Ceph – Status and Plans at the RAL Tier 1

Alastair Dewhurst, Tom Byrne, Bruno Canning, George Ryall

Introduction

- There are several different projects looking at Ceph on the Harwell site:
  - Grid Cluster – RAL Tier 1's disk only mass storage for the LHC VO's funded by GridPP.
  - Cloud Cluster – Scientific Computing departments
  - ISIS – The Neutron source.
  - Castor – Version 2.1.16 plans to use Ceph as its underlying storage.
- This talk is exclusively about the 'Grid Cluster' which will potentially be an order of magnitude larger than then next biggest cluster.
- This talk has an ATLAS bias. This is because we are at an early stage in the testing. Other VOs will be included shortly.

Terms of reference

- Deploy a production quality disk only storage for use by the LHC VOs from late 2015 onwards.
  - The system should be scalable to 40 PB+ and 100 million+ files
- The system should be maintainable by 1 FTE
  - Ceph is the preferred solution as this will need to be deployed for Castor and this will allow sharing of FTEs.
- The system should be no more expensive hardware wise than Castor.
  - To keep cost down, RAID 6 (or equivalent erasure coding) will be necessary.

Available Resources

- Manpower:
  - Alastair Dewhurst - 20% FTE.
  - Tom Byrne – 100% FTE for 6 months.
  - Bruno Canning – 50% FTE (possibly more).
  - Significant best effort help from others.
- Hardware:
  - ~30 x 40TB disk servers from 2009 procurement.
  - 14 x 120TB disk servers from 2013 procurement.
  - 2015 procurement – expected to be around 6PB raw storage, dual 10GB/s network cards, 1 CPU and 1GB RAM per disk (OSD), SSD for journaling.

CephFS vs Object Store

- Originally thought we were going to go with CephFS because POSIX. However:
  - Requires Meta Data Servers to run.
  - Requires Kernel updates to WN.
  - Still not officially supported.
  - Currently requires cache tier if using erasure coding*.
  - LHC VOs don't actually use the file system.
- Object Stores offer lots of potential benefits but we need to re-design the way VOs access their data!

ATLAS Object Store

- How might an ATLAS Object store look?
- Pool/Bucket per Spacetoken (this is where file permissions can be set)
  - Can we write 100 million files to a Pool without breaking it?
- Within these Pools objects stored as "scope:name"
  - Might make it easier to use with Rucio (and FAX).
- Can write a pilot mover script (few lines of python) to use RADOS to directly access files from WN (unix style permissions).
  - Will need to performance test this (Direct I/O or Copy to scratch?).
- FTS transfersvia S3 or WebDAV?

Time line

Testing Now

- Decide between CephFS and Object store.
- Decide on supported protocols.
- Setup monitoring and performing hardware benchmarking.

Deployment March 2015+

- New hardware can be deployed into Ceph.
- Refine operational procedures and tune settings.
- 2013 hardware moved into Castor to meet GridPP pledge.

Production September 2015

- Decision needs to be made about 2015 procurement.
- Need to start migrating at least 1 LHC VO.
- Last chance to back out.

Testing so far

- At GridPP33 James Adams gave a talk on the Ceph testing that has taken place.
- Initial plan was for ARC CEs to use CephFS as a shared file system to help with job submission.
  - Jobs never succeeded due to an unrelated problem with PID namespaces. By the time this was figured out CephFS had gone down.
- GridFTP server did work, but demonstrated the need for separate networks.
- Gabriela Fidyk (Erasmus student) did some benchmarking tests on SSD which were inconclusive.
  - George Ryall plans on repeating them.

Current status

- 7 node Ceph cluster up (using 13 generation hardware)
  - Plan to add equivalent amount of old hardware to demonstrate effectiveness of crush maps
- 34 disks per node, but only 4 physical cores, may have to reduce number of disks being used for testing.
- S3 gateway on virtual machine. Hope to have firewall holes + x.509 authentication working by next week.
- Current erasure coding level M=2 K=1.
  - Will switch to more realistic setting (M=14, K=2) once older hardware added.

FTS

- The LHC VOs use FTS for large amount of data transfer.
  - BNL have requested S3↔S3 and S3↔gridftp/SRM support which should be added in November.
  - It won't be third-party however, it will use tunneling through the FTS3 host itself!
- There is no reason why it should not be possible to setup a WebDAV gateway for Ceph.
- It would be very interesting to see if we could get FTS transfers working between Tier 2s using WebDAV.

Networking and IPv6

- Ceph recommends that network traffic is split between replication traffic and production + monitor traffic.
  - New procurement will come with 2x NICs.
  - Early testing has already demonstrated what happens if the network link gets saturated.
- ATLAS and CMS want sites to provide dual stack storage sooner rather than later.
- The Gateways can be made dual stack.
  - Need to wait on RAL to be able to provide IPv6 addresses although Tiju is working on it.
- Ceph supports IPv4 or IPv6 but not dual stack?!

Monitoring & Operations

- We will setup and play around with Calamari.
  - Currently need to compile source code into package. Share this task?
- Will integrate with Nagios and Ganglia monitoring.
- Philosophy will be to treat the Ceph cluster as if it were a (low priority) production service as soon as possible.
  - We want to always know the state of the cluster and not leave it down without a plan to have it back up.
  - We have an internal Ceph RT queue setup (email lcg-support@gridpp.rl.ac.uk with Ceph in the title)

Summary

- Provided an introduction to the work going on at the Tier 1 on Ceph.
  - Will try and present updates at monthly intervals to either Storage or Technical meeting.
- We feel that an SRMless Object Store is a realistic possibility.
  - It will be a lot of work, will greatly appreciate sharing knowledge/testing with Tier 2s.