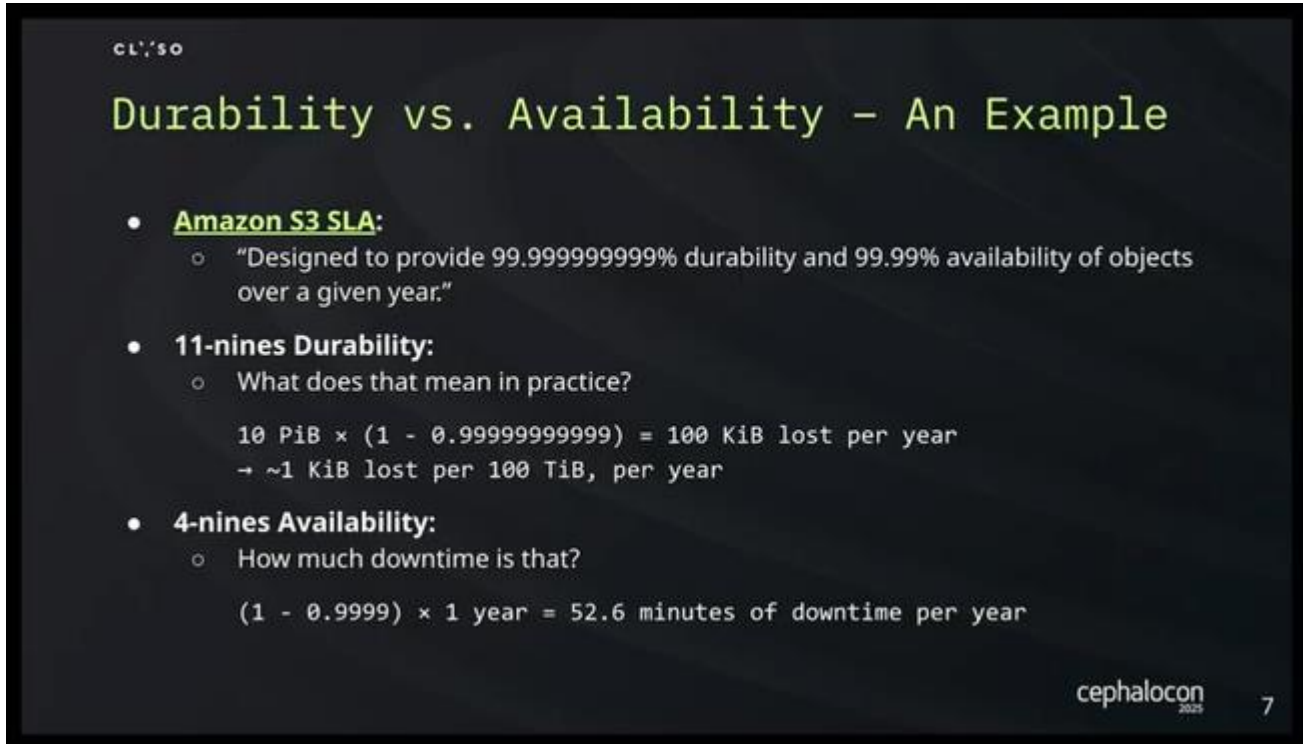
- Large NVMe's typically have larger Indirection Units to reduce DRAM requirements
- Support for automatic setting of bluestore min_alloc size to the underlying device IU to prevent excessive write amplification
- Various improvements for support of block devices >100TB and clusters >65PiB



https://cephalocon2025.sched.com/event/27f3A/demystifying-hyper-dense-ceph-how-to-design-for-large-capacity-nvme-alessandro-goncalves-solidigm-mark-nelson-clyso

Science and Technology Facilities Council

Ceph pool durability



- Large storage devices and widening erasure code layouts could be changing the durability of our clusters unexpectedly
- New CLI calculator to simulate *mean-time-to-data-loss* on your cluster's pools using MC simulations of PG mappings and OSD failures

https://cephalocon2025.sched.com/event/27f34/ceph-durability-how-safe-is-my-data-dan-van-der-ster-clyso

Crimson update

- Complete OSD rewrite for the NVMe era aimed at improving IOPS/cycle and removing lock contention where possible
- Steady progress with promising results but still early days
  - Expected to support a mixed cluster mode i.e. filestore -> bluestore transition

https://cephalocon2025.sched.com/event/27f4Z/re-architecting-ceph-one-crimson-osd-at-a-time-matan-breizman-ibm

Science and Technology Facilities Council

CPU Stats Comparison

cephalocon 2025

	Classic	Crimson
Context switches / Sec	40,878	**98**
CPU migrations / Sec	6,524	**None**
Lock contention	455,000	**400**
Page faults / Sec	5	**0.7**
Cache Miss rate	12.86%	**4.18%**
CPUs utilized (out of 32)	24.2	**31.78**

14

Crimson Roadmap

cephalocon 2025

Squid

- **Tech Preview** (Bluestore)
- RBD with Replication
- Automated Test Suite
- Snapshots
- MultiCore Stability
- Recovery/Backfill
- Initial Scrub

Tentacle

- **SeaStore Tech Preview**
- Read Performance
- PG Splits
- Seamless deployment
- User Docs

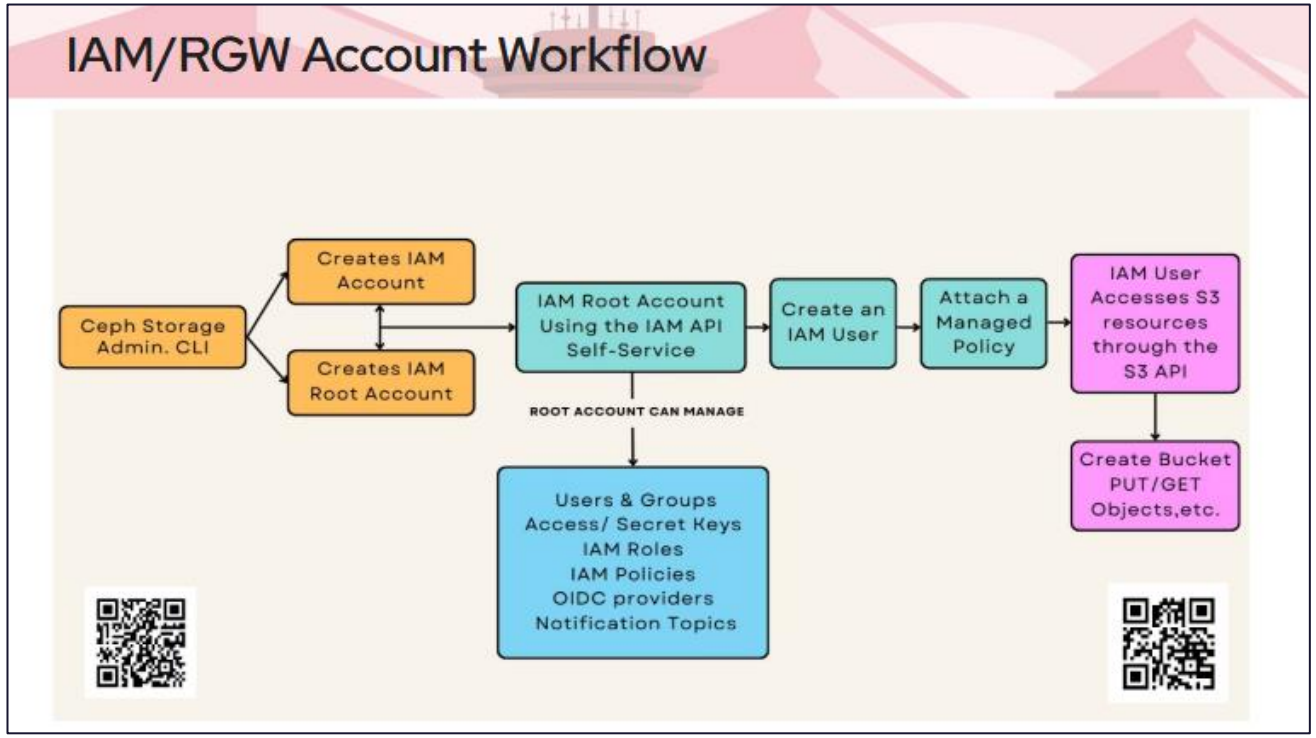
Umbrella

- Write Performance
- Cephadm / Seastore
- QoS / Mclock
- Full Scrub support
- Erasure coding
- PG Merging

18

RGW roadmap

- Squid brings:
  - Multisite improvements, replication status headers
  - Self contained, self service IAM account support
  - Hybrid cloud support – policy-based archive and restore from local or cloud based non-Ceph S3 storage
- Tentacle brings initial deduplication support for large objects



https://cephalocon2025.sched.com/event/27f2y/ceph-object-storage-roadmap-session-at-cephalocon-2025-daniel-parkes-matthew-benjamin-ibm

Science and Technology Facilities Council

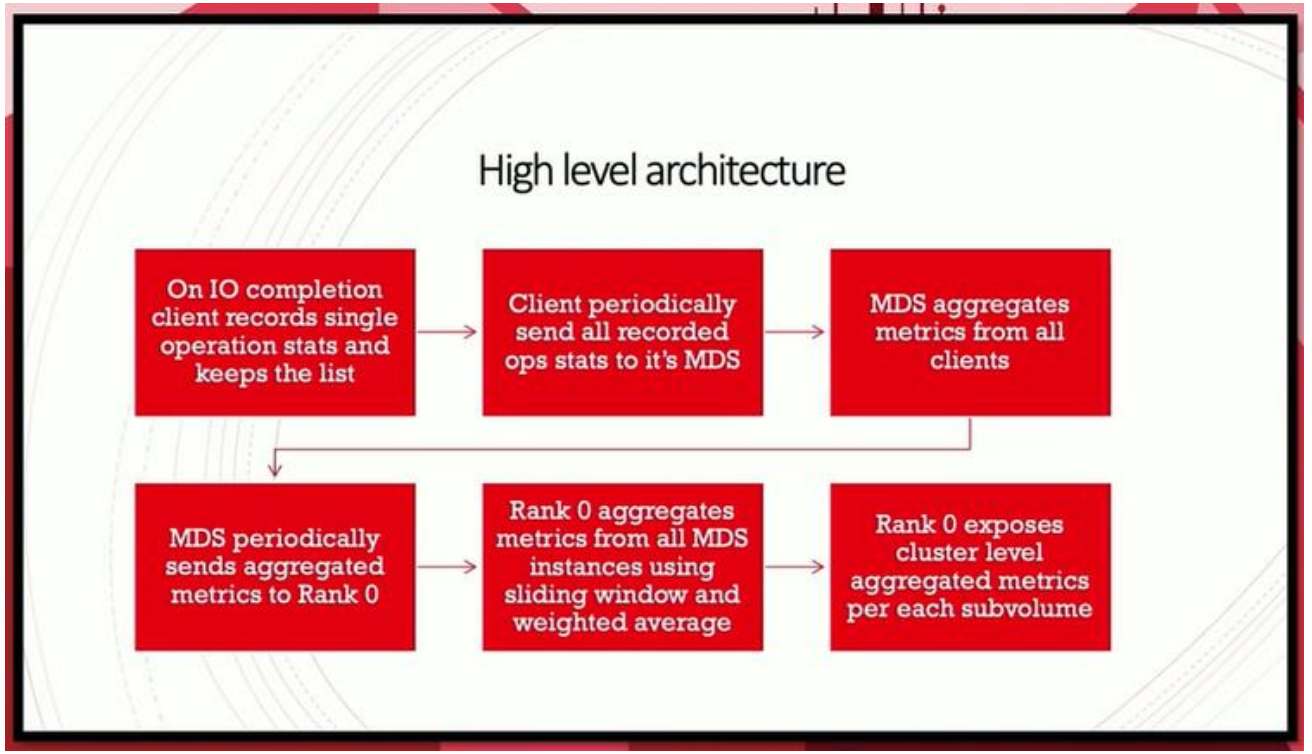
Other RGW/Object storage talks

- Bloomberg - Multi process port binding for seamless restarts
- Bloomberg - Open-source distributed rate limiting for RGW
- Intel/IBM – RGW for K/V Caching to speed up vLLM inference

CephFS talks

Per SubVolume IO metrics

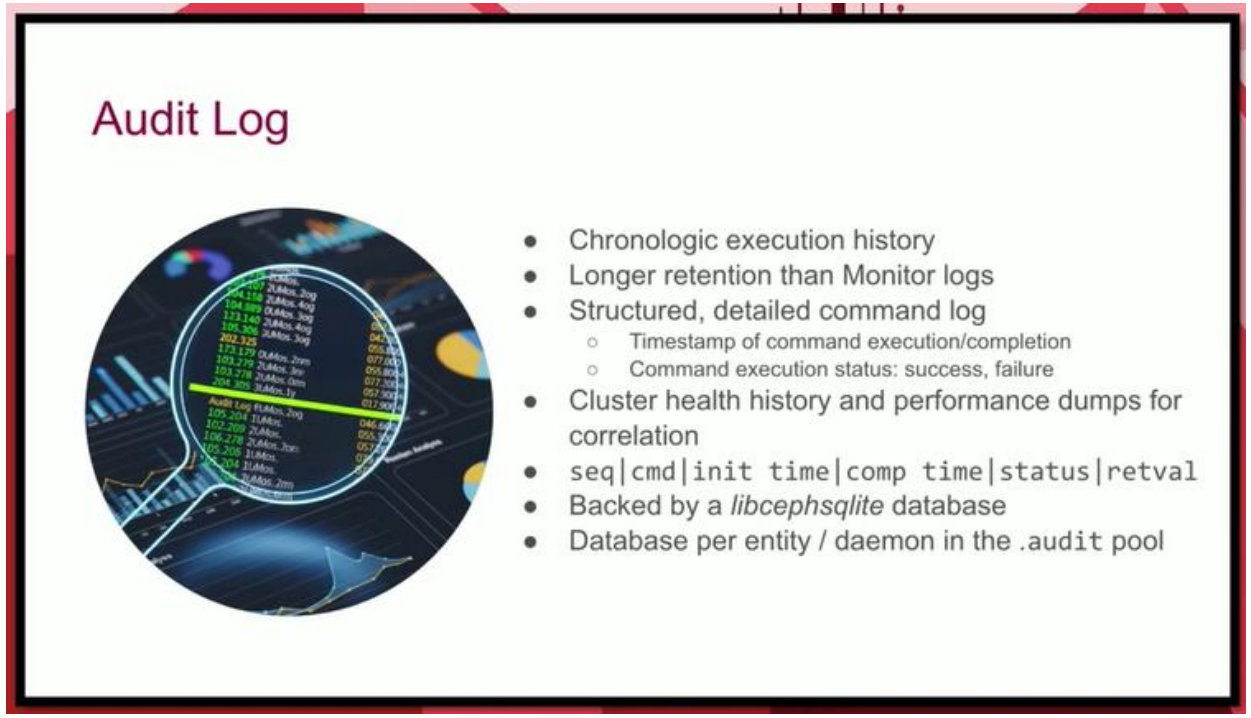
- CephFS improvements to allow monitoring of IO activity related to subsets of a CephFS filesystem
- Lightweight, client driven approach to work with the decentralised IO paths
- Early work currently



https://cephalocon2025.sched.com/event/27f45/per-subvolume-io-metrics-in-cephfs-igor-golikov-alex-markuze-ibm

Science and Technology Facilities Council

Command history logging for CephFS



- A lot of Ceph issues are hard to diagnose and fix due to lack of detailed admin operation history
- Leverage `libcephsqlite` to store a chronological command log along with cluster state at the time
- Starting with CephFS admin commands, but can be expanded to cover all aspects
- Anticipated for Umbrella

https://cephalocon2025.sched.com/event/29vfg/cephfs-tell-me-what-happened-yes-i-really-really-mean-it-anthony-datri-ibm

Science and Technology Facilities Council

Science user stories

Ceph @ CERN

Concerns from daily CephFS operations

1. **Performance / QoS Shaping**
   - No throttling possible for metadata requests by clients
   - *"Something has happened…"*
2. **Poisonous patterns from clients**
   - Stuck requests due to lock contention among clients
     - Might lead to MDS not trimming journal, which is bad…
     - Evicting the client(s) at fault solves the problem, but…
   - MDS might crash when evicting
     - In `"finisher"`, if client has outstanding `"batch getattr"` requests (70624, 70769)
     - Further down the road (`num_rdlock`, `num_pins`), attempting to fix the crash in `"finisher"` (72941, 73006)
   - Other ranks crashing while failed rank is in `rejoin` (62036)
3. **Backups**
   - Walking large directory hierarchies is expensive for the MDS: Cache thrashing, cap locking rtts
   - Even with snapshots, MDS is exposed to extra load for stat/read operations

cephalocon 8

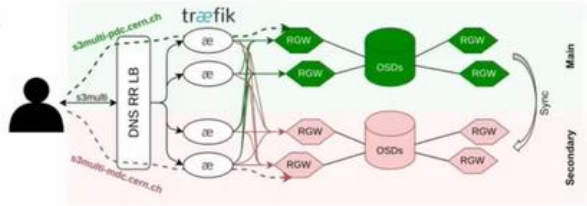
https://cephalocon2025.sched.com/event/27f4B/modernizing-ceph-deployments-at-cern-cephfs-and-object-storage-across-data-centres-enrico-bocchi-cern

- Heavy users of CephFS, lots of interesting observations
- Investigating moving away from large multi-MDS filesystems where possible for simplicity
- Distributed S3 multisite setup

S3 Multisite

"Traditional" multisite configuration

- 2 clusters, 1 zonegroup, 2 zones
  - Full replica main (rw) to secondary (ro)
- Dedicated groups of RadosGWs for:
  - Client traffic: Traefik frontend (L7 routing, TLS termination, health check, failover), and OpenStack Keystone for credentials
  - Sync traffic: Simple RGW with beast



- Recent addition: Each zone is directly addressable through LBs (`s3multi-{pdc,mdc}.cern.ch`)
  - Allows for inferring replication status with new `x-amz-replication-status` header (blog article)

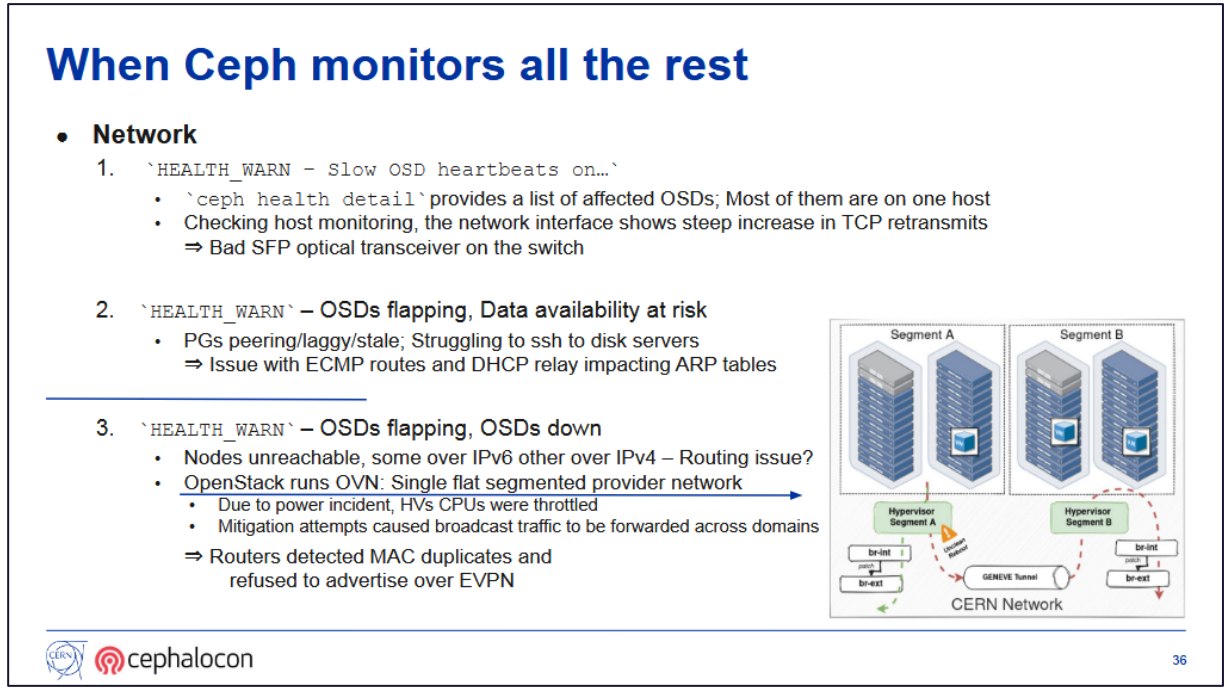
```
# s3cmd --host=s3multi-mdc.cern.ch --host-bucket=s3multi-mdc.cern.ch --debug info s3://enrico/passwd 2>&1 | grep replica
            'x-amz-replication-status': 'REPLICA',
            'x-rgw-replicated-at': 'Thu, 23 Oct 2025 13:33:01 GMT',
            'x-rgw-replicated-from': 'l3ebcc99-db9e-46e0-81db-90a1556ce13b:enrico:l3ebcc99-db9e-46e0-81db-90a1556ce13b.2534364.2'),
```

cephalocon 33

Science and Technology Facilities Council

Monitoring @ CERN

- Great talk about CERN's Ceph monitoring stack built on prometheus+thanos
- Lots of real-world examples of Ceph issues encountered and how monitoring helped

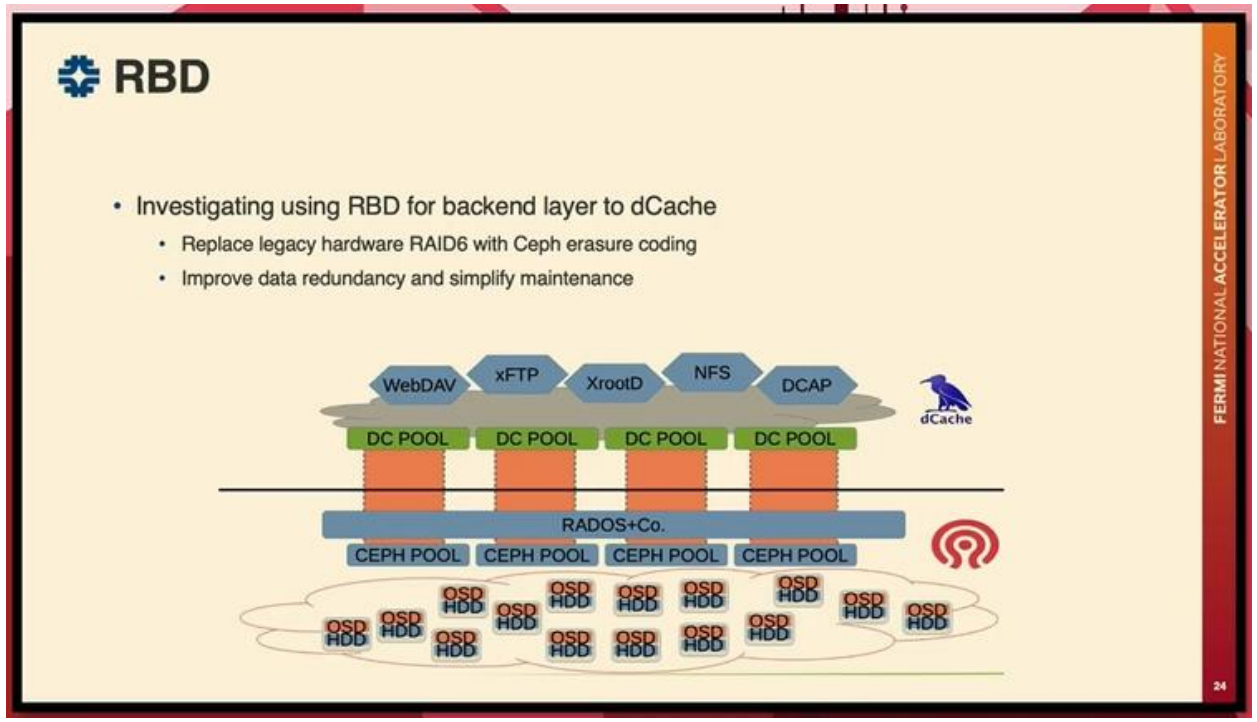


https://cephalocon2025.sched.com/event/27f2F/enhancing-ceph-monitoring-at-cern-our-approach-and-solutions-roberto-valverde-cameselle-enrico-bocchi-cern

Science and Technology Facilities Council

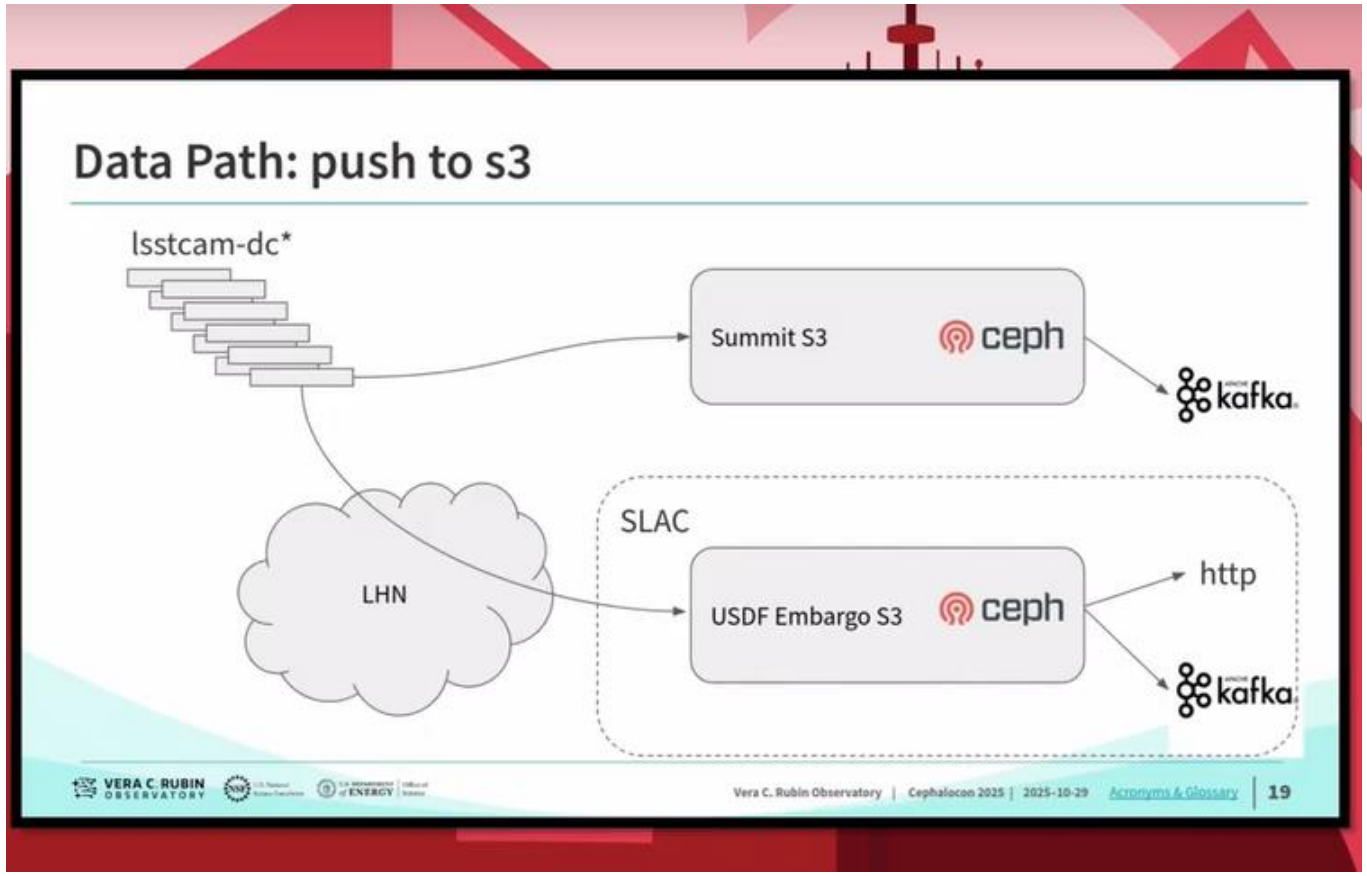
FNAL

- Ceph used as general purpose filesystem for interactive analysis
- Growing interest from CMS in using RGW for scalable object storage
- Investigating replacing dCache RAID arrays with RBD devices



Rubin Observatory

- Ceph RGW used as initial storage of telescope images
- Bucket notifications used to trigger next steps of the data pipeline
- Rook used to run and manage Ceph due to k8s environment



https://cephalocon2025.sched.com/event/27f3w/how-rookceph-enables-science-rubin-observatory-joshua-hoblitt-rubin-observatory

Science and Technology Facilities Council

Ceph project + community update

- Another strong year of development work on the Ceph project – *majority from IBM*
  - Ceph is now used in ~15 IBM storage products
  - IBM also funding new infrastructure for Ceph project labs
  - A rising tide lifts all boats…
- New/rejoining Ceph foundation members – Sony, WD, Samsung
- More important than ever that our interests are represented – we must participate

Science and Technology Facilities Council

Community update

- >2EiB of Ceph clusters reporting public stats
- 5 Ceph-days + a Cephalocon in 2025
  - Numerous virtual events
  - Similar planned in 2026 – look out for the Ceph day London 2026
- New Ceph Foundation community manager appointed – Anthony Middleton

Science and Technology Facilities Council