Ceph at Lancaster University

Matt Doidge, **Gerard Hand**, Steven Simpson, Peter Love

Lancaster University

g.hand@lancaster.ac.uk

Configuration

- Ceph Squid 19.2.3
- 12.2PB Raw
- 5 MON+MGR
- 32 OSD nodes (768 OSD Daemons)
- 2 MDS - 1Active, 1 Standby - 8+3 EC
- 6 XRootD nodes
- 332 Compute nodes
- Monitoring: Grafana, Prometheus, Loki

OSD Node Configuration

- 2 x 1.6TB NVMe to hold OSD BlueStore.
- Each NVMe holds the BlueStore for half the HDDs.
- Not using RAID reduces NVMe wear (25% of RAID)
- Downside: No redundancy when NVMe fails so half an OSD node lost.
- 1 NVMe failure in 4.5 years.

g.hand@lancaster.ac.uk

Observations

- Versions used: Pacific, Reef, Squid.
- Pacific was stable:
  - Scrubbing Problems
- Reef proved to be problematic:
  - Slow OSD Ops
  - Orchestration crashing
  - Poor mClock performance. Recovery rates of 1 Obj/s. Enable *osd_mclock_override_recovery_settings* to improve performance.
  - Permission problems on crash directories.
- Squid more stable but not perfect.
  - Occasional MON crashes - 100% memory usage, Machine lockup.
  - mClock does a better job of balancing client data, backfill/recovery, scrubbing but still room for improvement. Occasionally using *osd_mclock_override_recovery_settings* to increase backfill/recovery when low client transfer load.
  - PGs stuck in scrubbing state.
- Ceph has proved to be resilient. A data centre power cut causing overheating and uncontrolled shutdown of machines resulted in 2 damaged ceph objects.

g.hand@lancaster.ac.uk