PROCEEDINGS OF SCIENCE

First Experiences with Ceph on the WLCG

Rob Appleyard [1]
Science and Technology Facilities Council, Scientific Computing Department
STFC Rutherford Appleton Laboratory, Harwell Oxford, OX110QX UK
E-mail: Rob.Appleyard@stfc.ac.uk

Shaun de Witt
Science and Technology Facilities Council, Scientific Computing Department
STFC Rutherford Appleton Laboratory, Harwell Oxford, OX110QX UK
E-mail: Shaun.de-Witt@stfc.ac.uk

James Adams
Science and Technology Facilities Council, Scientific Computing Department
STFC Rutherford Appleton Laboratory, Harwell Oxford, OX110QX UK
E-mail: James.Adams@stfc.ac.uk

Much interest has been shown in the Ceph storage platform and this paper presents an introduction to Ceph, object stores, and the initial results of testing Ceph for possible use at large sites. We look at its performance when used as a disk cache for a traditional HSM and compare its performance with competing systems. In addition, we present a model comparing hardware usage for various Ceph configurations with RAL's existing CASTOR instance and discuss future plans for Ceph deployment at RAL.

The International Symposium on Grids and Clouds (ISGC) 2013
March 17-22, 2013
Academia Sinica, Taipei, Taiwan

1 Speaker

PoS(ISGC2014)037

1. Introduction

The Science and Technology Facilities Council's (STFC) Scientific Computing Department (SCD) hosts the UK's WLCG[1] Tier 1 computing centre for the GridPP[2] project at the Rutherford Appleton Laboratory (RAL). The RAL Tier 1 serves all four major LHC experiments, and the same building hosts computing services for various other users, including T2K, MICE, na62, and local facilities, such as the ISIS neutron spallation experiment and the Diamond Light Source.

The RAL Tier 1 (henceforth referred to as 'the Tier 1') currently runs the CASTOR (CERN Advanced STORage)[3] storage system for both disk-only storage and as a cache in front of a tape robot. At the time of writing, the total capacities of the Tier 1 CASTOR/Tape system are 17 petabytes (PB) of disk storage and 13PB of tape. The Tier 1's intentions are to maintain CASTOR in its optimal role as a system for managing the disk cache in front of the tape system and to find another system that is better suited for the demands of disk-only storage.

CASTOR is a distributed hierarchical mass storage system developed in-house at CERN that dates back to 1999 in its current form. It has been running in production at RAL since 2005. The RAL implementation is a PB-scale storage platform that uses a central Oracle database to manage transactions. There are no built in measures for data redundancy in the event of hardware failure, and so to remain fault-tolerant, sites must use some form of hardware RAID underneath CASTOR. RAID levels vary from site to site depending on budget and user requirements. RAID 1 is used at CERN, while RAL uses RAID 6.

Experience of running the current configuration has led to an intention on the part of the Tier 1 to retain CASTOR for management of the disk cache in front of the tape system and to move to a better-suited system for disk-only storage. A previous project, undertaken in 2012-13, sought to evaluate a variety of storage systems for use by the Tier 1[2], but its results were inconclusive. A decision was taken to preserve the status quo while monitoring the two most promising candidates, namely Ceph and HDFS.

Moving forward to the present day, based on observation of the project's progress, the Tier 1 now intends to create a petabyte-scale Ceph test instance for evaluation by both the local administration team and by LHC experiments. The reasons for the Tier 1's choice will be detailed in the following sections.

2. What is Ceph?

'Ceph' refers to set of technologies that form an open-source, highly scalable distributed storage system. The system's fundamental element is a collection of Object Storage Daemons (OSDs), of which there are typically one per underlying drive. The OSDs store data, co-operatively handle the various data management tasks necessary for the system (replication,

2 This included: dCache[4], OrangeFS[5], HDFS[6], Lustre[7], Ceph[8] and EOS[9]

recovery, rebalancing), and send monitoring information on both themselves and other daemons to the Ceph Monitors. The Ceph Monitors, in turn, maintain maps of the state of the Ceph cluster and compile historical information on state changes of the various elements.

The location at which a given object is placed is determined using an algorithm known as CRUSH. CRUSH uses the map of the cluster maintained by the Ceph Monitors to uniformly distribute data across the OSDs into distributed 'placement groups' in a reproducible pseudo-random fashion. The system is able to mitigate against common storage failure modes (such as a multi-drive failure on a single node) by maximizing the logical distance between each element of a placement group, thus placing the replicated copies of data in physically different locations to the originals.

In the event of hardware failure (e.g. a failed hard drive), the system demonstrates 'self-healing' properties, as the monitors determine that there are an insufficient number of replicas of the objects that were on the failed drive, and replicate the files in question until a sufficient number of replicas exists.

The final element of the system is the Ceph Filesystem (CephFS) component provides a file system above the object store using Ceph Metadata Servers to store metadata. The reason that CephFS uses dedicated metadata servers is performance-related, as they improve the system's performance for simple POSIX query commands like ls and find.[10]

2.1 Object Stores

The term 'object store' has been used above, and an explanation is in order. An object store is a system that manages each 'chunk' of data that goes into it as an object or objects, with an associated identifier. There is thus a flat structure with no hierarchy, merely a set of IDs and objects. This allows the system to abstract away the lower levels of the storage system and manage a pool of storage that may consist of very many distinct elements as if it were a single large element, and distribute the metadata separately from the data to allow improved performance. One can then impose a filesystem or namespace above the object store as desired. In Ceph's case, the object store is known as RADOS and the filesystem is known as CephFS. This structure allows systems to potentially remain performant even when scaled to sizes in the multi-petabyte range.

3. Why is the Tier 1 interested in Ceph for storage?

The Tier 1's requirements for a storage system are specific. A relevant selection of the agreed-upon set of requirements from the SCD 2012-13 storage evaluation project includes:

- The storage system SHOULD be independent of restrictively licensed software (such as a licensed database or scheduler), although the solution itself may be licensed.
- The storage system SHOULD be resilient to hardware failure, on levels of disk, node and memory.
- The storage system MUST make effective use of existing hardware
- The storage system MUST NOT unduly restrict hardware purchasing.

- The storage system MUST be competitive with CASTOR in terms of cost.

We believe that Ceph has the potential to fit all of these requirements, as detailed below.

3.1 Licensing Requirements

Ceph is an open-source project developed primarily by Inktank[10]. Inktank use a two-tier model, where the open-source project is freely available to all, but the company offers paid subscription services for support, design and deployment consulting. Individual contributors retain ownership in the same manner as contributions to the Linux kernel, making a future license change very difficult. A Ceph client has been incorporated into the Linux kernel since 2010[10]. This model is entirely compatible with our requirement for freedom from restrictive licensing.

3.2 Resilience

The resilience of storage systems against hardware loss is often a trade-off between performance and cost. The more replicas of a file one keeps, the less likely it is for a catastrophic failure to wipe out all of the replicas. At the time of writing, Ceph offers replication as its only scheme for resilience, but the developers have announced that an erasure coded back-end will be included in the latest release [10]. All of these would satisfy the requirement for the system to be resilient to hardware failure, although at different points on the cost-resiliency scale.

3.3 Use of existing hardware and hardware cost

Ceph is a hardware-agnostic system, and broadly speaking its hardware requirements are similar, but not identical to those of CASTOR. The Tier 1 CASTOR disk-only system uses large (16-24 drive) storage nodes with 3-4TB drives, and provides resiliency to disk failures by running hardware RAID 6 arrays underneath the file system. This requires specialised RAID cards to implement, something that is not necessary for Ceph. However, as will be shown later, this node profile offers a favourable price/terabyte ratio regardless of the storage system used, and to use similar nodes for Ceph would require some sort of hardware disk management in any case. Thus Ceph satisfies the requirements for good use of existing hardware and the lack of restriction on hardware purchasing.

The question of cost is one that does bear further consideration. To this end, an investigation was carried out into the ratio of raw to usable storage for CASTOR and a variety of possible Ceph configurations given a variety of different node sizes. The ratio for CASTOR configurations was determined by the authors' assessment of what they would consider to be reasonable disk configurations (in terms of the numbers of parity disks and hot spares) for a single CASTOR node with the given number of drives (the investigators acknowledge that this is not a precise measurement), plus 1% the overhead required by CASTOR, and the Ceph configurations were a variety of possible Ceph deployment scenarios. These were:

- Ceph running on current RAID 6 nodes bought for CASTOR with zero replication at the software level.
- Ceph configured to keep 1 original copy and 2 replicas.
- Ceph configured to keep 1 original copy and 1 replica.
- Ceph configured so that 1 drive in each placement group of 18 drives contained parity information.
- Ceph configured so that 2 drives in each placement group of 18 drives contained parity information.

The results of this exercise are shown in Fig 1.

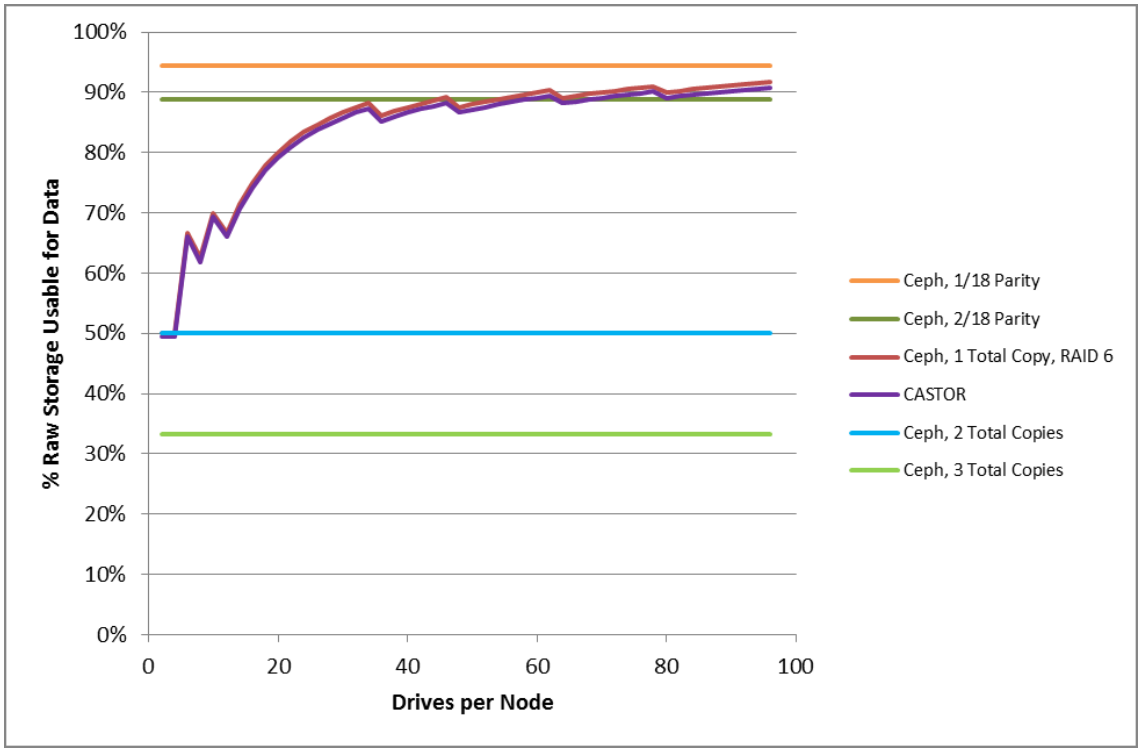

Fig 1: Graph showing the relative amounts of usable storage provided by CASTOR and a variety of different Ceph configurations for various node sizes.

As shown, non-RAID configurations are agnostic on the number of drives per node. The graph also shows that erasure coding is a necessity if the Tier 1 were to run Ceph within CASTOR's budget, and that CASTOR is only competitive with the proposed erasure coded Ceph configurations given very large nodes (>34 drives/node).

This is not a comprehensive model of costing – hardware cost is far from the only cost inherent in running a data centre. Neglected are cooling, power (although both of these could be expected to scale linearly with the number of nodes bought), and significantly, staffing. The staff cost of maintaining a Ceph cluster is not easily estimated; this is something that the Tier 1 will find out through experience while running the proposed Ceph test instance.

4. Future plans at the Tier 1

As previously noted, the Tier 1 intends to prepare a ~1PB test instance of Ceph with the intent to conduct a realistically-scaled evaluation of the file system. The initial configuration will use one original copy and one replica. A shift to erasure coding will be made in time for performance testing to be carried out. If all goes well, then the hope is to deploy a production instance as a replacement for CASTOR disk-only during early-to-mid 2015. This is a risky operation; the scale of the change to operations will be large, and as shown above the change is dependent on an as-yet unreleased feature in order to be cost-effective. If significant problems arise with the erasure coding then these plans will be disrupted.

4.1 Non-High Energy Physics Use Cases

In a separate development, a project is currently running within SCD to develop a private cloud for both data management development and scientific use by staff and researchers at STFC. This cloud has requirements for persistent and shared storage that match exceptionally well to Ceph's RBD (Rados Block Device) and RADOSGW (Object Gateway) systems.

RBD allows Ceph to provide a striped set of objects as if they were a block device. Support in the Linux kernel and QEMU (Quick EMUlator)[12] allow these to be used as if they were block devices on a SAN (Storage Area Network). We intend to use this to provide virtual machines with a high-availability decoupled storage backend.

RADOSGW provides a translation layer between Ceph's internal object APIs and industry standard interfaces such as Amazon S3 or OpenStack Swift. We intend to deploy these gateways to provide S3 persistent shared storage buckets for our private cloud users.[10]

5. Conclusion

This paper gives an account of the Ceph object storage platform and argues that it offers a potentially compelling platform for future usage by the RAL Tier 1 site. It provides an explanation of object storage and why this is a useful technology for storage systems. It shows that Ceph meets many of the requirements for the RAL Tier 1 storage site, including those relating to licensing, resilience and cost. Finally, it discusses other use cases for Ceph and sets out a future upgrade path for the Tier 1.

References

[1] http://wlcg.web.cern.ch/

[2] D. Britton, D, et al. *GridPP: the UK grid for particle physics*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367.1897 (2009): 2447-2457.

[3] J.P. Baud, et al. *CASTOR status and evolution,* arXiv preprint cs/0305047.

[4] Fuhrmann, Patrick. *dCache, the Commodity Cache*, MSST. 2004.

[5] Bonnie, Michael Moore David, et al. *OrangeFS: Advancing PVFS*.

[6] Borthakur, Dhruba. *HDFS architecture guide*. HADOOP APACHE PROJECT http://hadoop. apache. org/common/docs/current/hdfs design. pdf (2008).

[7] Braam, Peter J. *The Lustre storage architecture* (2004).

[8] Weil, Sage A., et al. *Ceph: A scalable, high-performance distributed file system*, Proceedings of the 7th symposium on Operating systems design and implementation. USENIX Association, 2006.

[9] Peters, A. J. *The EOS disk storage system at CERN* ACAT conference proceedings, 2011.

[10] http://www.inktank.com/

[11] Bellard, Fabrice. *QEMU, a Fast and Portable Dynamic Translator.* USENIX Annual Technical Conference, FREENIX Track. 2005.