The deployment of a large scale object store at the RAL Tier-1

A Dewhurst, I Johnson, J Adams, B Canning, G Vasilakakos, A Packer
STFC - Rutherford Appleton Lab. UK
E-mail: alastair.dewhurst@cern.ch

Abstract. Since 2014, the RAL Tier-1 has been working on deploying a Ceph backed object store. The aim is to replace Castor for disk-only storage. This new service must be scalable to meet the data demands of the LHC to 2020 and beyond. As well as offering access protocols the LHC experiments currently use, it must also provide industry standard access protocols. In order to keep costs down the service must use erasure coding rather than replication to ensure data reliability. This paper will present details of the storage service setup, which has been named Echo, as well as the experience gained from running it.

The RAL Tier-1 has also been developing XrootD and GridFTP plugins for Ceph. Both plugins are built on top of the same libraries that write striped data into Ceph and therefore data written by one protocol will be accessible by the other. In the long term we hope the LHC experiments will migrate to industry standard protocols, therefore these plugins will only provide the features needed by the LHC experiments. This paper will report on the development and testing of these plugins.

1. Introduction
Since before the start of LHC Run 1, the RAL Tier-1 has been running Castor for storing both disk-only and tape backed data. In 2012, CERN switched to using EOS for disk-only storage. ASCG, who were also running Castor, switched to using DPM for their disk-only storage. This left the RAL Tier-1 as the only site running Castor for disk-only storage. The RAL Tier-1 conducted its own review of available storage software which reported in May 2013. The conclusions were:

- That Castor should be able to meet the requirements for LHC Run 2.
- dCache was the only solution that met the requirements and was deployable at the time. However it did not offer sufficient improvements over Castor to be worth the additional operational cost involved.
- Hadoop and Ceph were both seen as interesting possibilities, but concerns remained about missing functionality in CEPH and instabilities in Hadoop.

It was therefore decided to remain with Castor. In June 2014, due to increasing concerns with Castor, the previous decision was reviewed. In the intervening year, Ceph had matured significantly and other sites (including CERN) were investigating it. It had also recently been decided that the new cloud infrastructure at RAL would use Ceph for the backend storage. In October 2014 a graduate student started work on a prototype Ceph cluster with the aim of...
eventually replacing Castor for disk-only storage. They key requirements for the new storage cluster were:

- The hardware cost per usable TB of storage should be kept in line with Castor.
- Once it had been deployed it would require 1 - 1.5FTE effort to maintain.
- It must scale to meet the demands of the LHC Run 3

The original prototype was made from old Castor storage nodes. In April 2015, dedicated hardware for the test cluster was delivered. In April 2016, the first batch of hardware that would be part of the production Ceph cluster was delivered. In July 2016, the Echo storage service was brought into existence.

This paper will focus on the testing and work done in the last year and a half and is organised as follows. Section 2 will describe the setup of the Echo cluster. Section 3 will describe the work done on the GridFTP and XRootD plugins that are necessary for the LHC experiments. The paper ends with a short summary in section 4.

2. Echo Cluster

Echo is a Ceph cluster which provides access to its storage via gateways that support the XRootD and GridFTP protocols. A RADOS Gateway, providing S3 and Swift compatible API access is also provided. The name Echo comes from the key design principles behind the cluster: Erasure coded, Ceph, High throughput and Object store.

The Echo cluster is made up of 3 Ceph physical monitors as well as 62 storage nodes which each have 36 x 6TB drives. The overall raw storage capacity is 13.4 PB. There are currently 3 machines which will act as gateways for external traffic. The number of gateways will be increased before the service enters production.

2.1. Network design

To access data stored in a Ceph cluster, all that is required is access to the CRUSH map as well as a keyring. This means that for external access to the cluster must go via gateway machines. Careful consideration must be given to the deployment of these machines so as to not create a bottleneck. Figure 1 shows the physical network topology of the Echo cluster within the RAL Tier-1. It also shows a simplified network topology of some of the other relevant service the Tier-1 provides.

There are up to eight storage nodes in each rack. Each storage node is connected by a pair of 10Gb/s links to a network switch. One network switch provides connectivity to three racks of storage nodes. Each of these network switches are connected by two 40Gb/s links to the Tier-1 network core.

Each storage node actually has four 10Gb/s network interfaces and the hardware in the racks has been arranged such that it would be possible to purchase another set of network switches which would not only double the maximum throughput to the machine but also provide resilience in case of a network switch failure.

The gateways and the monitors are connected to a separate pair of switches which are both connected to the Tier-1 network core via two 100Gb/s links. The gateways themselves have four 10Gb/s network connections. For simplicity it was decided to connect the four links to the same network switch, which would allow them to be bonded. For the monitors, it is critical for the service that, they stay up and therefore they were connected to both switches. Only one of the links is active at any time. In the event of a failure of a switch, while some of the gateways would become inaccessible, the monitors should all stay up and the service as a whole would stay up (with a reduced throughput).
2.2. Erasure Coding

One of the key specifications when building Echo was that the hardware cost per TB of usable storage needed to be kept in line with Castor. In Castor storage nodes were normally 24 or 36 bay machines with a RAID6 or RAID60 configuration. It is immediately apparent that using replicated pools would not provide anywhere near the same cost per TB of usable storage. Erasure coding was therefore the only choice.

In Ceph, erasure coding splits an object into $k$ chunks and then writes a total of $n$ chunks (where $k < n$) of data to the storage. The object can be re-created from any $k$ chunks. $n$ is known as the set size. The difference, $m = n - k$, is the number of chunks that can be lost before data integrity is compromised. The $n$ disks that a particular piece of data is written to is referred to as its placement group. Ceph storage pools are made up of large numbers of placement groups so as to spread the data evenly across the storage nodes. It should be noted that any particular disk will be part of many different placement groups.

The RAID configuration normally used on storage nodes in Castor would be equivalent to $k = 16$, $m = 2$. Table 1 shows what percentage of raw storage capacity is actually usable under different erasure coding profiles.

Another specification for Echo was that it needed to be easier to maintain than Castor. With RAID6, while single disk failures are easily dealt with, double disk failures, which happen on average a little over once a month, will trigger a call-out to minimise the risk of data loss. The number of placement groups in Echo is already an order of magnitude higher than the number of RAID arrays in Castor so the chance of two random disk failures being part of the same PG is significantly higher. By increasing $m$ to 3, even if two disks in an acting set fail in quick...
succession, there is still resilience against a further disk failure and the need for call-outs is reduced.

Initially a $k = 16$, $m = 3$ erasure coding profile was chosen. It quickly became obvious that this wouldn’t work as the disks in each placement group found it very difficult to remain synced with each other and the cluster was very unstable. The number of messages sent is proportional to the square of the set size. The vast majority of known production Ceph clusters use 3 replicas of the data which has a set size of three. There is ongoing work by the Ceph developers to improve this, however it was decided to reduce $k$ to 8. This was primarily chosen because Yahoo runs a modified version of Ceph to store their data and use an $k = 8$, $m = 3$ erasure coded profile [1]. Even with this reduction in the set size, it was non-trivial to keep all the disks in the placement groups synced.

The default jerasure erasure code plugin is being used with the Reed, Solomon, Vandermonde algorithm [2]. CERN had previously performed testing comparing the ISA and jerasure plugins [3] and found little difference between the two.

### 2.3. CRUSH Map and ruleset

In Ceph, the CRUSH Map[4] describes the layout of the storage cluster. The CRUSH ruleset is a method to select which disks make up a placement group. The CRUSH ruleset that has been chosen for Echo is to require every disk in a placement group to be on a separate host. This means that the loss of an entire storage node will not cause any data to be lost. In future, this should mean that a hardware problem on a single storage node will no longer need to trigger a call out. It also means that rolling upgrades of the cluster can happen in a manner that is transparent to the users.

More complicated CRUSH maps and rulesets were considered but it was decided they did not offer sufficient benefit for the increased complexity they introduced to be worth deploying. One possibility that was considered was to spread the disks in each placement group across racks. Unfortunately it is not possible to simply specify that no more than 3 disks in a placement group can be on one rack, instead one must specify a series of rules. If the cluster had eleven or more racks it would be simple to specify that each disk needed to be on a different rack, however as the cluster is currently composed of 8 racks, the rule would need to specify that 6 racks are selected, and then two disks are selected from each rack, with one disk being discarded from the placement group. An analysis was performed on how the placement groups were spread across the cluster and it was found that only 0.05% of data would be compromised in the event an entire rack failed.

### 3. GridFTP and XrootD Plugins

One of the first steps when designing Echo, was to identify what features were necessary to support the LHC experiments. In order to maximise the reliability while minimising the amount of development work needed, it was decided to provide only the minimum required feature set necessary for the LHC experiments to run their production work. It was also decided that, if necessary, it would be better to try to change the experiments workflows than add complicated

| m = 2 | 80 | 83 | 86 | 88 | 89 |
|-------|----|----|----|----|----|
| m = 3 | 73 | 77 | 80 | 83 | 84 |
| m = 4 | 67 | 71 | 75 | 78 | 80 |

**Table 1.** Table showing the percentage of usable storage with different EC profiles.
features to Echo. The RAL Tier-1 supports all four LHC experiments and has a dedicated contact for ATLAS, CMS and LHCb so is well placed to do this.

There has been a long standing desire within the WLCG community to remove the need to run an SRM. It was therefore decided very early on in the development process to not provide one at all for Echo. It is also known that the LHC experiments have central database which record the amount of data stored at each site. Complicated space reporting is being replaced by a JSON file that can be easily downloaded and that simply lists the total space allocated to the particular experiment.

In the long term RAL intends to provide industry standard protocols (S3 / Swift) for experiments to use. However, after a brief consultation with the LHC experiments it was realised that on the time scale available, they would be unable to properly utilise storage exclusively via these protocols. There is, however significant interest in supporting these new protocols and considerable development work is being put towards these goal.

The LHC experiments make use of a range of storage protocols, however it was found that only GridFTP and XRootD are necessary. GridFTP is required by ATLAS, CMS and LHCb for WAN transfers mediated by the FTS service. XRootD is the preferred / required protocol by all the LHC experiments, for accessing data by local jobs. Some experiment jobs use XRootD simply to transfer the data from the storage element to the local disk of the worker node, while other jobs read data directly from the storage as it is needed.

An XRootD plugin for Ceph has been developed by CERN. In order to build the XRootD plugin, CERN had also developed a method for striping data into Ceph using librados. This was called libRadosStriper and became part of the official Ceph build in Hammer. Any plugin that made use of libRadosStriper would be able to read and write data from any other plugin also using it. In January 2015, as a proof of concept, a GridFTP plugin was also created using libRadosStriper. RAL took over the development of the GridFTP plugin in April 2015 and has had a stable version since October 2016.

3.1. Authentication and Authorisation

While it would be necessary to develop authentication and authorisation for the GridFTP plugin, various tools to do this already existed within XRootD. To minimise the amount of development work required it was decided to find a configuration that worked with XRootD and then develop just this for GridFTP. It was decided to use a Gridmap file for authentication and an authDB for authorisation. Figure 2 is a diagram showing how the authentication and authorisation works.

A conscious decision was made to use a Gridmap file over VOMS roles. Castor uses a Gridmap file and we wanted to keep the same mechanism in place. There have been no operational issues observed with using the Gridmap file with Castor.

The authDB is a simple text file that contains a list of users followed by directories and the authorisation they have on it. The Ceph object store does not provide a file system so the ‘directories’ were actually just the first part of the object ID. Numerous small changes had to be made to the XRootD code to account for the fact that Ceph is an object store rather than a filesystem. e.g. there was an assumption that all path names would begin with a “/”.

3.2. Buffer re-assembly for FTS transfers

The GridFTP protocol can transfer files either as a single stream of data or as multiple streams of data. LAN transfers usually use the single stream mode while, for WAN transfers between sites (normally using the FTS service), the multiple stream method is used. With data arriving in multiple streams a method to re-assemble the data in the correct order before copying it into the storage is required. Similar work had been done for the CMS Tier 2 site in Nebraska for their Hadoop storage[5].
Figure 2. Figure showing how data is written to Echo using either the GridFTP or XrootD plugin

The data is re-assembled using two equally sized buffers. Once the first buffer has been filled, it is written to Ceph and then it becomes the second buffer. Even with multiple streams most data arrives quite close together and so the buffers can be kept small. Should data arrive that is further ahead than allowed by the second buffer, then the transfer will fail, however there is a patch available for GridFTP that slows down the streams if they start getting too far ahead. The observed failure rate for this problem is extremely low and even if there is a failure, the retry is likely to succeed.

For performance reasons it is very important to ensure that the size of the objects being written by the plugin to Ceph align with the libRadosStriper settings. Testing has shown that Ceph provides maximum throughput when transferring objects of around 8 MB. A file written to an erasure coded pool, with $k = 8$, will be broken into 8 equally sized objects, therefore for optimal performance the file should be 64MB. This is possible to do by configuring libRadosStriper to split large files into 64MB stripes. The buffers that are reassembling the data to write to Ceph should therefore be an integer multiple of 64MB. The buffer size chosen is 128MB. By ensuring that these object sizes are all aligned, an individual write will progress around a factor of 10 faster, at around 70MB/s as opposed to 6−7MB/s without alignment.

4. Summary
This paper describes the deployment of a new Ceph backed storage service call Echo at the RAL Tier-1. The major challenge in the deployment of the cluster has been getting the erasure coded pools to work reliably and work is still ongoing. Alongside the deployment of the cluster, GridFTP and XRootD plugins have also being developed. These plugins have been designed to work interchangeably and make use of a Gridmap file and authDB for authentication and
authorisation. To maximise throughput it is essential to ensure that the objects being written to Ceph have the correct size.

Acknowledgments
The authors would like to acknowledge the support provided by the GridPP Collaboration, CERN and STFC. In particular we would like to thank: Sebastien Ponce, who developed the XRootD plugin for Ceph and provided significant support during the development of the GridFTP plugin. Brian Bockelman who assisted with the development of the buffer re-assembly code as well as offering advice with XRootD. Dan van der Star who provided a huge amount of advice on running a Ceph cluster. We would also like to thank Tom Byrne and George Ryall who previously worked on Ceph at RAL.

References
[1] Yahoo Ceph Storage https://www.nextplatform.com/2015/04/16/inside-the-ceph-exascale-storage-at-yahoo/ accessed Jan 2017.
[2] J Plank, K Greenan 2014 Jerasure: A Library in C Facilitating Erasure Coding for Storage Applications Technical Report UT-EECS-14-721 University of Tennessee
[3] CERN Ceph Testing https://cds.cern.ch/record/2015206/files/CephScaleTestMarch2015.pdf accessed Jan 2017.
[4] S Weil, S Brandt, E Miller, C Maltzahn 2006 CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data Proc., of the 2006 ACM/IEEE Conf., on Supercomputing Article No. 122
[5] B Bockelman 2009 Using Hadoop as a Grid Storage Element J. Phys.: Conf. Ser. 180 012047