Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Current Status of the Ceph Based Storage Systems at the RACF

To cite this article: A. Zaytsev *et al* 2015 *J. Phys.: Conf. Ser.* **664** 042027

Current Status of the Ceph Based Storage Systems at the RACF

A. Zaytsev[1], H. Ito[1], C. Hollowell[1], T. Wong[1], T. Rao[1]

[1]Brookhaven National Laboratory, Upton NY 11973, USA

Email: hito@bnl.gov

Abstract. Ceph based storage solutions are becoming increasingly popular within the HEP/NP community over the last few years. With the current status of Ceph project, both object storage and block storage (RBD) layers are production ready on a large scale, and the Ceph file system storage layer (CephFS) is rapidly getting to that state as well.

This contribution contains a thorough review of various functionality, performance and stability tests performed with all three (object storage, block storage and file system) levels of Ceph by using the RACF computing resources in 2012-2014 on various hardware platforms and with different networking solutions (10/40 GbE and IPoIB/4X FDR Infiniband based). We also report the status of commissioning a large scale (1 PB of usable capacity, 4k HDDs behind the RAID arrays by design) Ceph based object storage system provided with Amazon S3 complaint RadosGW interfaces deployed in RACF, as well as performance results obtained while testing the RBD and CephFS storage layers of our Ceph clusters.

1. Introduction

Over a last decade, the RHIC and ATLAS Computing Facility (RACF) at BNL have been using conventional storage system that provides the access to its namespace operation as well as the content of the data. During the same period, the new type of storage, which separates the operations of the data from those of its namespace, has emerged. These new kinds of storage systems operate on a principle that the storage service should provide the stable and scalable service accessing the content of data without performance limitations of manipulating the metadata associated with the namespace. They are commonly called object storage systems, emphasizing the data over its metadata, such as Amazon S3 [1], GlusterFS [2], and Ceph [3].

At RACF, Ceph based storage clusters were chosen to be tested for the following reasons. At first, it is free software for us to deploy on top of our older or even retired hardware without any licensing and other financial obligations. Secondly, based on reliable autonomic distributed data store (RADOS) [4], Ceph provides three distinctive use cases: object storage with S3 interfaces, block devices and a file system. The support of S3 APIs, which were originally developed by Amazon to access their services, reduces the learning curve of using the new type of storage. The RADOS block device layer (RBD) [5] can be used as any part of existing mounted storage system. Finally, the Ceph file system (CephFS) [6] allows us to study the use case of a resilient and scalable distributed file system for possible applications in our HEP specific data storage environment.

2. Hardware

At the moment the RACF is hosting two Ceph clusters of total usable capacity of 1 PB (assuming factor of 3 data replication) built out of similar hardware components, but provided with different types and layouts of storage interconnect and network interconnect between the cluster head nodes. These hardware components essentially are:

1. Twenty nine Sun Thor x4540 storage servers each enclosing 48 1 TB 7.2krpm SATA HDDs and provided with a single optical 10 GbE uplink, up to four 4 Gbps Fibre Channel ports and 32 GB of RAM installed. In our Ceph environment the raw capacity of these units was exported to the Ceph cluster head nodes (running Linux OS) as a single ZFS pool via iSCSI. All the Sun Thor servers are deployed under Solaris 10 10/08 OS in our setup.
2. Fifty six Nexsan SataBeast disk arrays each enclosing 42 1 TB or 2 TB 7.2krpm SATA HDDs and configured with one embedded hardware RAID controller provided with two 4 Gbps Fibre Channel ports. Only 5 out of 56 Nexsan SataBeast arrays are loaded with 2 TB drives while the majority of them are loaded with 1 TB drives. Depending on the Ceph cluster to which a particular Nexsan array belongs, its raw capacity is exported to the cluster head node either directly via Fibre Channel (as a block device) or via iSCSI, using one of the Sun Thor x4540 storage servers as an iSCSI target export node. Each Nexsan array is configured as a single RAID set containing 40 HDDs in RAID-60 plus two hot spares with four volumes of equal capacity of 9.5 TB deployed on top. The volumes are exported via both 4 Gbps Fibre Channel uplinks two one of the Ceph cluster head nodes (two LUNs associated with each FC link configured into Point-to-Point mode).
3. Sixteen Dell PowerEdge R420 servers each provided with 48 GB of RAM, three 1 TB 7.2krpm SATA HDDs configured in a hardware RAID-10 plus one hot spare for the OS, one extra 1 TB 7.2krpm SATA HDD or 500 GB SATA SSD (depending on which Ceph cluster these nodes are deployed in), plus a dual port Mellanox Virtual Protocol Interconnect (VPI) enabled 4X FDR IB/40 GbE/10 GbE PCI-E card configured in dual port 10 GbE mode. These nodes are used as the Ceph cluster RadosGW/S3 gateways nodes and also the Ceph cluster head nodes (exclusively for the Federated Ceph cluster, which historically was commissioned earlier). These servers are deployed under Scientific Linux (SL) 6.5 x86_64 with the custom built Linux kernel 3.19.1 (3.14.22 for the Federated Ceph cluster).
4. Eight Dell PowerEdge R720XD servers each provided with 80 GB of RAM, ten 4 TB 7.2krpm SAS HDDs configured in a hardware RAID-10 plus two hot spares, two extra 250 GB SATA SSDs to be used exclusively by the Ceph OSD journals, one dual port Mellanox VPI 4X FDR IB/40 GbE/10 GbE PCI-E card configured into the mixed 4X FDR IB/optical 40 GbE mode, and 5 additional Emulex and Qlogic PCI-E Fibre Channel cards featuring four 8 Gbps FC ports and eight 4 Gbps FC ports. These nodes are deployed under Scientific Linux (SL) 6.5 x86_64 with the custom built Linux kernel 3.19.1 and serving as the cluster head nodes for the RACF Main Ceph cluster. With Nexsan storage arrays attached to all 12 Fibre channel ports each of these nodes is experimentally tested to be able to push up to 3 GB/s of data to the disks and up to 7 GB/s through both 40 GbE and IPoIB/4X FDR network uplinks simultaneously.
5. Two Mellanox SX6018 4X FDR IB/40 GbE VPI switches provided with 18 QSFP ports each used for providing each of RACF Ceph clusters with the high speed/low latency private interconnect fabric used for handling OSD-to-OSD communications. Both switches are configured in Infiniband mode (4X FDR IB, 56 Gbps per port) on all ports and each of them is hosting its own instance of the Infiniband Subnet Manager (IB SM) for each of these fabrics. All the head nodes connected by the Infiniband fabrics in both of our Ceph clusters are using additional IPoIB layer to make it possible for the versions of Ceph lacking RDMA support to exploit the benefits of the Infiniband interconnect [7].
6. Dell Force10 Z9000 switch featuring 32 QSFP ports used as the main Ethernet switch for both clusters, with each of these ports configured either to 40 GbE mode or 4x 10 GbE

mode. The Z9000 switch was fully dedicated to the Ceph installation and carried all the internal network traffic for the RACF Ceph clusters (all but OSD-to-OSD communications), serve as network transport for the raw storage exported from Sun Thor servers via iSCSI and enabled all the client connectivity to the Ceph MON and RadosGW components for clients inside and outside BNL. The transition is now on-going to move all the 40 GbE uplinks of the RACF Main Ceph cluster head nodes to the central Arista 7508E switch provided by the RACF facility.

Thus, both of our Ceph installations together have 3.9k spinning drives (95% of which are 1 TB SATA HDDs), which with all RAID and partitioning / file system overheads results in 3 PB of raw storage capacity available for out Ceph clusters. Our two Ceph cluster installations are configured as follows:

1. RACF Federated Ceph Cluster consisting of six head nodes mounting 29 storage volumes from Sun Thor servers via iSCSI/10 GbE and connected by 10 GbE (public) / 56 Gbps (private Infiniband) uplinks, plus two RadosGW/S3 gateways. Because of the limited performance of iSCSI links and throughput limitation of the 10 GbE links this installation is limited to 11 GB/s of maximum data throughput on the storage backend and 7.5 GB/s of aggregated traffic on the client side. This Ceph cluster is deployed with Ceph 0.80.1 Firefly release and factor of 3 data replication for all the pools providing approximately 0.4 PB of usable disk space. Each head node in this cluster is configured to carry one MON, one MDS and up to eight OSD instances. Because of the limited aggregated amount of RAM (288 GB) this cluster can safely operate with up to 48 OSD instances, but the maximum we ever used with this installation historically never exceeded 45 OSDs (6.4 GB RAM per OSD). All the OSDs are deployed with XFS with the OSD journals placed either on the same file system or the dedicated SSD drive.
2. RACF Main Ceph cluster consisting of eight head nodes mounting volumes via Fibre Channel from 48 Nexsan disk arrays (8 more arrays can be connected to this setup in the future) and eight more RadosGW/S3 gateways. Because of the limited performance of the hardware RAID controllers on the Nexsan arrays this setup is limited to 26 GB/s of maximum data throughput at the storage backend. All the network interconnects used on the head nodes in this setup (40 GbE on the public cluster network and 56 Gbps on the private Infiniband interconnect for each head node) are completely non-blocking (about 64 GB/s if both interconnects are loaded simultaneously) with respect to that storage performance limitation. This Ceph cluster is deployed with Ceph 0.94 Hammer release and factor of 3 data replication for all the pools providing approximately 0.6 PB of usable disk space. Each head node is configured to carry one MON, one MDS and up to 13 OSD instances. The aggregated amount of RAM on the head nodes of this Ceph cluster (640 GB) permits the safe operations of up to 108 OSD instances. All the OSDs are deployed with XFS with the OSD journals placed on the dedicated SSD drive (up to 7 journal partitions per SSD).

The networking and storage layout of both RACF Ceph clusters is show in Figure 1. More details on the configuration and performance of these systems can be found in the reports [8] and [9].

3. Ceph Object Gateway

Ceph Object Gateway (RadosGW) provides two interfaces to Ceph storage clusters: S3 and Swift. The S3 Interface provides the compatible APIs used by Amazon S3 while the Swift interface is compatible with OpenStack Swift APIs [10]. The common interface used in Amazon and Ceph makes the use of S3 APIs are convenient for developers to create new and adapt the exiting application. The object store does not provide block devices and/or file system by itself – it only provides containers, call buckets, for data to be stored, requiring the use of external file catalog to retrieve/modify the content of the containers.

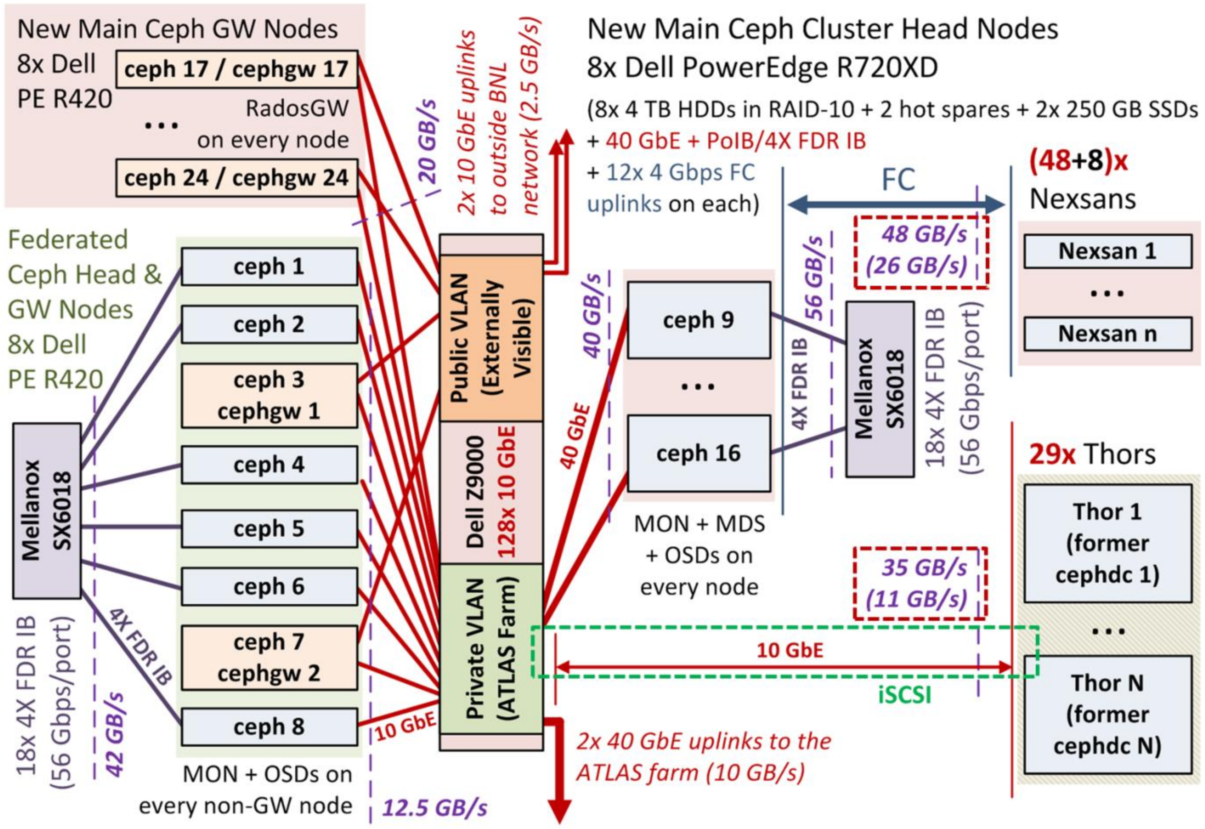

Figure 1. Current layout of two Ceph clusters of RACF: the Federated Ceph cluster (0.4 PB of usable capacity) and the Main Ceph cluster (0.6 PB of usable capacity).

ATLAS experiment has already started to implement the use of S3 storage in the computing framework called ATLAS Event Service [11] as a data store for a very large number (~10^9) of small data objects (~a few MB each), utilizing the advantage of not having a file system.

4. Ceph Block Device

Ceph (RADOS) block device (RBD) provides one of the most convenient ways to integrate the Ceph storage clusters into any existing storage services. It provides the regular mountable device to OS. Once such a device is mounted, it can be formatted to be used as a normal space.

One example of using Ceph block device on our existing dCache system is shown in Figure 2. First, multiple RBD objects are created in the Ceph storage cluster. Then, they are provided to multiple hosts where they can be mounted and formatted to be used as storage on those hosts, allowing to be added as dCache [12] storage pools. With dCache as the top layer of the storage service, Ceph storage immediately becomes a part of any production service, providing all necessary, existing access interface without complication of learning how to access a new type of storage. One disadvantage of the using RBDs is that it does not fully utilize the redundancy of Ceph storage system by only allowing single-mount of RBD on a particular host, resulting in the reliance in the stability of RBD-mounted host system to serve the data regardless of the availability of the Ceph storage cluster.

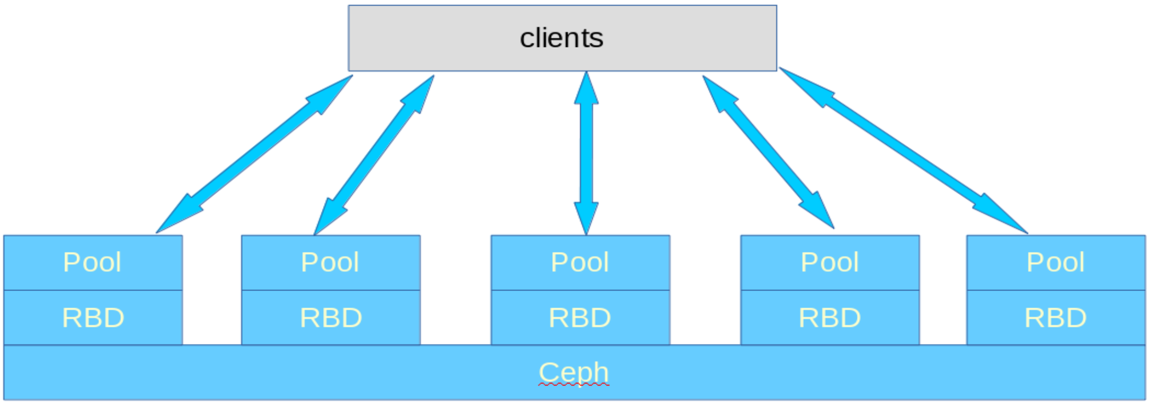

Figure 2. Ceph RBDs as dCache pools.

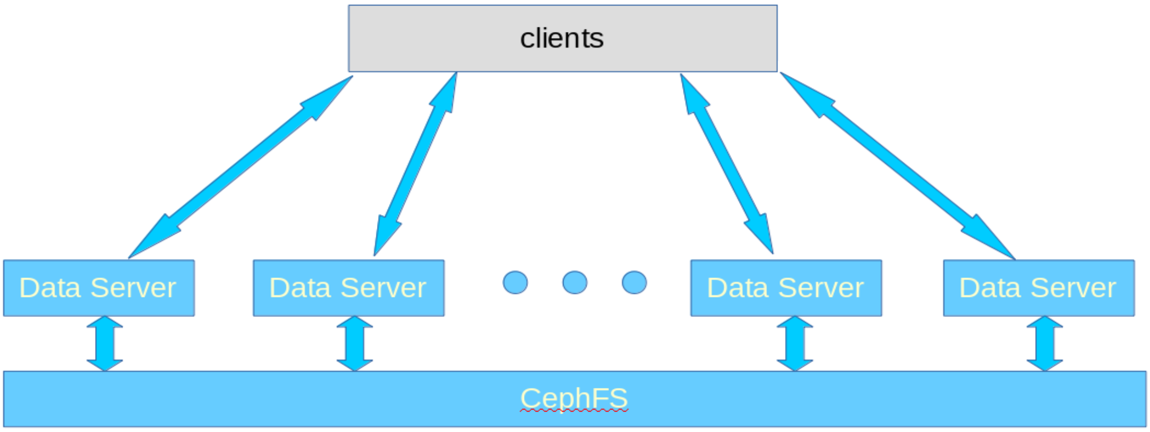

Figure 3. CephFS provided with additional layer of the with XRootD data servers (gateways).

5. Ceph File System

Ceph file system, CephFS, is a POSIX compliant (with minor exceptions summarized in [13]) file system built on top of the Ceph storage cluster mention in the previous sections. Using the Ceph storage technology, the CephFS provides the familiar interface to users without any complication of learning new APIs with all the benefits of Ceph storage cluster. The distributed nature of the CephFS makes the service to be reliable and resilient. In addition, unlike the RBDs mentioned in the last section, the same storage can be accessed on the multiple hosts, making some interesting use case.

Since the same storage can be presented to multiple hosts, when the above layer of a storage service are state-less in serving the data, it can be configured to increase the reliability of the service horizontally by adding any additional hosts. For example, one of the candidates for such a layer of a state-less storage service on top of CephFS is the XRootD data access service [14]. Figure 3 illustrates this configuration. CephFS provides a common file system. This same name space is mounted in multiple XRootD data server hosts. Since any data can be accessed through any of the data servers, CephFS with XRootD can provide a very reliable storage services with horizontal scalability.

6. Performance Tests with CephFS

6.1. Test setup configurations

The throughput tests were conducted using two setups:

- *Setup S1* consisting of 74 client hosts each provided with 1 GbE network interfaces plugged to a non-blocking networking core of the RACF ATLAS Farm. Each of the client nodes was deployed under Scientific Linux (SL) 6.5 x86_64 OS provided with the custom built Linux kernel 3.19.1 and was mounting CephFS directly from the Main Ceph cluster. The maximum network throughput was limited by 74 Gbps = 9.25 GB/s on the client side.
- *Setup S2* making use of up to 100 client hosts each provided with 1 GbE network interfaces plugged to a non-blocking networking core of the RACF ATLAS Farm and four or eight 10 GbE attached XRootD data server (XRootD gateway) nodes. Each of the client nodes was deployed under Scientific Linux (SL) 6.5 x86_64 OS with standard kernel packaged with the distributive. Each gateway XRootD gateway was deployed under Scientific Linux (SL) 6.5 x86_64 OS provided with the custom built Linux kernel 3.19.1 and was mounting CephFS directly from the Main Ceph cluster and making it contents available to the clients via locally installed instances of the XRootD server v4.2.3 (x86_64). The maximum achievable network throughput was limited by 40 Gbps = 5.0 GB/s or 80 Gbps = 10 GB/s on the level of the XRootD gateway nodes, depending on the number of the gateway nodes used.

Simple replication factor of 3 was configured for both *data* and *metadata* pools backing up the CephFS instance used both test setups. The following additional performance optimizations were used on the side of the Ceph cluster involved:

1. Clustered configuration consisting of eight Ceph MDS components working together was added to the Main Ceph cluster (one MDS per physical Ceph cluster head node) in order to maximize performance of this subsystem.
2. Since the tests involves a relatively small set of large files (up to 100 files of up to 40 GB in size), the following extended attributes (*xattrs*) were set for all the directories and files deployed in CephFS: *stripe_unit = 16777216, stripe_count = 36, object_size = 4194304*. The optimal value for the *stripe_count* parameter was chosen as the number of OSDs in the cluster (108) divided by the data pool replication factor (3). The optimal values for the *stripe_unit* and *object_size* parameters were determined experimentally while trying to maximize performance of I/O operations with a single large file in a series of preliminary CephFS performance tests with one client node mounting the CephFS directly from the cluster involved.

6.2. Read and write throughput with test setup S1

The largest scalability test for CephFS directly mounted on the client nodes was performed with the group of 74 1 GbE network attached clients, each mounting the CephFS directly from the Main Ceph cluster. The test was conducted in the following 3 consequent stages:

1. Measure the network bandwidth available between the clients and the Ceph cluster head nodes with iperf3 in one and several simultaneous iperf3 TCP session established on every client node. The maximum aggregated network throughput of 8.9 GB/s was reached in these tests with two sessions per client. This corresponds to 96% of the throughput limitation on the client side, showing that there is no obstruction on the network level between the client nodes and the Ceph cluster.
2. Write individual files each 40 GB in size to CephFS from all 74 client nodes. Maximum sustained bandwidth of 4.5 GB/s (5.5 GB/s peak) was observed in both of these tests in the relatively short bursts of traffic up to 10-15 minutes long.
3. Read one of the files previously written to CephFS in the second stage of the test from all 74 nodes. Secondly, read a distinct file from all client nodes simultaneously. The OS level read buffers were flushed on the client hosts before every step of the test. Maximum sustained

bandwidth of up to 8.7 GB/s was observed, which corresponds to 98% of the value measured with iperf in the first stage of the CephFS scalability test.

Thus, the CephFS read performance was limited only by the network capabilities of the client nodes, while the sustained write performance observed was about 50% of the network throughput limit on the client side. Further Ceph cluster level optimizations and similar scalability tests are needed to fully understand the ways in which the write performance can be further improved in our setup.

6.3. Read and write throughput with test setup S2

To measure write throughput of CephFS, a set of distinct files with relatively large size (approximately 3.6 GB per file) from BNLs production dCache storage system were used as the data source. They were written to CephFS instance in order to be later retrieved via XRootD protocol. Figure 4 (top) shows the network activity during the one hour long test. The network values in the figure include network activity within the Ceph storage cluster associated with internal replications. The actual write throughput values are about 1/3 of those indicated by the green line. Although the tests were conducted repeatedly with varying number of clients, the maximum write throughputs did not change, indicating the limit in the current configuration of the Ceph cluster. Additionally, we have observed that, as more data is written, the slow degradation of performance was seen Figure 4 (bottom). It might be attributed to the performance issues associated with the Ceph Metadata Server (MDS), even though the clustered MDS configuration was used. Even faster degradation was observed with just a non-clustered MDS configuration with just one active Ceph MDS component in the cluster.

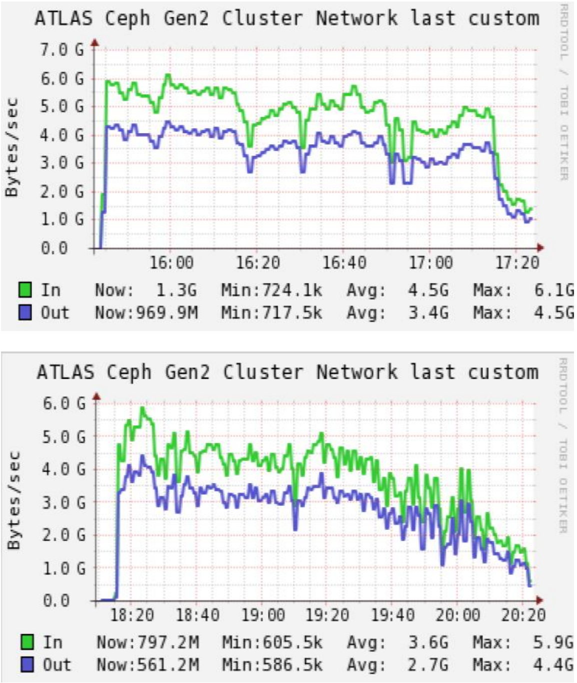

Figure 4. Write throughput test. Values include network activity associated with replication of data. Actual write (marked as "In") is about 1/3 of values shown above: (top) 1 hour long test and (bottom) 2 hour long test. Note that the blue line corresponds to Ceph internal replication traffic.

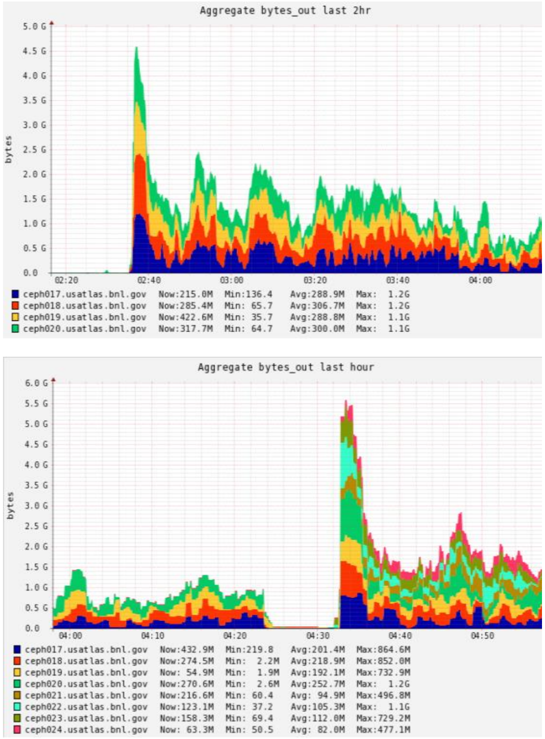

Figure 5. Network traffic observed on the CephFS/XRootD data servers during the CephFS tests. Read throughput test: (top) four CephFS mounting data servers, (bottom) eight CephFS mounting data servers used.

In order to measure read throughput of CephFS, source test files written in the previous section (write tests) were used as source data to be read by clients. Again, XRootD protocol was used here. Figure 5 (top) and (bottom) show network activities seen in the Ceph Storage clusters during the read tests using four or eight XRootD data servers. The colors indicate different hosts used for the data server. The read tests show that in the current setup, four data servers were sufficient to reach the saturation of the stable read-throughput of about 2 GB/s. The tests also revealed the existence of the initial large spike, indicating the effect of cache in the CephFS and OSD read cache on the Ceph cluster head nodes.

7. Conclusion

Two Ceph storage clusters were successfully implemented in RACF at BNL using the existing hardware and reaching the scale of 4k spinning drives and 1 PB of aggregate usable capacity across both clusters while maintaining factor of 3 internal data replication. All three types of Ceph storage layers (object storage/S3, RBD and CephFS) were incorporated as a part of the HEP specific storage services provided by RACF for ATLAS experiment. The experience gained while operating these two clusters indicates that while RBD and CephFS storage layers of Ceph can be used as a part of common storage service, the S3 storage layer provides the opportunity to utilize the data store without the limitations typically associated with file systems. The stability of services provided (including those based on CephFS) is now verified on the timescale of about 10 months, starting from the moment when our first large scale Ceph cluster entered production in 2014.

References

[1] Amazon Simple Storage Service (S3): http://aws.amazon.com/s3/

[2] Gluster File System (GlusterFS): http://www.gluster.org/documentation/About_Gluster/

[3] Ceph storage system: http://ceph.com

[4] Ceph Reliable Autonomic Distributed Object Store (RADOS): https://ceph.com/dev-notes/the-rados-distributed-object-store/

[5] Ceph RADOS Block Device: http://ceph.com/docs/master/rbd/rbd/

[6] Ceph File System (CephFS): http://ceph.com/docs/master/cephfs/

[7] Zaytsev A 2014 Evaluating Infiniband Based Networking Solutions for HEP/NP Data Processing Applications *Report presented at the HEPiX Fall 2014 conference (Lincoln, Nebraska, USA)* https://indico.cern.ch/event/320819/session/4/contribution/46

[8] Zaytsev A 2014 Ceph Based Storage Systems for the RACF *Report presented at the HEPiX Fall 2014 conference (Lincoln, Nebraska, USA)* https://indico.cern.ch/event/320819/session/6/contribution/39

[9] Rind O 2015 Status Report on Ceph Based Storage Systems at the RACF *Report presented at the HEPiX Spring 2015 conference (Oxford, UK)* https://indico.cern.ch/event/346931/session/4/contribution/23

[10] OpenStack Object Storage component: https://wiki.openstack.org/wiki/Swift

[11] Wenaus T 2015 The ATLAS Event Service: A new approach to event processing *To be published in proceedings of the CHEP 2015 conference (J. Phys.: Conf. Series, 2015)*

[12] dCache storage system: http://www.dcache.org

[13] Summary on CephFS differences from POSIX: http://docs.ceph.com/docs/infernalis/dev/differences-from-posix/

[14] XRootD data access service: http://xrootd.org