GridPP
UK Computing for Particle Physics

# Data Management at Lancaster University

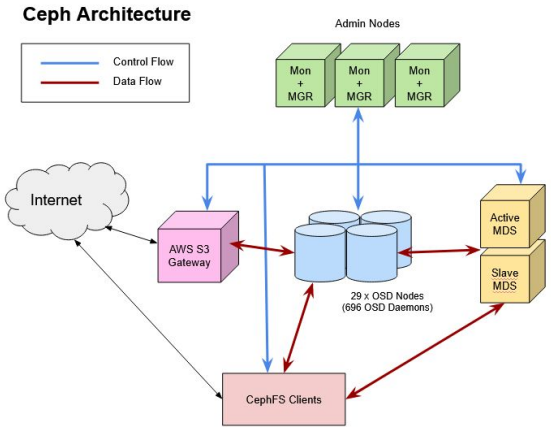
Matt Doidge, Gerard Hand, Steven Simpson, Peter Love

Lancaster University

g.hand@lancaster.ac.uk

# A Step into the Void - XRootD and CephFS

- Existing Ceph+XRootD installations used xrdceph as the interface between Ceph and XRootD. Lancaster chose to trailblaze the use of CephFS to interface to XRootD. https://indico.jlab.org/event/459/contributions/11358/
- After advice from Dan van der Ster we chose Ceph Pacific and configured our cluster with one active MDS server with a single rank backed up by a single standby.
- Data is stored using 8+3 Erasure Coding.
- Installation was done using cephadm.
- Single XRootD server.
- An S3 gateway - light usage.




g.hand@lancaster.ac.uk

# Testing and Tooling

- Initially a ceph cluster was created on a handful of VMs providing a few GB of storage before moving onto the new hardware. We now run a larger VM cluster for testing purposes.
- The existing DPM storage produced the Storage Resource Reporting. We wrote Python scripts to generate the report. https://github.com/lancs-gridpp/cephfs-srr
- Ceph monitoring and alerting using Grafana, Prometheus, Loki.
- Installed XRootD Shoveler.
- ATLAS requires a periodic list of stored files which is created using a simple BASH script.

# Experiences

- We had teething problems with XRootD: Low throughput, file descriptor exhaustion, file permissions. Throughput limitations due to the checksumming load is still our bottleneck in the system. We increased the number of XRootD servers which introduced load balancing issues.  We now have 6 XRootD servers + 1 redirector.
- Ceph provides a reliable platform for servicing data.  Hardware failures/maintenance doesn't mean a loss of service.
- Documentation for Ceph can be difficult to follow and sometimes lacks detail.
- Working out what Ceph is doing can sometimes be difficult.
- We experience intermittent "Slow Ops" which occasionally causes problems with XRootD and causes hammercloud failures.
- We had had a few drive failures which mostly have been straightforward to replace.
- The upgrade from Pacific to Reef was straight forward but has proved problematic.


g.hand@lancaster.ac.uk

# To Infinity and Beyond

- The amount of storage we have has proved to be least we can get away with.
- We are moving to a 100GB/s link so more XRootD servers will be needed to make the most of the connection.
- On the fly checksumming by XRootD would be useful.
- Look at ways to optimize our Ceph/hardware setup to reduce Slow-Ops incidents.
- Upgrade to Reef 18.2.4.
- It would be nice to have a channel dedicated to ceph discussions and documentation.

Lancaster University

g.hand@lancaster.ac.uk