

# Ceph Project at FNAL

---

Firstname Lastname

October 16, 2014

## 1 INTRODUCTION TO CEPH

Ceph is a open source storage platform designed to present object, block, and file storage from a single distributed computer cluster. Ceph's main goals are to be completely distributed without a single point of failure, scalable to the exabyte level, and freely-available. The data is replicated, making it fault tolerant. Ceph software runs on commodity hardware. The system is designed to be both self-healing and self-managing and strives to reduce both administrator and budget overhead.

Ceph's software libraries provide client applications with direct access to the reliable autonomic distributed object store (RADOS) object-based storage system, and also provide a foundation for some of Ceph's features, including RADOS Block Device (RBD), RADOS Gateway, and the Ceph File System.

### 1.1 CEPH STORAGE CLUSTER

A Ceph Storage Cluster requires at least one Ceph Monitor and at least two Ceph OSD Daemons. The Ceph Metadata Server is essential when running Ceph Filesystem clients.

- Ceph OSDs: A Ceph OSD Deamon stores data, handles data replication, recovery, backfilling, rebalancing, and provides some monitoring information to Ceph Monitors by checking other Ceph OSD Daemons for a heartbeat. A Ceph Storage Cluster requires at least two Ceph OSD Daemons to achieve an active + clean state when the cluster makes two copies of your data.
- Monitors: A Ceph Monitor maintains maps of the cluster state, including the monitor map, the OSD map, the Placement Group (PG) map, and the CRUSH map. Ceph maintains a history (called an "epoch") of each state change in the Ceph Monitors, Ceph OSD Daemons, and PGs.
- MDSs: A Ceph Metadata Server (MDS) stores metadata on behalf of the Ceph Filesystem (i.e., Ceph Block Devices and Ceph Object Storage do not use MDS). Ceph Metadata Servers

make it feasible for POSIX file system users to execute basic commands like ls, find, etc. without placing an enormous burden on the Ceph Storage Cluster.

## 2 CEPH STORAGE CLUSTER AT FNAL

The Ceph Storage Cluster at FNAL consists of 3 Monitors, 8 OSDs and 1 MDS. There are four hosts in the cluster, each of them has 8 cores, 16GB RAM, 15TB HDD, and they are connected by 10GB Ethernet. As shown in figure 1, in each host machine, there reside 1 MON(MDS) and two OSDs. Since MON(MDS) require little cpu or storage resources, they won't affect the performance of OSD that reside in the same host.

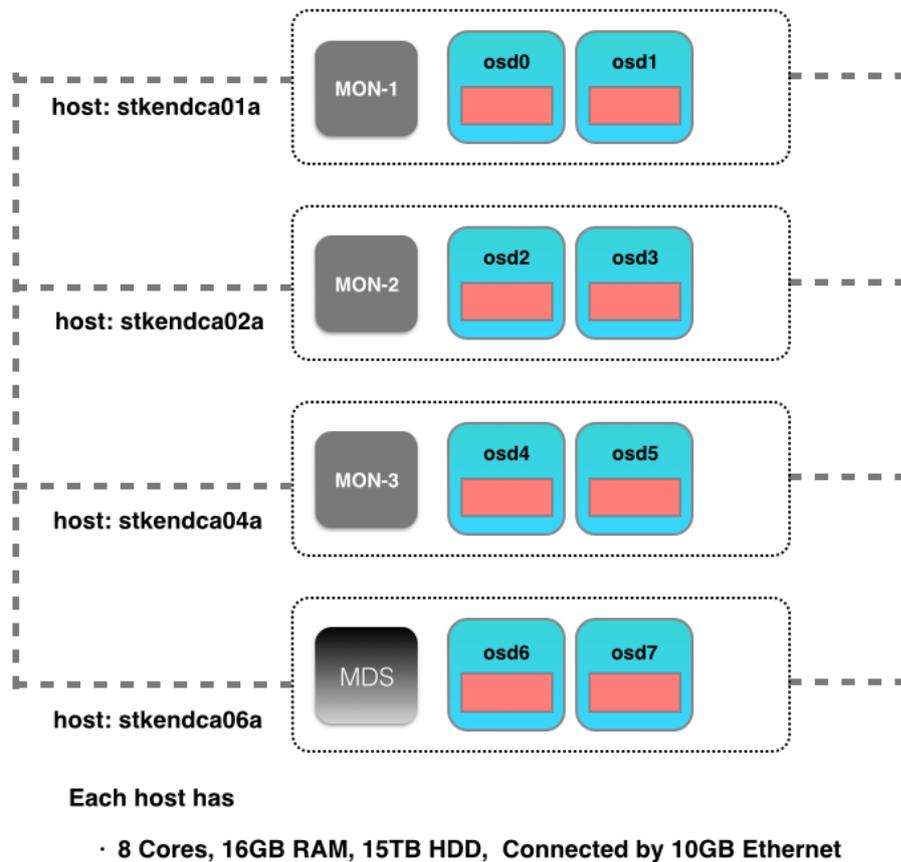


Figure 1: Ceph Storage Cluster at FNAL–Stage 1. Each OSD contributes about 7TB storage capacity to the ceph storage cluster.

## 3 CEPH INSTALLATION AND OPERATION AT FNAL

The Ceph storage cluster at FNAL is implemented and operated by using an automatic tool **ceph-deploy**, which can also automatically add OSD, monitor node, metadata server node and

Table 1: iozone test results of 10 VMs

I/O rate (kB/sec)	Max	Min	Avg	Total
Writer	1208.32	1404.13	1321.477	66073.85
Re-writer	1288.06	1470.5	1363.449	68172.45
Reader	2886.46	4126.75	3672.29	183614.5
Re-reader	3868.67	5290.6	4386.161	219308.05

ceph-client node. **ceph-deploy** is installed on host **stkendca01a**. It only need to be installed on one machine. To use **ceph-deploy**, user need to log in **stkendca01a** as root and work under the directory */home/ceph/ceph-cluster*.

## 4 EVALUATION RESULTS

We have conducted three sets of experiments to evaluate the performance of Ceph storage cluster. In our first experiment, we use 10 VMs as ceph-client to create file system using rbd block device image. Then we conduct iozone tests with these 10 VMs to analysis the I/O performance of Ceph Storage Cluster. In our second experiment, we replace the 10 VMs with 10 physical machines to see the best I/O performance we can get from Ceph Storage Cluster without the virtualization layer. In the third experiment, we try to export 3 copies of rbd block device images on each physical machine simultaneously.

### 4.1 IOZONE TEST WITH VMs

In this experiment, we use 10 VMs that been added in the Ceph storage cluster as clients. Each of them creates a 60GB rbd block device image and map the image to its block device. Then, the VM will use the block device by creating a file system and mount the file system on itself.

The iozone test start with 10 VMs and each VM initiates 5 process perform 10GB I/O operations write/re-write, read/re-read to the Ceph storage cluster simultaneously. The results are presented in table 1.

### 4.2 IOZONE TEST WITH PHYSICAL MACHINES

The prepare work is the same as in 4.1. The iozone test start with 10 physical machines and each initiates 5 process perform 10GB I/O operations write/re-write, read/re-read to the Ceph storage cluster simultaneously. The results are presented in table 2.

Table 2: iозone test results of 10 physical machines

I/O rate (kB/sec)	Max	Min	Avg	Total
Writer	1244.13	1489.29	1387.963	62582.35
Re-writer	1246.1	1636.93	1378.424	60938.1
Reader	1980.19	2407.94	2196.304	99310.25
Re-reader	2234.9	3172.75	2553.519	116336.95

### 4.3 RBD IMAGE EXPORT TEST WITH PHYSICAL MACHINES

The rbd image export tests are conducted by making 10 physical machines export 3 copies of their own rbd image simultaneously. The average bandwidth performance of in this test is 139.78MB/s.