WHITE PAPER

May 2014

Deploying Ceph with High Performance Networks, Architectures and benchmarks for Block Storage Solutions

Contents

Executive Summary

As data continues to grow exponentially storing today's data volumes in an efficient way is a challenge. Many traditional storage solutions neither scale-out nor make it feasible from Capex and Opex perspective, to deploy Peta-Byte or Exa-Byte data stores. A novel approach is required to manage present-day data volumes and provide users with reasonable access time at a manageable cost.

This paper summarizes the installation and performance benchmarks of a Ceph storage solution. Ceph is a massively scalable, open source, software-defined storage solution, which uniquely provides object, block and file system services with a single, unified Ceph Storage Cluster. The testing emphasizes the careful network architecture design necessary to handle users' data throughput and transaction requirements. Benchmarks show that a single user can generate read throughput requirements to saturate a 10Gbps Ethernet network, while the write performance is largely determined by the cluster's media (Hard Drives and Solid State Drives) capabilities. For even a modestly sized Ceph deployment, the usage of a 40Gbps Ethernet network as the cluster network ("backend") is imperative to maintain a performing cluster.

The reference architecture and cost of solution is based on online reseller, single unit offering, on the date this paper is published, and should be adjusted to actual deployment, timing and customer equipment sourcing preference.

Background

Software defined storage solutions are an emerging practice to store and archive large volumes of data. Contemporary web, cloud and enterprise organizations' data grows exponentially, and data growth of Terabytes per day are common practice. Legacy solutions do not suffice to meet these storage needs at a reasonable cost; thus driving customers to find more efficient solutions, such as scale-out software defined storage. One of the leading solutions in the scale-out, software defined storage solutions is Ceph.

Ceph uniquely delivers object, block, and file storage in one unified system. Ceph is highly reliable, easy to manage, and open-source. The power of Ceph can transform your company's IT infrastructure and your ability to manage vast amounts of data. Ceph delivers extraordinary scalability –thousands of clients accessing petabytes or even Exa-bytes of data. A Ceph Node leverages commodity hardware and intelligent daemons, and a Ceph Storage Cluster accommodates large numbers of nodes, which communicate with each other to replicate and redistribute data dynamically. A Ceph Monitor can also be placed into a cluster of Ceph monitors to oversee the Ceph nodes in the Ceph Storage Cluster, thereby ensuring high availability.

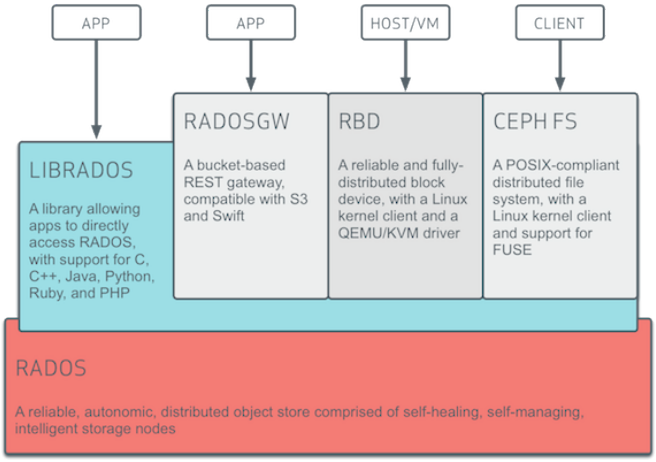

Figure 1. *Ceph Architecture*

Ceph Storage Clusters are dynamic– like a living organism. Whereas, many storage appliances do not fully utilize the CPU and RAM of a typical commodity server, Ceph does. From heartbeats, to peering, to rebalancing the cluster or recovering from faults, Ceph offloads work from clients (and from a centralized gateway which doesn't exist in the Ceph architecture) and uses the distributed computing power of the Ceph Object Storage Daemons (OSDs) to perform the work.

Network Configuration

For a highly scalable fault tolerant storage cluster, the network architecture is as important as the nodes running Ceph Monitors and the Ceph OSD Daemons. The major requirements for Ceph Storage Clusters are high scalability and high availability. So networks obviously must have the capacity to handle the expected number of clients and per-client bandwidth. The networks must also handle Ceph OSD heartbeat, data replication, cluster rebalancing and recovery traffic. In normal operation, a single write to the primary Ceph OSD Daemon results in additional writes to secondary daemons, based on the replication factor set. So, cluster (back-side) traffic significantly exceeds public (front-side) traffic under normal operating conditions.

The public network enables Ceph Client to read data from and write data to Ceph OSD Daemons as well as sending OSDs heartbeats; and, the cluster network enables each Ceph OSD Daemon to check the heartbeat of other Ceph OSD Daemons, send status reports to monitors, replicate objects, rebalance the cluster and backfill and recover when system components fail. When considering the additional loads and tasks on the cluster network, it is reasonable to suggest that the fabric interconnecting the Ceph OSD Daemons should have at least 2X-4X the capacity of the fabric on the public network.

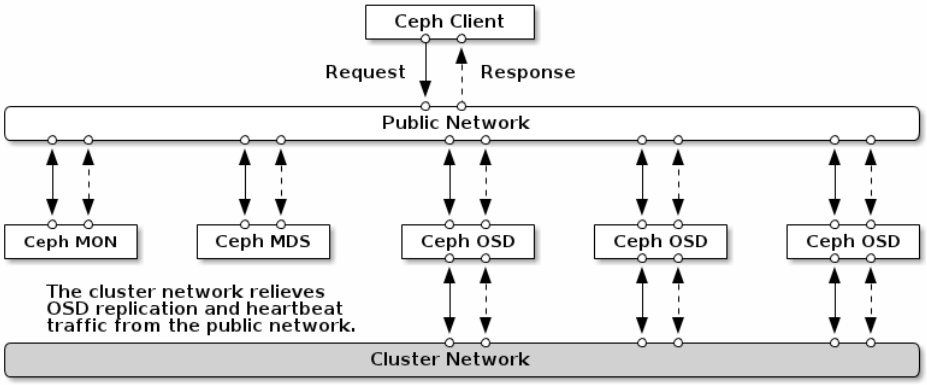

Figure 2. *Ceph Network Architecture*

Test Environment

We created a test environment to measure capabilities of Ceph Block Storage solution over 10Gbps and 40Gbps. Testing goal is to maximize data Ingestion and extraction from a Ceph Block Storage solution.

The environment contains the following equipment:

Ceph Installation:

- Ceph 0.72.2 (Emperor)
- Ceph-deploy 1.3.5
- Each node configured with 5 OSDs (HDDs), 1 Journal (PCIe SSD)
- 3 Monitors

Hardware:

- (5) Ceph OSD nodes
- CPU: 2x Intel E5-2680
- DRAM: 64GB
- Media:
  - 1x 256GB SSD, Boot drive
  - 5x 600GB, 10K RPM Hard drives
  - 1x800GB, PCIe Gen2x4 SSD Card

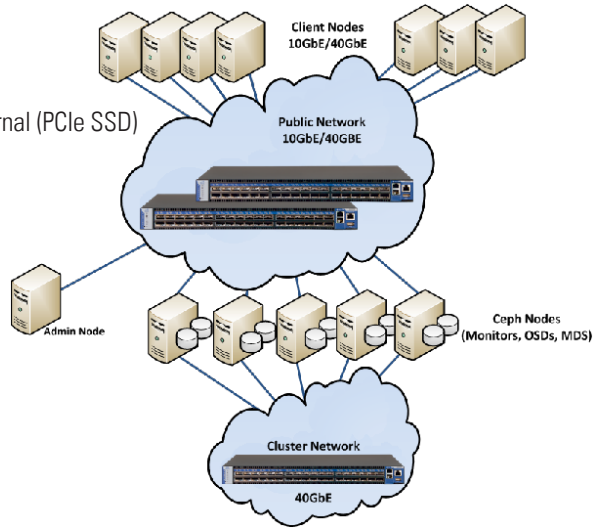

Figure 3. *Ceph Test Environment*

- Network: Mellanox ConnectX®-3 Dual Port 10G/40Gb Ethernet NICs, MCX314A-BCBT

(1) Admin Node:

CPU: 1x AMD Opteron 4180

DRAM: 32GB

Media: 2x 500GB HDD

Network: Mellanox ConnectX®-3 Dual Port 10G/40Gb Ethernet NICs, MCX314A-BCBT

(2) Client Nodes:

CPU: 2x Intel E5-2680

DRAM: 64GB

Media:

1x 256GB SSD, Boot drive

2x1000GB, 7.2K RPM Hard drives

Network: Mellanox ConnectX®-3 Dual Port 10Gb/40Gb Ethernet NICs, MCX314A-BCBT

Switching Solution:

40GbE: Mellanox® MSX1036B, SwitchX®-2 based 40GbE, 1U, 36 QSFP+ ports

10GbE: Mellanox® MSX1016X, SwitchX®-2 based 10GbE, 1U, 64 SFP+ ports

Cabling:

40GbE: MC2210128-003, Mellanox® passive copper cable, ETH 40GbE, 40Gb/s, QSFP, 3m

10GbE: MC3309130-003, Mellanox® passive copper cable, ETH 10GbE, 10Gb/s, SFP+, 3m

The Ceph Cluster Installation

The installation followed the instructions on Ceph.com documentation.

We defined a public and cluster network setting in the ceph.conf file, 3 monitors quorum setting and replication factor set to the default 2.

The testing ceph.conf file can be found in appendix B

The network performance is checked after the installation using iperf tool. The following are the commands used to measure network bandwidth:

Server Side: iperf –s

Client Side: iperf –c <server host IP> -P16 –l64k –i3

For the 10GbE network the bandwidth performance range achieved is 9.48Gb/s to 9.78Gb/s

For the 40GbE network the bandwidth performance range achieved is 36.82Gb/s to 38.43Gb/s

The benchmark testing conducted using fio tool, rev 2.0.13, http://freecode.com/projects/fio. The testing setting is:

fio --directory=/mnt/cephblockstorage --direct=1 --rw=$Action --bs=$BlockSize --size=30G --numjobs=128 --runtime=60 --group_reporting --name=testfile --output=$OutputFile

- $Action=read, write, randread, randwrite
- $bs=4k,128k,8m

The client setting of file system format is ext4. The setting and mounting commands on the client are:

#> rbd -c /etc/ceph/ceph.conf -p benchmark create benchmrk --size 6144000

#> rbd map benchmrk --pool benchmark --secret /etc/ceph/client.admin

#> mkfs.ext4 /dev/rbd1

#> mkdir /mnt/cephblockstorage

#> mount /dev/rbd1 /mnt/cephblockstorage

The testing is conducted with the cluster network set for its maximum bandwidth of 40Gb/s, and varying the public network performance the graph below shows the performance results achieved for sequential read throughput, complete test results for random read and write as well as sequential read and write performance can be found in appendix A

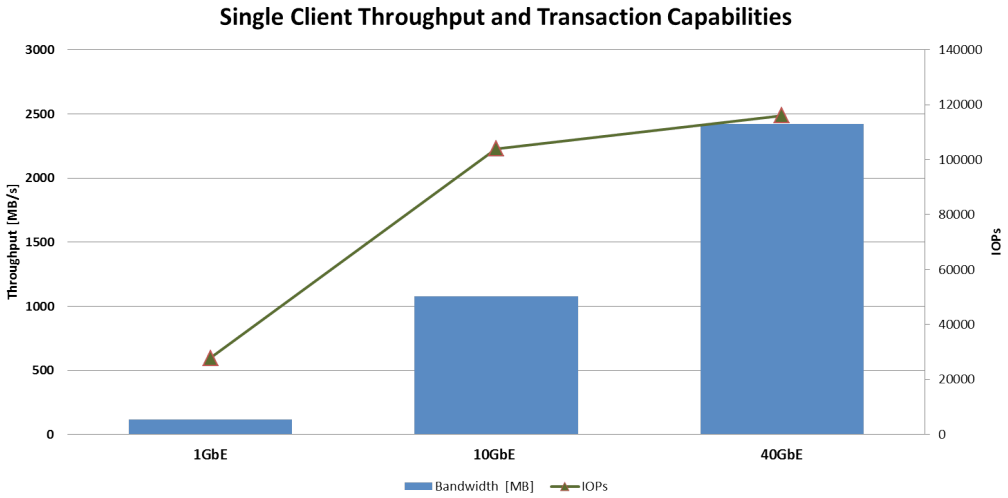

Figure 4. *Ceph Single Client, Sequential Throughput and Transaction Capabilities*

The need for high performance network in the public network is obvious to maintain a performing environment. We suggest a tiered network solution where the Ceph nodes are connected both on the cluster and the public network over a 40GbE. Using Mellanox SX1036B to connect the Ceph nodes, and using Mellanox SX1024 to connect the clients to the Ceph nodes as described in Figure 4.

The need for 40GbE in the cluster network comes from data write replication workload, with servers deployed with as few as 15HDDs, the write bandwidth exceeds 10GbE network capability. In our testing with only 5 drives per machines we experienced over 6Gbps of peak network usage. Obviously for systems incorporating flash based for solid state drivers 10GbE is inadequate and a minimum of 40 Gb/s backend bandwidth is required.

The proposed architecture can scale from 1TB to over 8PB, using 4RU Servers, with 24 media solutions per server. Since the clients are connected with 10GbE solution, concurrent client traffic will saturate the network capability if the Ceph nodes use the same 10GbE solution. To eliminate the bottleneck we suggest connecting the Ceph nodes to a semi-private network operating at 40GbE, connected to the public network in which clients are using 10GbE.

The images below show two examples of this deployment, the first one for 500TB of raw storage and the second one for 8PB of raw storage.

Each rack contains up to 10 servers, 4RU each with 72TB of raw storage using 24x 3TB HDDs and 6 SSD drives used as journals. The servers are connected using a dual port 40GbE card, port 1 connected to the semiprivate network while port 2 is connected to the cluster network. In the 8PB deployment we are using the increased throughput capability of Mellanox switches to drive 56Gbps of Ethernet traffic between the Top of Rack switch and the aggregation layer. The increased throughput reduces the number of cables needed to create non-blocking fat-tree architecture.

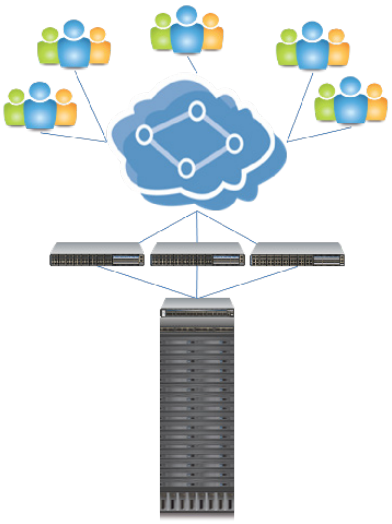

Figure 5. *500TB Ceph Storage Design*

500TB Raw storage configuration

Includes 7 Ceph nodes, each equipped with 24 HDDs, 3TB of raw storage each, 6 SSD drives 480GB each, Dual CPU and total of 20 cores per node, 64GB of DRAM per node, Mellanox ConnectX-3 based NIC Dual port operating at 40GbE. The switching infrastructure contains the following:

Public Network:

- 1x MSX1036B, based on Mellanox SwitchX-2, 36 ports, 40GbE QSFP
- 3x MSX1024B, based on Mellanox SwitchX-2, 48 ports 10GbE, 12 ports 40GbE

Cluster Network:

- 1x MSX1012B, Based on Mellanox SwitchX-2, 12 ports, 40GbE QSFP

This solution can connect over 150 client nodes to the storage nodes, at average client line rate speeds of over 1GB/s. The storage cluster can easily scale to over 800TB without the need to add any additional switching hardware. The cost per tera-byte of this solution is US$410, using a single unit purchase from online retailer at the date of this paper publication. The cost includes all hardware components for the Ceph cluster: Complete servers (CPU, DRAM, HDDs, SSDs, NICs), cluster and public network equipment including cables. The cost does not include the client nodes, any software licensing or professional services fees. Complete hardware list used in this configuration can be found in appendix C.

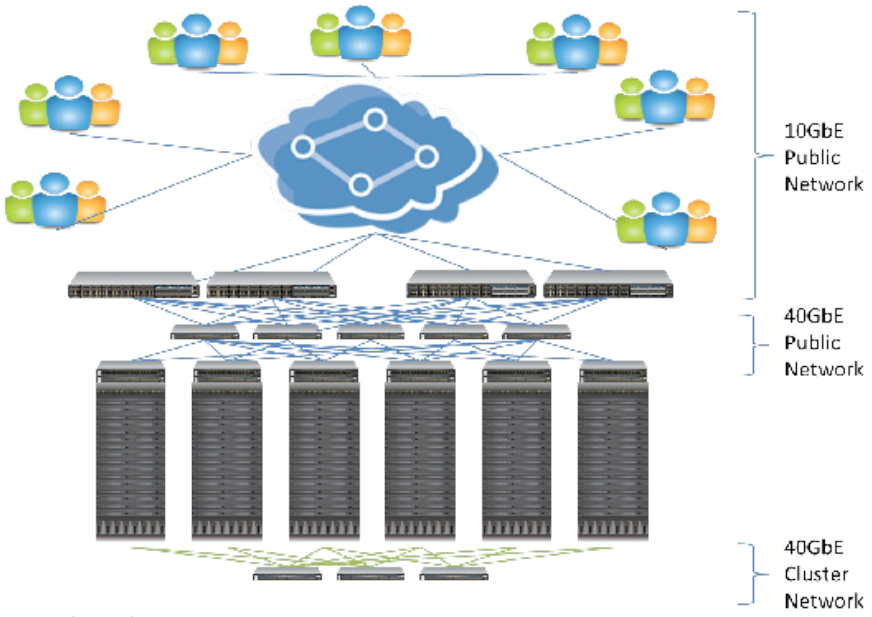

Figure 6. *8.5PB Ceph Storage Design*

8.5PB Raw storage configuration

The scalability of Ceph is indicated by this 8.5PB configuration which includes 120 Ceph nodes, each equipped with 24 HDDs, 3TB of raw storage each, 6 SSD drives 480GB each, Dual CPU and total of 20 cores per node, 64GB of DRAM per node, Mellanox ConnectX-3 based Dual port NIC operating at 40GbE. The switching infrastructure consists of the following:

Public Network:

- 11x MSX1036B, based on Mellanox SwitchX-2, 36 ports, 40GbE QSFP
- 10x MSX1024B, based on Mellanox SwitchX-2, 48 ports 10GbE, 12 ports 40GbE

Cluster Network:

- 9x MSX1012B, Based on Mellanox SwitchX-2, 12 ports, 40GbE QSFP

This solution can connect over 500 client nodes to the storage solution, at client line rate speed of over 1GB/s. The cost per tera-byte of this solution is US$350, using a single unit purchase from online retailer at the date of this paper publication. The cost includes all hardware components for the Ceph cluster: Complete servers (CPU, DRAM, HDDs, SSDs, NICs), cluster and public network equipment including cables. The cost doesn't include the client nodes, any software licensing or professional services fees. Complete hardware list used in this configuration can be found in appendix C.

Conclusion

Ceph software-defined storage deployed with Mellanox's 10G/40Gb Ethernet solutions enables smooth scalability and performance endurance. Creating fast public network tightly coupling Ceph nodes, equipped with multiple OSDs per node, allow clients to effectively use higher throughput and IOPs. The 40GbE cluster network offloads the replication throughput from the public network and provides with faster consistency and reliability of data.

Appendix A

Ceph block storage complete testing results

Network Speed	Block Size [Bytes]	Sequential Read	Sequential Write	Random Read	Random Write
1GbE	4K	108MB/s 27K IOPs	11MB/s 2.7K IOPs	25MB/s 6.4K IOPs	6.3MB/s 1.6K IOPs
	128K	115MB/s 903 IOPs	115MB/s 903 IOPs	115MB/s 903 IOPs	115MB/s 899 IOPs
	8M	115MB/s 14 IOPs	115MB/s 14 IOPs	115MB/s 14 IOPs	115MB/s 14 IOPs
10GbE	4K	408MB/s 102K IOPs	11MB/s 2.7K IOPs	28MB/s 7.2K IOPs	6.3MB/s 1.6K IOPs
	128K	1129MB/s 9K IOPs	262MB/s 2K IOPs	698MB/s 5.4K IOPs	188MB/s 1.4K IOPs
	8M	1130MB/s 141 IOPs	426MB/s 52 IOPs	1131MB/s 141 IOPs	413MB/s 50 IOPs
40GbE	4K	443MB/s 110K IOPs	11MB/s 2.7K IOPs	31MB/s 7.7K IOPs	8.4MB/s 2.1K IOPs
	128K	1894MB/s 15K IOPs	275MB/s 2.1K IOPs	757MB/s 5.9K IOPs	208MB/s 1.6K IOPs
	8M	2419MB/s 302 IOPs	434MB/s 53 IOPs	1773MB/s 221 IOPs	416MB/s 50 IOPs

Appendix B

Ceph Cluster configuration file, our nodes host names are indus001 to indus005

```
[global]
fsid = 1c44c28c-eb54-4013-9c5c-1f4ba5b7b609
mon_initial_members = indus001, indus002, indus003
mon_host = 192.168.140.11,192.168.140.12,192.168.140.13
auth_cluster_required = cephx
auth_service_required = cephx
auth_client_required = cephx
filestore_xattr_use_omap = true
public_network = 192.168.140.0/24
cluster_network = 192.168.190.0/24
# Monitors Section
[mon.indus001]
      host = indus001
      mon_addr = 192.168.140.11:6789
[mon.indus002]
      host = indus002
      mon_addr = 192.168.140.12:6789
[mon.indus003]
      host = indus003
      mon_addr = 192.168.140.13:6789
# OSDs Section
# START INDUS001 OSDs
[osd.0]
      host = indus001
      public_addr = 192.168.140.11
      cluster_addr = 192.168.190.11
[osd.1]
      host = indus001
      public_addr = 192.168.140.11
      cluster_addr = 192.168.190.11
[osd.2]
      host = indus001
      public_addr = 192.168.140.11
      cluster_addr = 192.168.190.11
[osd.3]
      host = indus001
      public_addr = 192.168.140.11
      cluster_addr = 192.168.190.11
[osd.4]
      host = indus001
      public_addr = 192.168.140.11
      cluster_addr = 192.168.190.11
# END OF INDUS001 OSDs
# START INDUS002 OSDs
[osd.5]
host = indus002
      public_addr = 192.168.140.12
      cluster_addr = 192.168.190.12
[osd.6]
      host = indus002
      public_addr = 192.168.140.12
      cluster_addr = 192.168.190.12
```

```
[osd.7]
        host = indus002
        public_addr = 192.168.140.12
        cluster_addr = 192.168.190.12
[osd.8]
        host = indus002
        public_addr = 192.168.140.12
        cluster_addr = 192.168.190.12
[osd.9]
        host = indus002
        public_addr = 192.168.140.12
        cluster_addr = 192.168.190.12
# END OF INDUS002 OSDs
# START INDUS003 OSDs
[osd.10]
        host = indus003
        public_addr = 192.168.140.13
        cluster_addr = 192.168.190.13
[osd.11]
        host = indus003
        public_addr = 192.168.140.13
        cluster_addr = 192.168.190.13
[osd.12]
        host = indus003
        public_addr = 192.168.140.13
        cluster_addr = 192.168.190.13
[osd.13]
        host = indus003
        public_addr = 192.168.140.13
        cluster_addr = 192.168.190.13
[osd.14]
        host = indus003
        public_addr = 192.168.140.13
        cluster_addr = 192.168.190.13
# END OF INDUS003 OSDs
# START INDUS004 OSDs
[osd.15]
        host = indus004
        public_addr = 192.168.140.14
        cluster_addr = 192.168.190.14
[osd.16]
        host = indus004
        public_addr = 192.168.140.14
        cluster_addr = 192.168.190.14
[osd.17]
        host = indus004
        public_addr = 192.168.140.14
        cluster_addr = 192.168.190.14
[osd.18]
        host = indus004
        public_addr = 192.168.140.14
        cluster_addr = 192.168.190.14
```

```
[osd.19]
        host = indus004
        public_addr = 192.168.140.14
        cluster_addr = 192.168.190.14
# END OF INDUS004 OSDs
# START INDUS005 OSDs
[osd.20]
        host = indus005
        public_addr = 192.168.140.15
        cluster_addr = 192.168.190.15
[osd.21]
        host = indus005
        public_addr = 192.168.140.15
        cluster_addr = 192.168.190.15
[osd.22]
        host = indus005
        public_addr = 192.168.140.15
        cluster_addr = 192.168.190.15
[osd.23]
        host = indus005
        public_addr = 192.168.140.15
        cluster_addr = 192.168.190.15
[osd.24]
        host = indus005
        public_addr = 192.168.140.15
        cluster_addr = 192.168.190.15
# END OF INDUS005 OSDs
```

Appendix C

Ceph Server Hardware Configuration

Chassis: 4RU
CPU: Dual Intel E5-2660v2
Total Server Memory: 64GByte
Hard Drives: 24x 3TB SAS, 7.2K RPM
Solid State Drives: 6x 480GB SATA
Boot Drives: 2x 120GB SSD SATA drives
Network card: MCX314A-BCBT, Mellanox ConnectX®-3 EN network interface card, 40GigE, dual-port QSFP, PCIe3.0 x8 8GT/s

Ceph Nodes Top of Rack (ToR) Switching Solution

The ToR solution includes 36 ports, 40GbE each with the optional 56Gbps licensing for ease of deployment. Public and cluster network uses the same building blocks and, where required, the aggregation layer also uses the same switching solution.
Switch: Mellanox SX1036B-1SFS, SwitchX®-2 based 36-port QSFP 40GbE 1U Ethernet Switch, 36 QSFP ports, 1 PS, Standard depth, PSU side to Connector side airflow, Rail Kit and RoHS6 (users should verify the required airflow directions)
Ceph node to ToR cabling solutions

1. MC2207130-001, Mellanox Passive Copper Cable, VPI, up to 56Gb/s, QSFP ,1meter long
2. MC2207130-002, Mellanox Passive Copper Cable, VPI, up to 56Gb/s, QSFP ,2meter long
3. MC2207128-003, Mellanox Passive Copper Cable, VPI, up to 56Gb/s, QSFP ,3meter long

ToR to aggregation/client switching cabling solutions

1. MC2207310-010, Mellanox Active Fiber Cable, VPI, up to 56Gb/s, QSFP, 10meter long

Ceph Client Switching Solution

The Clients are connected with a 10G/40GbE switching solution using the optional 56Gbps licensing for ease of deployment
Switch: Mellanox MSX1024B-1BFS, SwitchX®-2 based 48-port SFP+ 10GbE, 12 port QSFP 40GbE, 1U Ethernet switch. 1PS, Short depth, PSU side to Connector side airflow, Rail kit and ROHS6 (users should verify the required airflow directions)

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com