Science & Technology Facilities Council

Experience of Development and Deployment of a Large-Scale Ceph-Based Data Storage System at RAL

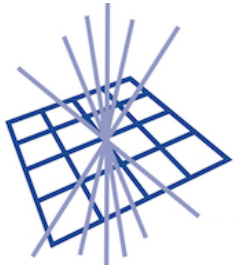
Bruno Canning

Scientific Computing Department

UK Tier 1, STFC Rutherford Appleton Laboratory

HEPiX Fall, 2016, LBNL

bruno.canning@stfc.ac.uk



GridPP

UK Computing for Particle Physics

Outline

- A bit about me
- Brief recap:
  - Ceph work in SCD at RAL
  - Echo cluster design brief
  - Cluster components and services offered
- Our experience of developing and deploying Ceph-based storage systems
  - Not a status report but a presentation of our experience

Hello!

- LHC Data Store System Administrator
  - 3 years at RAL in SCD
  - Based in Data Services Group, who provide Tier 1 storage, yet embedded in Fabric Team
  - Previously worked on CASTOR, now work on Ceph
  - Specialise in:
    - Linux system administration
    - Configuration management
    - Data storage
    - Server hardware
    - Fabric support

Science & Technology Facilities Council

The Story so Far

- Two preproduction Ceph clusters in SCD: Echo and Sirius
- Echo designed for high band-width file/object storage
  - LHC and non-LHC VOs
- Sirius designed for low latency block device storage
  - Departmental cloud service and local facilities
- Use case and architecture of both discussed previously in greater detail by our James Adams:
  - https://indico.cern.ch/event/466991/contributions/2136880/

Echo Cluster Components

- Cluster managed via Quattor and its Aquilon framework
- 3 × monitor nodes (ceph-mon)
  - first monitor is also ceph's deploy host
  - ncm-ceph automates ceph-deploy
    - Prepares and distributes ceph.conf to all nodes
    - Manages crushmap and OSD deployment
- 3 × gateway nodes (xrootd, globus-gridftp-server, ceph-radosgw)
- 63 × storage nodes (ceph-osd)
  - 36 × 6.0 TB HDD for data, total count = 2268
  - Capacity = 12.1 PiB, 13.6 PB Raw, 8.8 PiB, 9.9 PB usable
- No SRM

quattor

Science & Technology Facilities Council

XrootD Plugin

- XrootD plugin for Ceph written by Sébastien Ponce, CERN
  - https://indico.cern.ch/event/330212/contributions/1718786/
- Uses libradosstriper, contributed by Sébastien to Ceph project
  - Available since Giant release
- Plugin itself contributed to XrootD project
- First demonstration by our George Vasilakakos with Giant release in first half of 2015, but not packaged
- The xrootd-ceph RPM is not distributed, version of Ceph in EPEL (Firefly release) predates libradosstriper
- Needed to build xrootd RPMs against Ceph development packages to get xrootd-ceph plugin

XRootD

Science & Technology Facilities Council

Starting the XrootD Daemon

- Following upgrade to Ceph Hammer release, xrootd daemon will not start, reports an LTTng error
- Caused by the registration of tracepoint probes with the same name in both rados and libradosstriper code by their respective contributors
- Patch contributed (May 2015), merged into master after 0.94.5 released (October 2015)
- Needed to build Ceph RPMs to get patches and meet testing deadlines
- Demonstrated working xrdcp in and out of cluster (November 2015) with x509 authentication
- Patches incorporated into 0.94.6 release (February 2016)

Data Integrity - XrootD (1)

- Problems with files copied out with multiple streams
- Incorrect checksum of large (4 GiB, DVD iso) files returned in 81% of 300 attempts but always same size as original
- Created a test file containing 1069 × 4 MiB blocks
- Test file: 16 byte lines of text in 1069 × 4 MiB blocks, can identify line number in a block and block number
- Identified the following:
  - 1 to 3 blocks contained incorrect data, appearing at random
  - Incorrect blocks contained 2 MiB duplicated from another block
  - Overlapped exactly with either 1st or 2nd half of block
  - Typically from next block but could be from up to 11 blocks behind or 12 blocks in front of bad block

Data Integrity – XrootD (2)

- Communicated findings to Sébastien
- He was quickly able to reproduce the problem
- He determined a race condition occurred due to the use of non-atomic operations in a threaded context
- Patch committed to GitHub, rebuilt XrootD RPMs at RAL, problem solved
- Resolved in one week after initial contact with Sébastien
- Great team effort and collaboration with partners
- Happy sys admins and relieved project managers at RAL

GridFTP Plugin

- Started by Sébastien Ponce, continued by Ian Johnson at RAL
  - https://github.com/ijjorama/gridFTPCephPlugin
- Also uses libradosstriper
- Uses XrootD's AuthDB for authorisation
- Uses xrdacctest to return authorisation decision
- No problems with single stream functionality…
- … but out-of-order data delivery with parallel writes in MODE E used by FTS transfers problematic
- Erasure Coded pools in Ceph don't support partial writes, hence they don't support non-contiguous writing
- We now have a fix that is undergoing testing

Plugin Summary

- XrootD and GridFTP available and maturing
- Plugins talk directly to Ceph
- Plugins are interoperable
- Removed requirement for leading '/' in object name from both
  - Comes from history of use with POSIX file systems
  - This would require VO pool names to have a leading '/'
  - This character is supported, but concerned this may change
  - GridFTP plugin works around this
  - XrootD added support for object IDs
- We have working authentication and authorisation with both

Ceph Version Upgrades

- Typical sequence*: Monitors, storage nodes, meta data servers (if present) then gateways
- Restart daemons on node type, then upgrade next type
- Performed whilst the service is online
- Very easy from one version within a release to the next
- Just change RPM version number
- Can skip some intermediate versions
- Takes c. 30 minutes, can be performed by one sys admin
- Much simpler than CASTOR upgrades
  - * Change in Hammer OSDMap format with 0.94.7 requires different order, our thanks to Dan van der Ster
  - https://www.mail-archive.com/ceph-users@lists.ceph.com/msg32830.html

Ceph Release Upgrades

- Change RPM version and repository
- Firefly to Giant: Easy
- Giant to Hammer: Easy, but introduced LTTng problem
- Hammer to Jewel: More involved:
  - Jewel requires SL7
  - Daemons now separate, not one daemon with many roles
  - Daemons are now managed by systemd ...
  - … and run as ‘ceph’ user, not ‘root’ user
  - Needed to define ulimits for ‘ceph’ user (nproc, nofile)
  - Change ownership of /var/lib/ceph
- Upgraded to SL7 with Hammer first, then moved to Jewel

Science & Technology Facilities Council

New Operating System - Installations

- Upgrade to Jewel required us to support a new OS major version, before we were really ready for it
- Work required to adapt deployment configuration
  - Previously reliant on dhcp for network config during PXE and kickstart, quattor then updates to proper, static config
  - New (again) NIC naming convention with SL7, controls NIC names throughout entire installation
  - However, nodes in subnets other than the deploy server also need routing information in order to install, currently not configured in pxelinux.cfg or kickstart
- SL7 installations still need supervising and often need help, SL6 installations just work

New Operating System – Site Config

- Work required to adapt site-wide configuration for SL7 and for Ceph
  - /etc/rsyslog.conf
    - We send core system logs and certain application logs to central loggers
    - Fewer modules loaded by default with SL7, required modules (e.g. imuxsock) must be declared in file, modules for systemd journal must be added (imjournal), as must other directives
  - /etc/sudoers
    - Long-standing requirement to support ceph as a second use case of sudo
    - Use for nagios tests executed via NRPE, sudoers config in an RPM
    - Deploy host needs sudoers config on all nodes for ceph-deploy
    - ncm-ceph provides this via ncm-sudo
    - Conflict with RPM and ncm-sudo

New Operating System - Alerting

- Using nagios RPMs from EPEL
- Older than latest version but we don't need new features yet
- NSCA
  - Version difference between EPEL 6 (2.7) and 7 (2.9)
  - Versions are incompatible, packet size expected by server differs
  - Packaged nsca-client 2.7 for SL7 hosts
- NRPE
  - NRPE daemon runs as 'nrpe' user, not 'nagios' user
  - Reconfigured NRPE to run as 'nagios' user via systemd unit file
  - Add unit file under /etc/systemd/system/nrpe.service.d/
  - No need to modify existing sudoers config for SL7
- Would have preferred to upgrade site infrastructure to SL7 first

Upgrade to SL7 and Jewel

- After upgrade to SL7, OSD daemons start to crash
- First OSD crashes after c. 2 days of normal operation
- Crashes continue one at a time, but hundreds can build up
- Have to periodically start OSDs, but can restore full operation with patience
- Problem became much worse during upgrade to Jewel
  - Could not keep all OSDs running despite best effort
  - Over half the OSDs stopped running, client I/O halted
  - Ready to dismantle existing cluster (FY'14 storage) and build new cluster (FY'15 storage) anyway
  - Most interested in gaining upgrade experience
  - Decided declare the service dead to the test community and build new cluster
- But what was the cause?

The XFS Bug

- XFS reports possible memory allocation deadlock in kmem_alloc on storage nodes
- Causes XFS to hang which causes the OSD to crash
- Bug already reported:
  - http://tracker.ceph.com/issues/6301
  - http://oss.sgi.com/pipermail/xfs/2013-December/032346.html
- Bug present in kernels shipped with SL5 and SL6, fixed in kernels shipped with SL7
- File systems on data disks will need to be recreated on SL7
- Will affect all storage systems
- Our thanks to Dan van der Ster at CERN for communicating this to us

Deployment of FY'15 Storage

- Supermicro: 36 HDDs on HBA with one internal OS HDD
- Specified SL6 for ITT and used for pre-deployment testing
- Pre-deployment 'hammer' testing proceeded smoothly
- Deployment into Ceph with SL7 problematic:
  - OS disk now last element in list, /dev/sdak
  - Config management assumes OS disk is /dev/sda
  - Deploy host profile contains disk layout of all storage nodes, read from storage node profiles
  - Changed config management to programmatically identify boot disk for all nodes and exclude the boot disk from list of data disks
  - Simple change but can change profile of every node in data centre
  - Test, test and test again to get it right
- Deployment delayed but proceeded as expected once able
- Service now very stable

Reflections

- We can build services users want based on Ceph storage
- Challenge was always going to be technical in nature but…
- ... we did not appreciate the development effort required early on in the project
- Ceph administration is a learning curve
- Online documentation is generally good but the “really dangerous” features proved useful so could improve further
- Outlook is promising

Acknowledgements

- RAL Team:
  - George Vasilakakos, Ian Johnson, Alison Packer, James Adams, Alastair Dewhurst
  - Previous contributions: George Ryall, Tom Byrne, Shaun de Witt
- Collaborators:
  - Dan van der Ster and Sébastien Ponce at CERN
  - Andrew Hanushevsky at SLAC
  - Brian Bockelman at University of Nebraska-Lincoln
- And everyone who made this happen:

Thank you

Science & Technology Facilities Council

Backup Slides

Science & Technology
Facilities Council

Echo Cluster

- Brief: Provide a replacement for CASTOR disk-only storage
- Motivation:
  - Reaching limitations of CASTOR performance
  - CASTOR requires sys admins and a DBAs
  - Several components of the CASTOR system
  - Small community: cannot recruit experts, time consuming to train them
  - Reduced support: CERN have moved to EOS for disk storage
- Requirements:
  - Performance must continue to scale with work load
  - Only require one specialism (sys admins)
  - Must support established data transfer protocols used by the WLCG
- Benefits:
  - Larger community: can hire experts, reduced training time, Ceph specialises in disk storage
  - Less effort to operate and maintain

Science & Technology Facilities Council

Services Offered

- Ceph 10.2.2 (Jewel release, has LTS) running Scientific Linux 7x
- One pool per VO created with 8+3 erasure coding
- Custom gateways: GridFTP and XrootD
  - Plugin architecture, talk natively to Ceph
  - Both built on libradosstriper
  - Both endpoints are interoperable for puts/gets
- Authentication via x509 certificate, currently grid-mapfile
- Authorisation via XrootD's AuthDB for both GridFTP and XrootD
  - Naming scheme: <pool_name>:<space_token>/<file_name>
  - Grants rw access to production users, ro access to ordinary users for data space and rw access to all users for scratch space
- Also offering direct S3/Swift access to trusted users
- Testing Dynafed

Science & Technology Facilities Council

Networking Considerations

- Experience with CASTOR: Network is much like a utility to which we connect our servers, like power
- With Ceph, need to consider networking as part of your Ceph infrastructure
- Will require dual networks on storage nodes in large/busy clusters
  - Public or client network
  - Cluster or private network
- Networking diagram given in site report by Martin Bly:
  - https://indico.cern.ch/event/531810/contributions/2302103/

The SIRIUS Incident (1)

- Occasionally all OSDs crash on a storage node that never was “right in the head”
- No previous impact on service but a nuisance to service manager
- Repairs undertaken and returned to service but still not fixed
- OSDs shutdown, but node not removed from cluster
- Rolling reboots for kernel updates accidently brought node back into cluster
- OSDs crash again but leaves 4 PGs in a stale+incomplete state
- These PGs are unable to serve I/O
- Affects usability of every running VM, VMs need to be paused
- Try to fix the problem but Ceph admin procedures do not help
- Problematic storage node removed from cluster and new VMs created in new pool
- However, many people’s work is potentially lost

The SIRIUS Incident (2)

- Despite best efforts, no resolution
- Other deadlines divert attention
- Discuss incident with Dan at CERN and set him up with root access
- Dan is aware of undocumented procedure that tells OSD to ignore history and use the best information available
- Performs change on OSDs
- PGs spend some time backfilling but eventually go active+clean
- All old VMs can now recovered, however six weeks have passed
- Many not required but some users glad to recover work they feared was lost
- Reputation of cloud service somewhat diminished but did demonstrate we can recover from a disaster
- Personally glad we took the hard way out as considered deleting pool

Hardware

- 3 × monitor nodes:
  - Dell R420, cpu: 2 × Intel Xeon E5-2430v2, ram: 64 GiB.
- 3 × gateway nodes:
  - Dell R430, cpu: 2 × Intel Xeon E5-2650v3, ram: 128 GiB.
- 63 × storage nodes:
  - XMA (Supermicro X10DRi), cpu: as gateways, ram: 128 GiB.
  - OS Disk: 1 × 233 GiB 2.5” HDD, Data Disk: 36 × 5.46 TiB HDD (WD6001F9YZ) via a SAS HBA.
- Total Raw Storage = 12.1 PiB, 13.6 PB
- Usable Storage = 8.8 PiB, 9.9 PB