Science & Technology
Facilities Council

Ceph at the UK Tier 1

George Ryall (STFC)

James Adams (STFC), Alastair Dewhurst (STFC), Rob Appleyard (STFC), Kenneth Waegeman (UGent)

HEPiX Annecy-Le-Vieux, May 2014

Overview

- What is Ceph?
- RAL use cases
- Progress
- Future plans
- Concerns/problems encountered

What is Ceph?

“Ceph is an open-source, massively scalable, software-defined storage system which provides object, block and file system storage in a single platform.”

http://www.inktank.com/what-is-ceph/

Three Storage Types

- Object storage – using native API or RESTful Swift or S3 APIs
- Block storage
- File System – CephFS, not currently supported by inktank (“support coming in 2014”…)

Our Use Cases

1) To provide a storage backend for our departmental cloud.
2) For some time we have been looking for a new non-domain specific storage solution for grid data to replace our current castor based one.

Why we like Ceph

- Improved resilience over other solutions that use disc servers with RAID – the loss of a disk server does not cause any data loss when replication or erasure encoding is used.
- Support for CephFS incorporated into Linux Kernel (since 2.6.34), mounting CephFS as simple as running a mount command – no additional packages or configuration required.
- Open source
- Runs on commodity hardware

Progress- Development Cluster

- Development cluster: 6 nodes, ~100TB, Emperor.
- Has been running for several months, being used for familiraisation
- Currently upgrading to firefly.

Progress- Quattor Component

- Successfully using Ceph Quattor component (developed by Kennith Waegeman, UGent) for configuration management . Based on CephDeploy, the component is capable of:
  - Installing OSDs, Monitors, metadata servers
  - Pushing configuration and crush maps
  - Creating the initial cluster
- Like Ceph deploy it is not able to deploy object gateways for the RESTful APIs.
- Will be using this component from now on for the above cluster operations

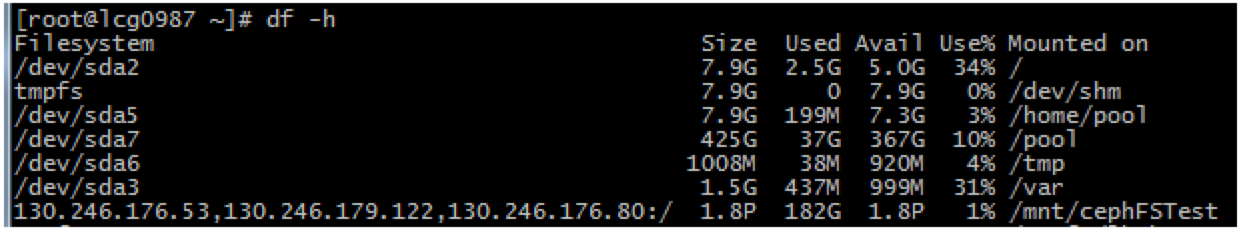
Science & Technology
Facilities Council

Progress – Cloud Storage

- Using the development cluster we have demonstrated the use of Ceph for both providing storage for machines instantiated in OpenNebula and for storing machine images.
- Hardware (~1PB) for a cluster to act as storage backend for the departments cloud offering has been delivered and is being installed.

Progress – Grid Storage

- Started to perform testing with 114 worker nodes and CephFS – required kernel upgrade
- Grid storage cluster – 1.8PB, currently running and exposed to some CMS and Atlas jobs through CephFS.



```
[root@lcg0987 ~]# df -h
Filesystem                                       Size  Used Avail Use% Mounted on
/dev/sda2                                        7.9G  2.5G  5.0G  34% /
tmpfs                                            7.9G     0  7.9G   0% /dev/shm
/dev/sda5                                        7.9G  199M  7.3G   3% /home/pool
/dev/sda7                                        425G   37G  367G  10% /pool
/dev/sda6                                       1008M   38M  920M   4% /tmp
/dev/sda3                                        1.5G  437M  999M  31% /var
130.246.176.53,130.246.179.122,130.246.176.80:/  1.8P  182G  1.8P   1% /mnt/cephFSTest
```

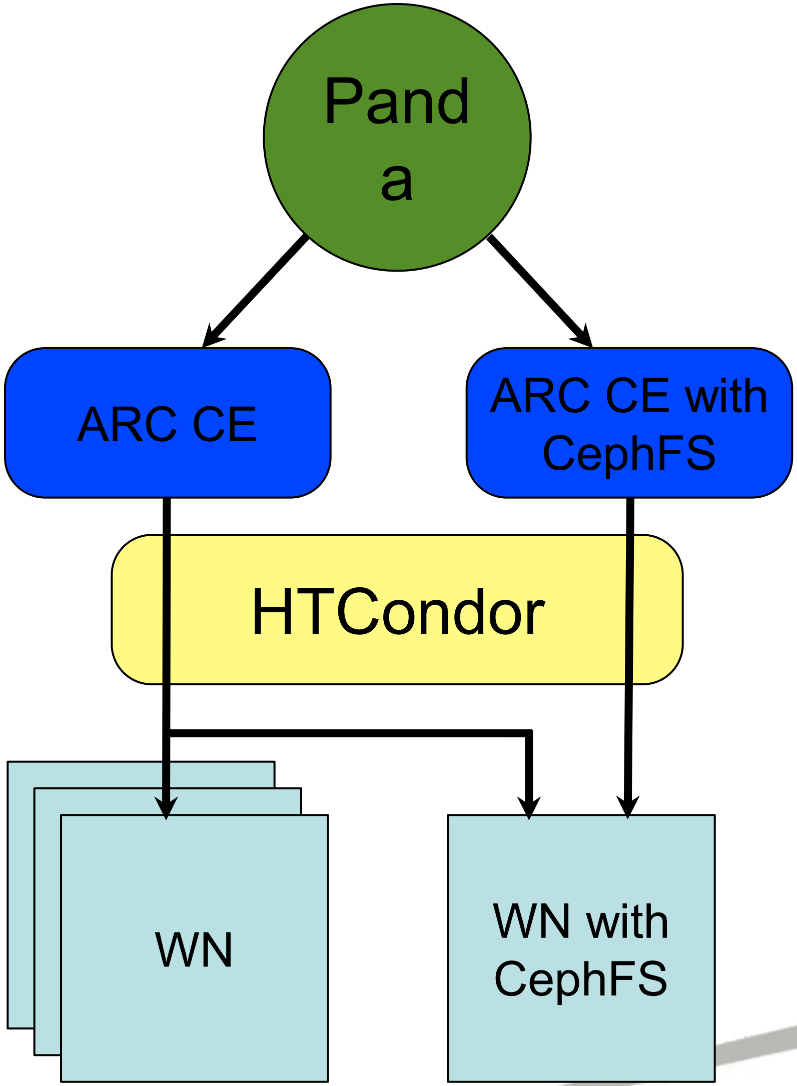
ATLAS testing

- ATLAS use ~5PB of disk storage at RAL (Castor).
- xRootD interfaces would allow testing of full range of ATLAS workflows – we are working on this.
- CephFS can be used by jobs running on the batch farm.

ATLAS testing

- Production ATLAS jobs are running on WN with upgraded Kernels - no problems observed so far.
- Plan to make use of an ARC CE feature which uses shared file system between WN and CE as a cache.
- Can put realistic load on Ceph without storing data permanently.

Test setup



- Queue within Panda submits jobs to ARC CE which has CephFS mounted
- HTCondor directs jobs to WN advertising that they have CephFS
- Normal jobs can use any WN.
- Will slowly upgrade WNs to enable CephFS
- Test jobs successful
- Preparing for production task.

Future plans

- Install firefly on our grid storage cluster
- We will be testing features offered in the firefly release, particularly erasure encoding. This will take less space than full replication, but still provide redundancy. There are concerns over what use cases erasure encoding currently supports.
- Test cache tiering feature of Firefly.

Cluster optimisation

- Optimising Ceph cluster operations, especially on large scale clusters, currently seems to be poorly documented within the Ceph community.
- We will be taking on an Erasmus student this year and giving them a project to optimise the cluster by testing different configurations. E.g. they will be looking at whether moving journaling onto SSDs has a positive performance gain, and if it does they'll quantify the gain so we can decide if it justifies the additional cost.

Problems We've Encountered

- Problems gathering stats in Emperor on our grid Cluster – going to install firefly and hope the problems go away!
- Recently found we had an incorrect (too recent) version of a package and that was preventing cluster installation, package came from Ceph repo.
- Ceph documentation is good, but has gaps. Not all of the functions provided to perform operations the cluster are fully documented.
- SL6 has kernel 2.6.32, CephFS requires 2.6.34

Problems We’ve Encountered

- Some inconsistencies in configuration. Pool numbers rather than names need to be specified, poor documentation on pools and assigning sections of file system to pools. Administrative interfaces often frustrating and not intuitive to use.
- We are concerned at the lack of support for using CephFs for production data and this represents a fairly big risk to us moving forwards.

Summary

- We currently have 3 PB of hardware earmarked for use with Ceph (and climbing)
- Ceph looks promising as a technology but currently has gaps in it's documentation – lack of support for production use of CephFS is concerning
- We will continue testing Firefly and will be looking to expose CephFS to more production jobs.