Reference Architecture: Red Hat Ceph Storage

Last update: 2 May 2019

Version 1.0

Describes the reference architecture for storage using Red Hat Ceph Storage

Describes Lenovo ThinkSystem servers, networking, and storage management software

Provides performance options for the storage solution

Includes validated and tested deployment and sizing guide

Finix Lei
Jay Bryant
Miroslav Halas
Mike Perks
Leo Liu

LENOVO PRESS

Table of Contents

1 Introduction 1

2 Business problem and business value 2

2.1 Business problem 2

2.2 Business value 2

3 Requirements 3

3.1 Functional requirements 3

3.2 Non-Functional requirements 3

4 Architectural overview 4

5 Component model 5

5.1 Core components 5

5.1.1 CRUSH ruleset 5

5.1.2 Pools 6

5.1.3 Placement groups 6

5.1.4 Ceph monitors 6

5.1.5 Ceph OSD daemons 6

5.2 Ceph access methods 6

6 Operational model 8

6.1 Deployment models 8

6.2 Hardware components 8

6.2.1 Servers 9

6.2.2 Network switches 10

6.3 Ceph storage node 12

6.3.1 Disk configuration modes 12

6.3.2 RAID controller configuration 13

6.4 Networking 14

6.5 Integration with Red Hat OpenStack Platform 15

6.6 Systems management 16

6.6.1 Lenovo XClarity Administrator 16

6.6.2 Lenovo Ceph Dashboard 17

6.7 Automated deployment 18

7 Appendix A: Lenovo bill of materials 21

7.1 Server BOM 21

7.2 Networking BOM 24

7.2.1 NE0152T 1GbE Switch 24

7.2.2 NE1032 10GbE Switch 24

7.2.3 NE1032T 10GbE Switch 24

7.2.4 NE1072T 10GbE Switch 24

7.3 Rack BOM 24

7.4 Red Hat Subscription options 24

Resources 25

1 Introduction

Red Hat Ceph Storage is a scalable, open, software-defined storage platform that combines the most stable version of the Ceph storage system with deployment utilities and support services. Red Hat Ceph Storage is designed for cloud infrastructure and web-scale object storage.

A Red Hat Ceph Storage cluster is built from two or more Ceph nodes to provide scalability, fault-tolerance, and performance. Each node uses intelligent daemons that communicate with each other to:

- Store and retrieve data
- Replicate data
- Monitor and report on cluster health
- Redistribute data dynamically (remap and backfill)
- Ensure data integrity (scrubbing)
- Detect and recover from faults and failures

Red Hat Ceph Storage provides effective enterprise block and object storage, supports archival, rich media, and cloud infrastructure workloads such as OpenStack. The advantages are:

- Recognized industry leadership in open source software support services and online support
- Only stable, production-ready code, vs. a mix of interim, experimental code
- Consistent quality; packaging available through Red Hat Satellite
- Well-defined, infrequent, hardened, curated, committed 3-year lifespan with strict policies
- Timely, tested patches with clearly-defined, documented, and supported migration path
- Backed by Red Hat Product Security
- Red Hat Certification and Quality Assurance Programs
- Red Hat Knowledgebase (articles, tech briefs, videos, documentation)
- Automated Services

The purpose of this document is to describe a storage solution based on Red Hat Ceph Storage software and Lenovo ThinkSystem hardware. This reference architecture mainly focuses on Ceph block storage functionality. File storage and object storage are also available after deploying a Red Hat Ceph Storage cluster, but are not discussed further in this document.

The target audience for this Reference Architecture (RA) is system administrators or system architects. Some experience with OpenStack and Ceph technologies may be helpful, but it is not required.

See also the Lenovo Reference Architecture for the Red Hat OpenStack Platform: lenovopress.com/lp0762.

2 Business problem and business value

2.1 Business problem

Traditional storage solutions have difficulties with petabyte level storage. Moreover, many traditional storage solutions do not exhibit high availability and have “single points of failure”.

For dealing with above problems, one storage solution, which can handle the following cases, is required.

- Be able to deal with petabyte level or even exabyte storage capacity
- Easy to scale out
- No single point of failure
- Cluster self-healing and self-managing

2.2 Business value

Red Hat Ceph Storage is a good solution to the business problem described above. Ceph is a unified, distributed, software defined storage system designed for excellent performance, reliability and scalability.

Ceph’s CRUSH algorithm liberates storage clusters from the scalability and performance limitations imposed by centralized data table mapping. It replicates and rebalances data within the cluster dynamically eliminating this tedious task for administrators, while delivering high-performance and excellent scalability.

The power of Ceph can transform organizations’ IT infrastructure and the ability to manage vast amounts of data. Ceph is mainly for organizations to run applications with different storage interfaces. Ceph’s foundation is the Reliable Autonomic Distributed Object Store (RADOS), which provides applications with object, block, and file system storage in a single unified storage cluster—making Ceph flexible, highly reliable and easy for management.

3 Requirements

The functional and non-functional requirements for this reference architecture are described below.

3.1 Functional requirements

Table 1 lists the functional requirements.

Table 1. Functional requirements

Requirement	Description
Support Cluster creation	Multiple storage nodes can be composed as a cluster to provide storage ability.
Support Cluster Expansion	The cluster can be extended by adding more nodes or disks.
Support volume	Users can create volumes from the cluster and use them.
Support snapshot	The snapshots of the volumes can be created/deleted. In addition, new volumes can be created from the snapshots.
Support data backup	Providing backup ability to back up the volumes/snapshots.
Support libvirt	OpenStack can attach the volumes by using libvirt.

3.2 Non-Functional requirements

Table 2 lists the non-functional requirements.

Table 2. Non-Functional requirements

Requirement	Description
No single failure on control plane	When one controller node fails, the cluster can still work well.
Data Redundancy	After store the data into the cluster, there are several copies of data, so that when one or multiple storage nodes down, the data can still be accessed.
Support failure domain	When hardware failure occurs in one failure domain, it will not affect the data accessing in other failure domains. One failure domain can be host, RACK, room, or even some logical region, etc.
Encryption	When data is transferred between clients and the storage cluster, the data should be encrypted.

4 Architectural overview

This chapter gives an architectural overview of Red Hat Ceph Storage (RHCS). Figure 1 shows the Architecture of a RHCS cluster on Lenovo ThinkSystem servers.

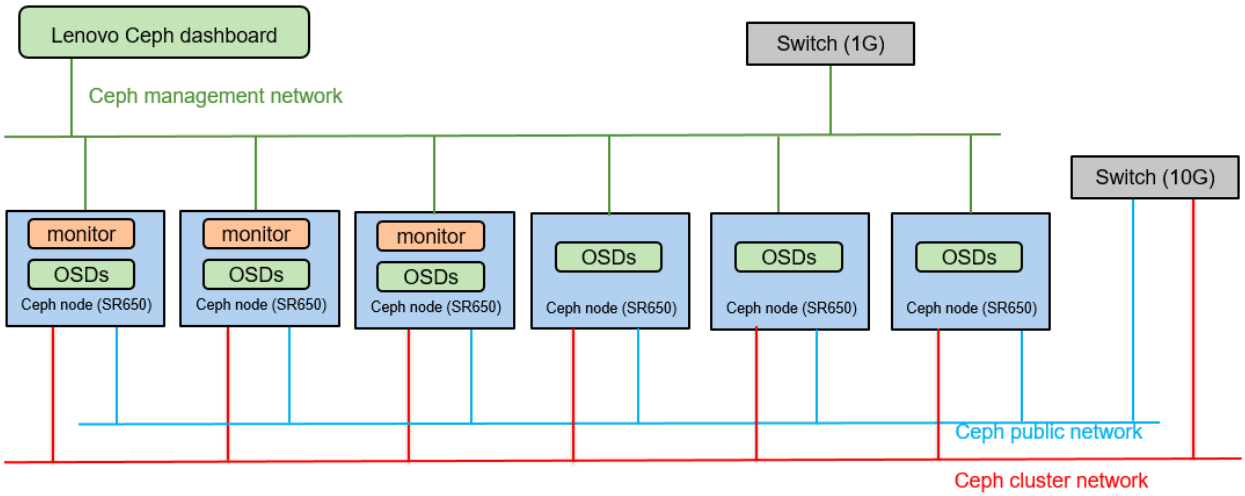

Figure 1: Architecture of RHCS cluster on Lenovo ThinkSystem servers

The RHCS cluster is composed of the following items:

OSD nodes	Object Storage Device (OSD) nodes represent storage nodes that store data. For data replication and failure domains, the minimum number of OSD nodes is 3. In theory, there is no hard limitation on the maximum number of OSD nodes. For more details about OSD daemons, please refer to section 5.1.5.
Monitor nodes	Monitor nodes contain the cluster map that can be used for Ceph client to communicate with OSD services. Monitor services are not computing sensitive or memory sensitive can be combined with OSD services on the same node to reduce cost. The minimum number of monitors is 3. When the number of OSD servers exceeds 15, the number of monitors should be increased to 5. For more details about monitor nodes, i.e. monitors, please refer to section 5.1.4.
Ceph cluster network	This 10Gb network connects all the Ceph nodes (OSD nodes and monitor nodes). This network is used for internal data transfers such as reads, writes, rebalancing, and data recovery.
Ceph public network	This 10Gb network connects all the Ceph nodes and Ceph clients. This network is used for transferring data between Ceph clients and RHCS cluster
Ceph provisioning network	This 1Gb network connects all Ceph nodes. It is for installing OS of each Ceph node.
Lenovo Ceph dashboard	Lenovo Ceph dashboard is an optional component to manage and monitor Red Hat Ceph Storage clusters. For more details, see “Lenovo Ceph Dashboard” on page 17.

5 Component model

Red Hat Ceph Storage (RHCS) is a robust, petabyte-scale storage platform and fit for public or private clouds. It is delivered in a unified self-healing and self-managing platform with no single point of failure, and handles data management, so businesses and service providers can focus on improving application availability.

Figure 2 shows an overview of RHCS components.

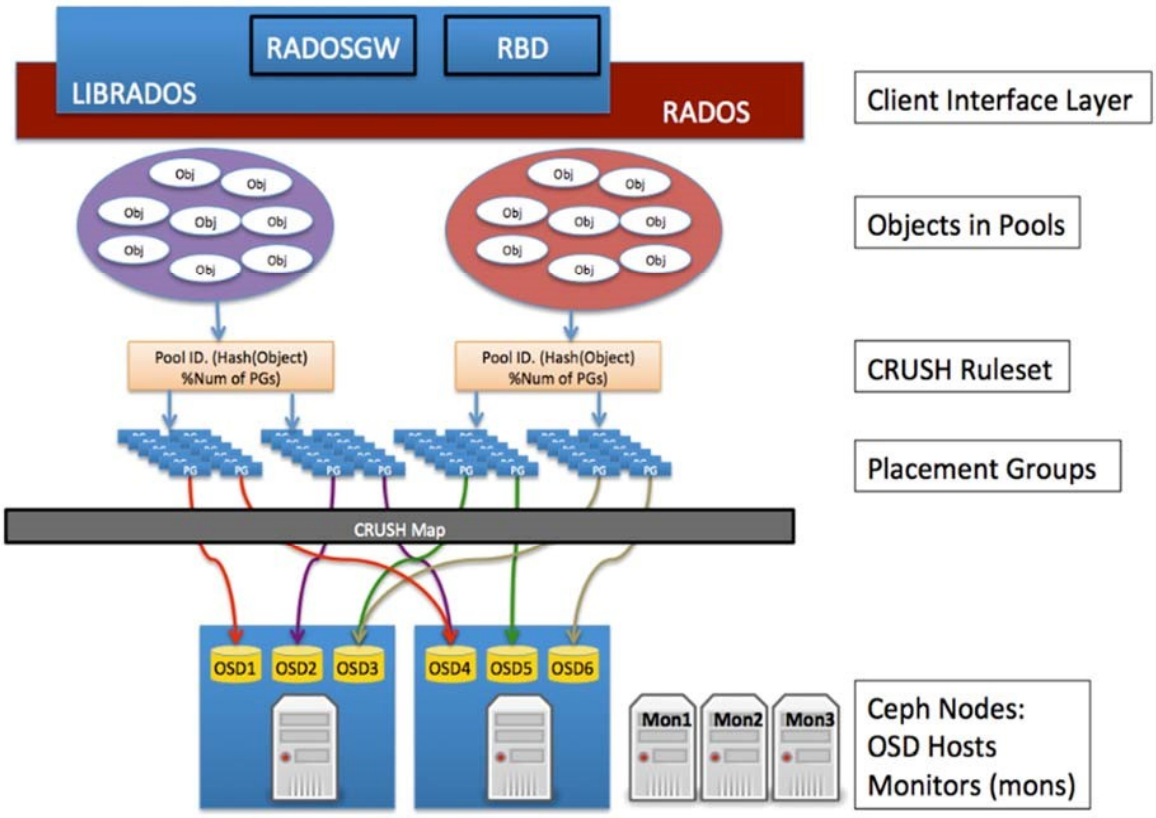

Figure 2: Red Hat Ceph Storage Overview

The following sections will introduce the components of RHCS and access methods.

5.1 Core components

For a Ceph client, the storage cluster is simple. When a Ceph client reads or writes data (referred to as an I/O context), it connects to a logical storage pool in the Ceph cluster. The following sub sections illustrate the core components of Ceph cluster.

5.1.1 CRUSH ruleset

CRUSH is an algorithm that provides controlled, scalable, and decentralized placement of replicated or erasure-coded data within Ceph and determines how to store and retrieve data by computing data storage locations. CRUSH empowers Ceph clients to communicate with Object Storage Devices (OSDs) directly, rather than through a centralized server or broker. By using an algorithm method of storing and retrieving data, Ceph avoids a single point of failure, performance bottlenecks, and a physical limit to scalability.

5.1.2 Pools

A Ceph storage cluster stores data objects in logical dynamic partitions called pools. Pools can be created for particular data types, such as for block devices, object gateways, or simply to separate user groups. The Ceph pool configuration dictates the number of object replicas and the number of placement groups (PGs) in the pool. Ceph storage pools can be either replicated or erasure-coded as appropriate for the desired application and cost model. Erasure-coded pools (EC pool) is outside the scope of this document.

Pools can also “take root” at any position in the CRUSH hierarchy allowing placement on groups of servers with differing performance characteristics—allowing storage to be optimized for different workloads.

5.1.3 Placement groups

Ceph maps objects to placement groups (PGs). PGs are shards or fragments of a logical object pool that are composed of a group of Ceph OSD daemons (see section 5.1.5) that are in a peering relationship. Placement groups provide a way of creating replication or erasure coding groups of coarser granularity than on a per object basis. A larger number of placement groups (for example, 200 per OSD) leads to better balancing.

5.1.4 Ceph monitors

Before Ceph clients can read or write data, they must contact a Ceph monitor (MON) to obtain the current cluster map. A Ceph storage cluster can operate with a single monitor, but this introduces a single point of failure. For added reliability and fault tolerance, Ceph supports an odd number of monitors in a quorum (typically three or five for small to mid-sized clusters). Consensus among various monitor instances ensures consistent knowledge about the state of the cluster.

5.1.5 Ceph OSD daemons

In a Ceph cluster, Ceph OSD daemons store data and handle data replication, recovery, backfilling, and rebalancing. They also provide some cluster state information to Ceph monitors by checking other Ceph OSD daemons with a heartbeat mechanism.

By default, Ceph keeps three replicas of the data. A Ceph storage cluster configured to keep three replicas of every object requires a minimum of three Ceph OSD daemons, two of which need to be operational to successfully process write requests. Ceph OSD daemons roughly correspond to a file system on a hard disk drive. More OSD daemons improves performance but requires increased system memory, e.g. each OSD daemon needs at least 8GB memory.

5.2 Ceph access methods

Choosing a storage access method is an important design consideration. As discussed, all data in Ceph is stored in pools—regardless of data type. The data itself is stored in the form of objects by using the Reliable Autonomic Distributed Object Store (RADOS) layer which:

- Avoids a single point of failure
- Provides data consistency and reliability
- Enables data replication and migration
- Offers automatic fault-detection and recovery

Figure 3 shows the RADOS layer in Ceph.

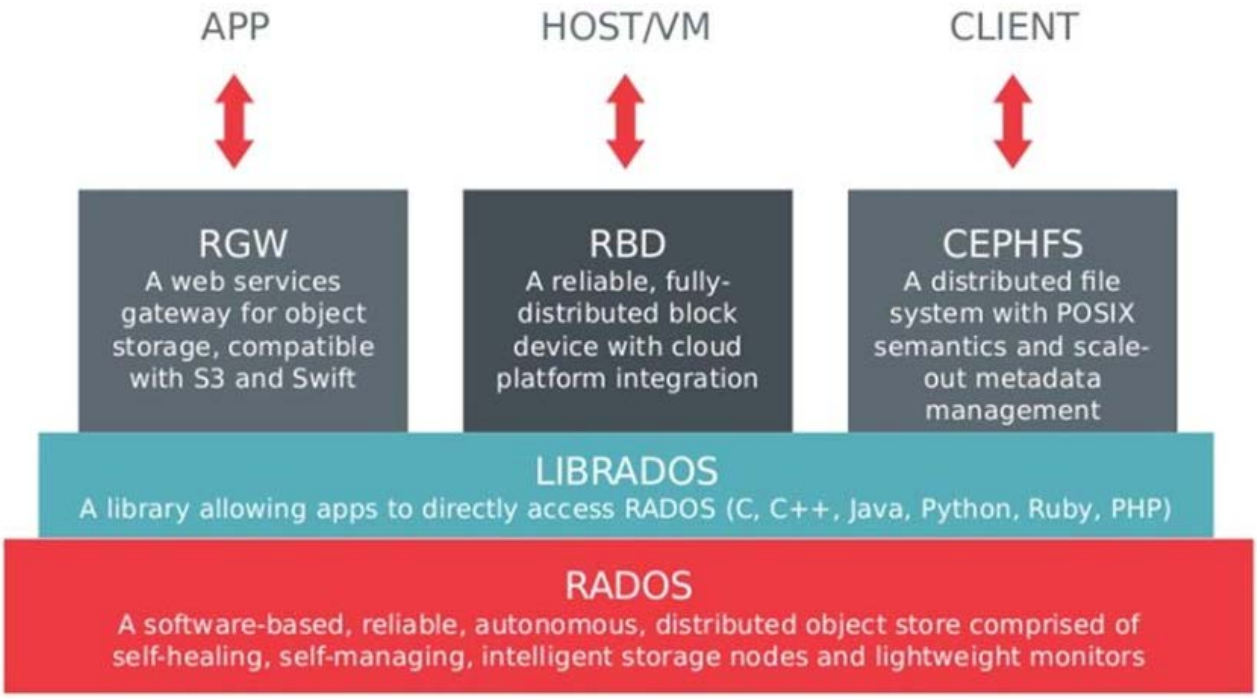

***Figure 3:** RADOS Layer in the Ceph Architecture*

Writing and reading data in a Ceph storage cluster is accomplished by using the Ceph client architecture. A range of access methods are supported, including:

- RADOSGW — Bucket-based object storage gateway service with S3 compatible and OpenStack Swift compatible RESTful interfaces.
- LIBRADOS — Provides direct access to RADOS with libraries for most programming languages, including C, C++, Java, Python, Ruby, and PHP.
- RDB — RADOS Block Device (RDB) offers a Ceph block storage device that mounts like a physical storage drive for use by both physical and virtual systems (with a Linux® kernel driver, KVM/QEMU storage backend, or user space libraries).

6 Operational model

This chapter describes the options for mapping the logical components of RHCS onto Lenovo ThinkSystem servers and network switches. Details of disaster recovery are also described in this chapter. The recommended version of Red Hat Ceph Storage is 3.0 (Luminous) and can downloaded from: access.redhat.com/documentation/en-us/red_hat_ceph_storage/3/.

6.1 Deployment models

Figure 4 shows a minimum RHCS cluster configuration of:

- 3 x Lenovo ThinkSystem SR650 servers
- 2 x Lenovo RackSwitch NE1032T
- 1 x Lenovo RackSwitch NE0152T

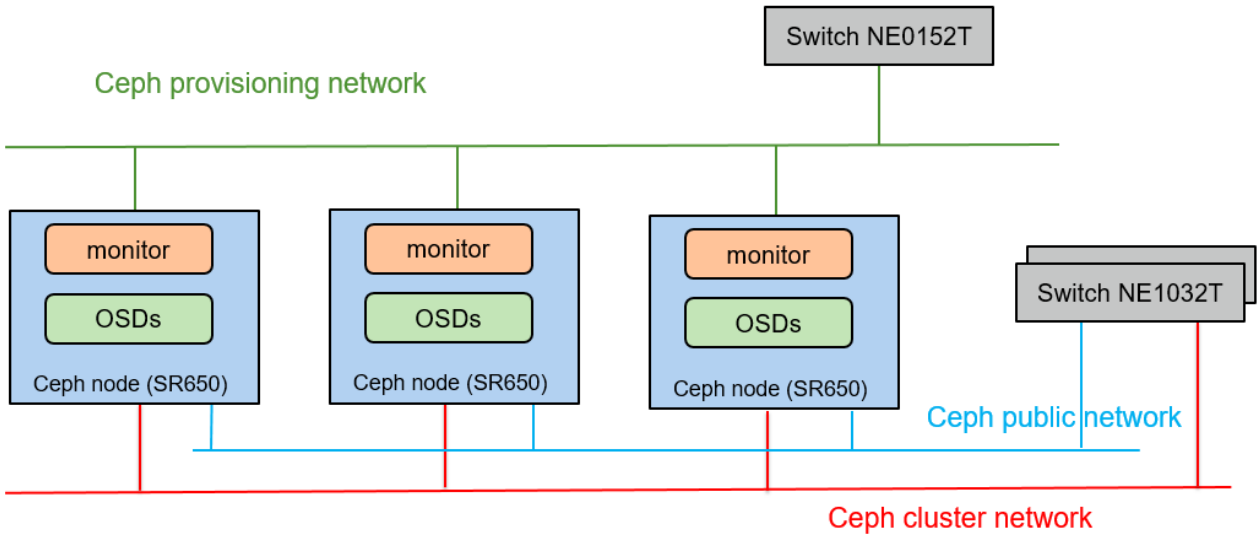

Figure 4: Minimum scale of Ceph cluster

Monitors and OSDs are collocated on each Ceph storage node to minimize the number of servers. The Lenovo ThinkSystem SR650 is more than adequate to support this configuration.

Larger scale deployments are also supported. When the number of Ceph storage nodes is less than or equal to 15, three monitors are sufficient; If the number of Ceph nodes exceeds 15, five monitors should be used.

When users need to expand their Ceph cluster to 15 or more Ceph nodes, Lenovo provides automation support to expand the RHCS cluster. In this process of expansion, the number of monitors is automatically increased from three to five.

A pair of 10GbE switches is needed for networking redundancy.

6.2 Hardware components

A RHCS cluster is composed of multiple Ceph nodes, which connect with each other via dedicated networks. For hardware, it contains the servers (Ceph nodes) and switches.

6.2.1 Servers

Lenovo ThinkSystem SR650 is the recommended server for Ceph nodes because of its overall cost effectiveness and storage capacity.

The Lenovo ThinkSystem SR650 server (as shown in Figure 5 and Figure 6) is an enterprise class 2U two-socket versatile server that incorporates outstanding reliability, availability, and serviceability (RAS), security, and high efficiency for business-critical applications and cloud deployments. Unique Lenovo AnyBay technology provides the flexibility to mix-and-match SAS/SATA HDDs/SSDs and NVMe SSDs in the same drive bays. Four direct-connect NVMe ports on the motherboard provide ultra-fast read/writes with NVMe drives and reduce costs by eliminating PCIe switch adapters. In addition, storage can be tiered for greater application performance, to provide the most cost-effective solution.

Figure 5: Lenovo ThinkSystem SR650 (with 24 x 2.5-inch disk bays)

Figure 6: Lenovo ThinkSystem SR650 (with 12 x 3.5-inch disk bays)

Combined with the Intel® Xeon® Scalable processors product family, the Lenovo ThinkSystem SR650 server offers a high-density combination of workloads and performance. Its flexible, pay-as-you-grow design and great expansion capabilities solidify dependability for any kind of virtualized workload, with minimal downtime. Additionally, it supports two 300W high-performance GPUs and ML2 NIC adapters with shared management.

The Lenovo ThinkSystem SR650 server provides internal storage density of up to 100 TB (with up to 26 x 2.5-inch drives) in a 2U form factor with its impressive array of workload-optimized storage configurations. The ThinkSystem SR650 offers easy management and saves floor space and power consumption for the most demanding storage virtualization use cases by consolidating the storage and server into one system. The Lenovo ThinkSystem SR650 server supports up to twenty-four 2.5-inch or fourteen 3.5-inch hot-swappable SAS/SATA HDDs or SSDs together with up to eight on-board NVMe PCIe ports that allow direct connections to the U.2 NVMe PCIe SSDs. The ThinkSystem SR650 server also supports up to two NVIDIA GRID cards for AI or media processing acceleration.

The SR650 server supports up to two processors, each with up to 28-core or 56 threads with hyper-threading enabled, up to 38.5 MB of last level cache (LLC), up to 2666 MHz memory speeds and up to 3 TB of memory capacity. The SR650 also support up to 6 x PCIe slots. Its on-board Ethernet solution provides 2/4 standard embedded Gigabit Ethernet ports and 2/4 optional embedded 10 Gigabit Ethernet ports without occupying PCIe slots.

For more information, see the following website: ThinkSystem SR650 Product Guide.

6.2.2 Network switches

The following Top-of-Rack (ToR) switches are recommended for the Lenovo Red Hat OpenStack Platform Solution:

- Lenovo ThinkSystem NE0152T RackSwitch
- Lenovo ThinkSystem NE1032 RackSwitch or Lenovo ThinkSystem NE1032T RackSwitch
- Lenovo RackSwitch G8272 or Lenovo ThinkSystem NE1072T RackSwitch

Lenovo ThinkSystem NE0152T Gigabit Ethernet switch

The Lenovo ThinkSystem NE0152T RackSwitch (as shown in Figure 7) is a 1U rack-mount Gigabit Ethernet switch that delivers line-rate performance with feature-rich design that supports virtualization, high availability, and enterprise class Layer 2 and Layer 3 functionality in a cloud management environment.

The NE0152T RackSwitch has 48x RJ-45 Gigabit Ethernet fixed ports and 4x SFP+ ports that support 1 GbE and 10 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables.

The NE0152T RackSwitch runs the Lenovo Cloud Networking Operating System (CNOS) that provides a simple, open and programmable network infrastructure with cloud-scale performance. It supports the Open Network Install Environment (ONIE), which is an open, standards-based boot code that provides a deployment environment for loading certified ONIE networking operating systems onto networking devices.

Figure 7: Lenovo ThinkSystem NE0152T Gigabit Ethernet switch

For more information, see the ThinkSystem NE0152T Product Guide

Lenovo ThinkSystem NE1032/NE1032T/NE1072T RackSwitch family

The Lenovo ThinkSystem NE1032/NE1032T/NE1072T RackSwitch family is a 1U rack-mount 10 Gb Ethernet switch that delivers lossless, low-latency performance with feature-rich design that supports virtualization, Converged Enhanced Ethernet (CEE), high availability, and enterprise class Layer 2 and Layer 3 functionality. The hot-swap redundant power supplies and fans (along with numerous high-availability features) help provide high availability for business sensitive traffic. These switches deliver line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data.

The NE1032 RackSwitch (as shown in Figure 8) has 32x SFP+ ports that support 1 GbE and 10 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables.

Figure 8: *Lenovo ThinkSystem NE1032 RackSwitch*

For more information, see the ThinkSystem NE1032 Product Guide

The NE1032T RackSwitch (as shown in Figure 9) has 24x 1/10 Gb Ethernet (RJ-45) fixed ports and 8x SFP+ ports that support 1 GbE and 10 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables.

Figure 9: *Lenovo ThinkSystem NE1032T RackSwitch*

For more information, see the ThinkSystem NE1032T Product Guide

The NE1072T RackSwitch (as shown in Figure 10) has 48x 1/10 Gb Ethernet (RJ-45) fixed ports and 6x QSFP+ ports that support 40 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. The QSFP+ ports can also be split out into four 10 GbE ports by using QSFP+ to 4x SFP+ DAC or active optical breakout cables.

Figure 10: *Lenovo ThinkSystem NE1072T RackSwitch*

For more information, see the ThinkSystem NE1072T Product Guide

Lenovo RackSwitch G8272

The Lenovo RackSwitch G8272 (as shown in Figure 11) uses 10Gb SFP+ and 40Gb QSFP+ Ethernet technology and is specifically designed for the data center. It is an enterprise class Layer 2 and Layer 3 full featured switch that delivers line-rate, high-bandwidth, low latency switching, filtering, and traffic queuing without delaying data. Large data center-grade buffers help keep traffic moving, while the hot-swap redundant power supplies and fans (along with numerous high-availability features) help provide high availability for business sensitive traffic.

The RackSwitch G8272 (as shown in Figure 11) is ideal for latency sensitive applications, such as high-performance computing clusters, financial applications and NFV deployments. In addition to 10 Gb Ethernet (GbE) and 40 GbE connections, the G8272 can use 1 GbE connections.

Figure 11: Lenovo RackSwitch G8272

For more information, see the RackSwitch G8272 Product Guide

6.3 Ceph storage node

This section describes the recommended configuration for the Ceph storage node. The Lenovo ThinkSystem SR650 should use:

- 2 x Intel Xeon Silver 4116 12C 85W 2.1GHZ processors
- 384 GB of system memory
- 1 or 2 ThinkSystem RAID 930-16i 8GB Flash PCIe 12Gb Adapters (depending on disk configuration)
- Local drives (number and type depend on the disk configuration)
- 2 M.2 480GB drives in a RAID 1 array for the Operating System

For system memory, 384GB should be sufficient for 12-24 OSDs and the Ceph daemons. Because these are storage servers, Xeon “Silver 4 series” processors are sufficient.

Lenovo recommends installing Red Hat Enterprise Linux 7.5 or later on the M.2 drives and configuring two partitions with the XFS filesystem as follows:

- 100GB mounted as /home
- 380GB mounted as “/”

6.3.1 Disk configuration modes

Disk configuration is an important consideration for Ceph storage nodes. A RHCS cluster is configured with only one disk configuration mode. Table 3 summarizes the three recommended disk configurations.

Table 3. Summary of disk configurations

Disk Configuration Mode	OSDs	Raw Storage size	Journal Type	Journal size	OSD Ratio
Mode-1 (Balanced)	22	52.8 TB/node	4 SSD drives	3.75 TB/node	5.5
Mode-2 (Performance)	24	64.8 TB/node	2 NVMe drives	2 TB/node	12
Mode-3 (Capacity)	12	168 TB/node	2 NVMe drives	2 TB/node	6

Disk configuration mode-1 is a balance of cost and performance, as it contains a high number of Object Storage Devices (OSDs) and uses SSDs for the journal. Disk configuration mode-2 is for performance, as it contains the most OSDs and uses NVMe drives for journaling. Disk configuration mode-3 is mainly for storage capacity purposes, as it contains the largest storage size, but retains performance by using NVMe drives for journaling.

Storage size in the table above refers to the raw storage size. For example using a replication factor of 3, the effective capacity of Mode-1 is 52.8/3=17.6TB per node.

The OSD ratio is the ratio between the number of OSDs and the number of journal drives. The recommended ratio for SSDs is between 4 and 5 to 1 and for NVMe drives it is between 12 and 18 to 1. For Mode-3, the NVMe journal ratio is 6 which is much higher than 12 to provide better performance.

Disk Configuration Mode–1

In this mode, all drives are 2.5” and up to 26 can be used in a SR650. Two 930-16i RAID controllers are needed.

Table 4. Disk Configuration Mode-1

Count	Position	Count	Capacity	Purpose
HDD	Front backplane (2.5”)	22	2.4 TB	OSD
SSD	Front backplane (2.5”)	2	960 GB	Journal
SSD	Rear backplane (2.5”)	2	960 GB	Journal

Disk Configuration Mode–2

In this mode, the front 24 drives are 2.5” and two drives in the rear are 3.5”. Two 930-16i RAID controllers are needed.

Table 5. Disk Configuration Mode-2

Count	Position	Count	Size	Purpose
HDD	Front backplanes (2.5”)	22	2.4 TB	OSD
HDD	Rear backplane (3.5”)	2	6 TB	OSD
NVMe	Front backplane (AnyBay)	2	1 TB	Journal

Disk Configuration Mode–3

In this mode, all drives are 3.5” and up to 14 can be used in a SR650. One 930-16i RAID controllers is needed.

Table 6. Disk Configuration Mode-3

Type	Position	Number	Size	Purpose
HDD	Front backplane (3.5”)	10	14 TB	OSD
HDD	Rear backplane (3.5”)	2	14 TB	OSD
NVMe	Front backplane (AnyBay)	2	1 TB	Journal

6.3.2 RAID controller configuration

Lenovo recommends using the 930-16i 8GB flash RAID adapter because the large cache improves performance. The Lenovo XClarity Controller (XCC) can be used to configure the RAID controller for optimal performance. Lenovo recommends the following settings:

- Read policy set to “Read Ahead”
- Write policy set to “Write Back” (uses the cache with battery to store the cache if there is a power failure)
- Each drive is configured as a separate RAID 0 drive

6.4 Networking

Figure 12 shows an overview of the recommended networking for RHCS.

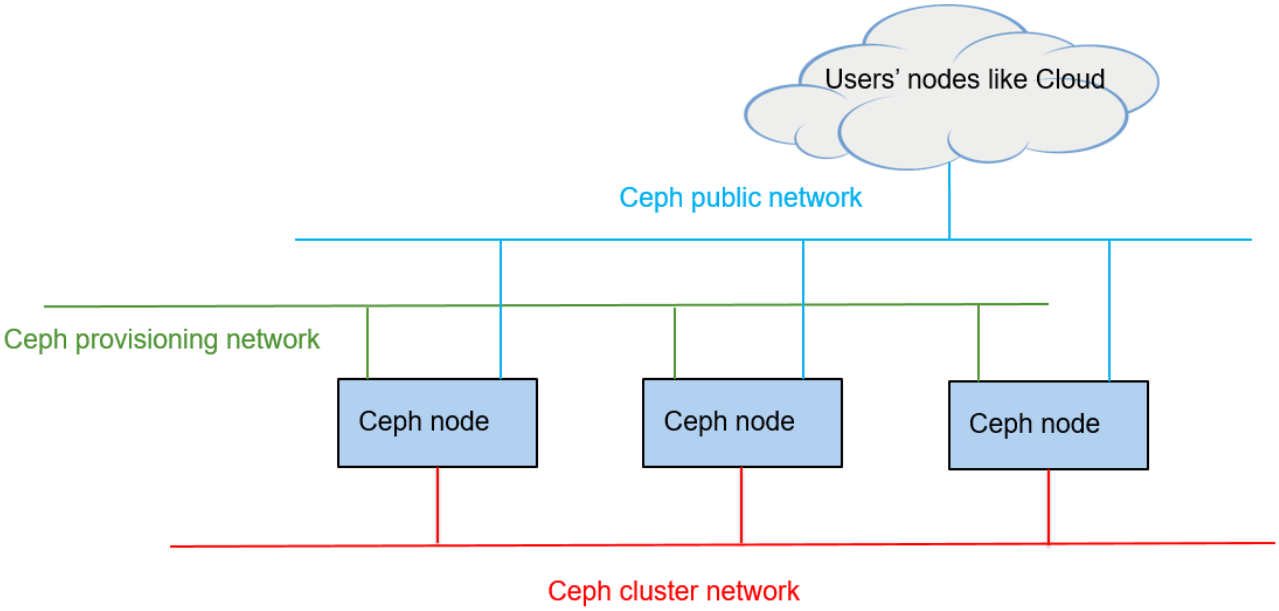

Figure 12: Ceph cluster networking

The Ceph storage nodes should be connected to at least three networks.

Ceph cluster network	This 10Gb network connects all the Ceph nodes (OSD nodes and monitor nodes). This network is used for internal data transfers such as reads, writes, rebalancing, and data recovery.
Ceph public network	This 10Gb network connects all the Ceph nodes and Ceph clients. This network is used for transferring data between Ceph clients and RHCS cluster
Ceph provisioning network	This is a 1Gb network for installing OS of Ceph nodes.

To improve the stability and performance of network data transmission, network bonding should be configured on the Ceph nodes as follows:

- For Ceph cluster network, two physical network ports of 10 Gb should be configured in bonding mode 4, which is 802.3ad mode. This mode follows IEEE 802.3ad Dynamic link aggregation policy. It creates aggregation groups that share the same speed and duplex settings, and utilizes all slaves in the active aggregator according to the 802.3ad specification.

- For Ceph public network, two physical network ports of 10 Gb should be configured in bonding mode 4, which is 802.3ad mode.
- For Ceph provisioning network, two physical network ports of 1 Gb should be configured in bonding mode 1, which is active-backup mode. This mode establishes that only one slave in the bond is active. A different slave becomes active if, and only if, the active slave fails.

A pair of 10 Gb switches are used for networking redundancy. The Virtual Link Aggregation Group (vLAG) feature on these switches should be enabled to allow a pair of switches to act as a single endpoint for aggregation. It provides improved high availability compared to a single switch, enhances performance and increases the available bandwidth by splitting loads across the aggregated links and switches.

Table 7 shows the recommended VLANs corresponding to the networks described above.

Table 7. Logical networking

Network	Speed	VLAN	Members	Functionality
Ceph cluster	10GbE	401	Ceph nodes	Ceph private data transferring, e.g. rebalance, recovery
Ceph public	10GbE	301	Ceph nodes, Ceph clients	Data transferring between Ceph clients and Ceph nodes
Ceph provisioning	1GbE	10	Ceph nodes	OS installation for Ceph nodes.

Table 8 shows the recommended network bonding configuration for the storage nodes.

Table 8. Bond configuration on physical level

Name	Physical Ports	Speed	Mode	VLAN	Network
bond0	NIC1-port1, NIC2-port1	1GbE	1	10	Ceph provisioning
bond1	NIC3-port1, NIC4-port1	10GbE	4	301	Ceph public
bond2	NIC3-port2, NIC4-port2	10GbE	4	401	Ceph cluster

6.5 Integration with Red Hat OpenStack Platform

One popular use case for Ceph block storage is to serve as the backend of a cloud implementation such as Red Hat OpenStack Platform, which needs at least five storage pools:

1. pool for Nova: storing virtual machine images (pool_nova)
2. pool for Glance: storing template images of virtual machine OSes (pool_image)
3. pool for Cinder: storing virtual disks/volumes of virtual machines (pool_volume)
4. pool for Cinder backup: storing backup volumes (pool_backup)
5. pool for Metrics: storing metrics for Ceilometer (pool_metrics)

Here is an example of how to plan the pools for a 3-node Ceph cluster using the disk configuration Mode-1 (see section in 6.3.1). Since the maximum number placement groups (PG) is static for a Ceph cluster, the number of pools needs to be planned. Three nodes is a maximum of 66 OSDs. Generally, one OSD can map to 100-200 PGs. Assuming a replica count of 3, the maximum number of placement groups is 4400.

A deployment may have hundreds of VMs, each of which needs storage for its image and any volumes it uses. For cloud deployments, often multiple ocpies of the same VM are used and this requires much less storage, as only “golden images” are needed.

Therefore, it can be assumed that pool_nova and pool_volumes should be the two biggest pools. For example, pool_volume takes 40% of storage, pool_nova takes 30% of storage, pool_backup takes 15% of storage, pool_images takes 8% of storage and pool_metrics takes 7% of storage.

The pgcalc tool from the official Ceph website can be used to calculate the PG count for each pools. Figure 13 shows the results of this calculation using the input described above. The total number of PGs is less than the maximum of 4400. The recommended number of PGs for pool_nova is 1024, for example.

Pool Name	Size	OSD #	%Data	Target PGs per OSD	Suggested PG Count
pool_volume	3	66	40.00	200	2048
pool_nova	3	66	30.00	200	1024
pool_backup	3	66	15.00	200	512
pool_images	3	66	8.00	200	512
pool_metrics	3	66	7.00	200	256
			Total Data Percentage: 100.00%		**PG Total Count: 4352**

Figure 13: Example PGs for pools

For more information, see the Lenovo Reference Architecture for the Red Hat OpenStack Platform: lenovopress.com/lp0762

6.6 Systems management

Systems management is an important aspect of any deployment. Lenovo provides two tools to help with this:

- Lenovo XClarity Administrator
- Lenovo Ceph Dashboard

6.6.1 Lenovo XClarity Administrator

Lenovo XClarity™ Administrator is a centralized resource management solution that reduces complexity, speeds up response, and enhances the availability of Lenovo® server systems and solutions.

The Lenovo XClarity Administrator provides agent-free hardware management for Lenovo’s ThinkSystem® rack servers, ThinkAgile VX servers, System x® rack servers, and Flex System™ compute nodes and

components, including the Chassis Management Module (CMM) and Flex System I/O modules. Figure 14 shows the Lenovo XClarity administrator interface, in which Flex System components and rack servers are managed and are seen on the dashboard. Lenovo XClarity Administrator is a virtual appliance that is quickly imported into a virtualized environment server configuration.

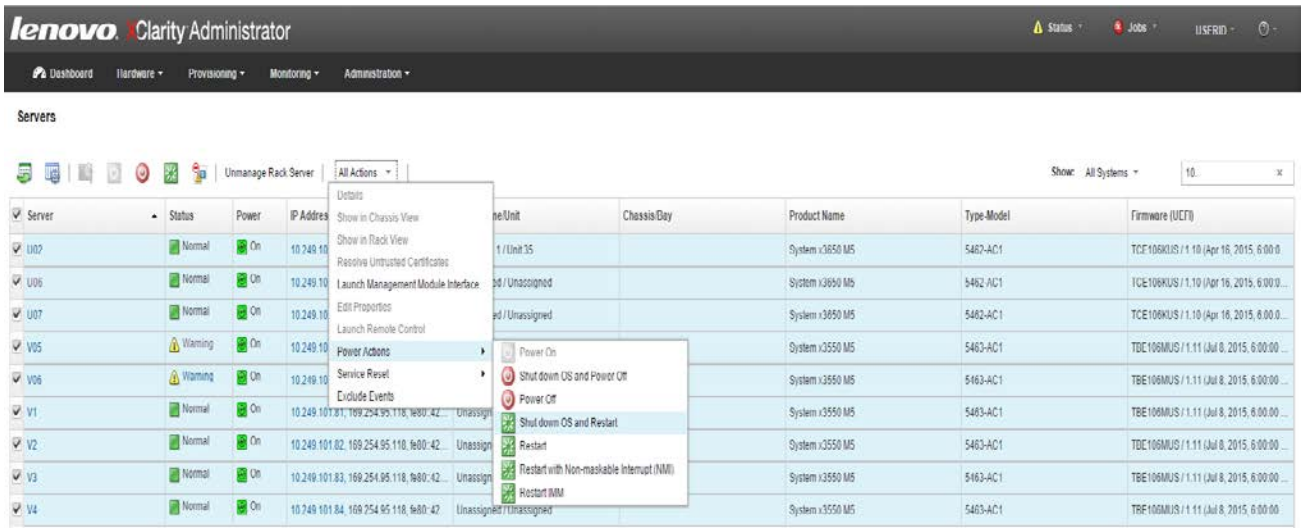

Figure 14: XClarity Administrator interface

6.6.2 Lenovo Ceph Dashboard

The Lenovo Ceph dashboard, developed by Lenovo, is the recommended management software for Ceph Storage. Figure 15 shows an example of the Lenovo Ceph dashboard.

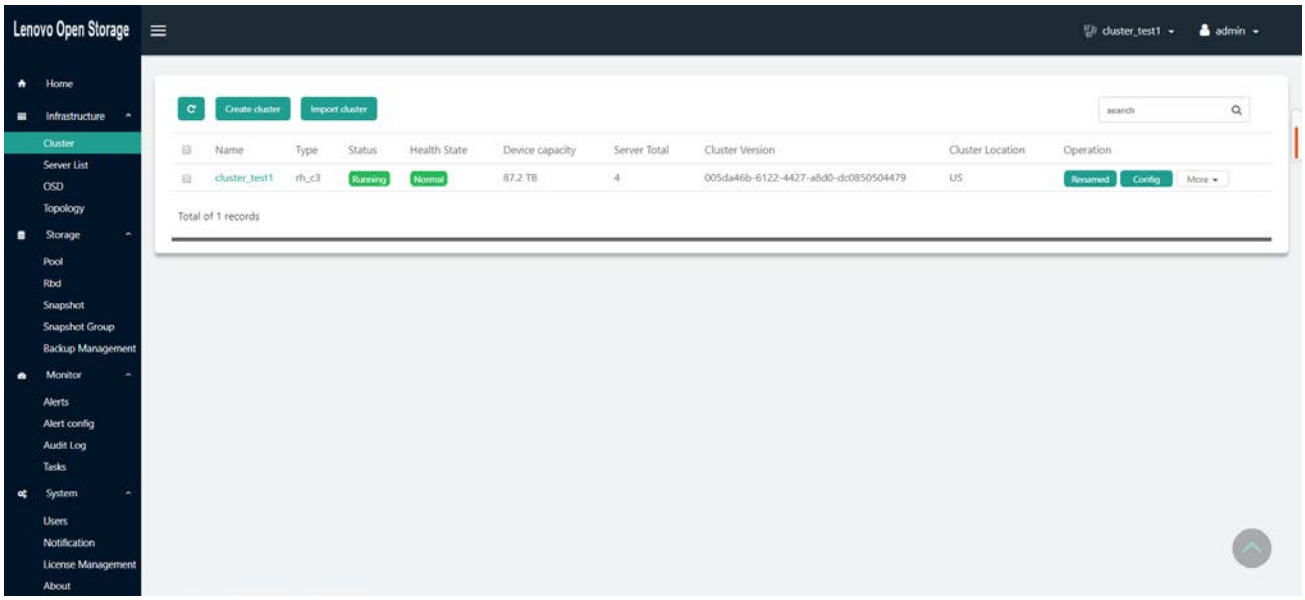

Figure 15 Lenovo Ceph dashboard

The Lenovo Ceph dashboard supports many monitoring and management functions for a Ceph cluster including:

- Cluster management including showing the topology and performance graphs

- Management of Ceph nodes, disks, network interfaces
- Management of OSDs
- Management of pools, placement groups and volumes
- Management of RADOS Block Devices (RDBs)
- Management of snapshots
- Back up of incremental data or full data to a remote Ceph cluster, a S3 cluster or an AWS S3 cluster.
- Alerts and alert history

Lenovo Ceph dashboard can be installed either on a physical machine or in a virtual machine located in the Ceph management network. The minimum hardware requirements for the Lenovo Ceph dashboard are:

- 2 CPU cores
- 8 GB memory
- 60 GB disk mounted on /
- 1 Gb network interface card

The operating system for Lenovo Ceph dashboard is the same as the operating system of Ceph nodes. It is Red Hat Enterprise Linux 7.5 or later.

Lenovo Ceph dashboard is available from Lenovo Professional Services. Please contact a Lenovo representative for more details about the Lenovo Ceph Dashboard.

6.7 Automated deployment

Deploying a Red Hat Ceph Storage (RHCS) cluster can be time-consuming and error prone. Lenovo provides a fully automated solution that covers the end-to-end process of deploying RHCS cluster on bare metal machines or for expanding an existing cluster with additional nodes or drives. This automation solution allows clicking one button or running one command to complete the deployment from scratch. Ansible Tower, which provides a GUI for customers, is supported.

Here are the deployment steps:

1. Set up one virtual machine or real machine as Lenovo XClarity server. Lenovo XClarity is a Lenovo organic software for managing ThinkSystem servers and can be used to install the OS on bare metal machines.
2. Install Ansible on another virtual machine, which is used to access the Ceph nodes via the provisioning network.
3. Ansible playbooks are available from Lenovo to perform the following steps:
    a) Install OS for the specified bare metal machines;
    b) Set up networks of these servers;
    c) Register these servers to Red Hat Satellite;
    d) Deploy Ceph on the these servers;
    e) Optionally set up pools as storage backend for Cloud, e.g. Red Hat OpenStack Platform.

Ansible playbooks are usually run with one inventory file and one extra variable file. The inventory file

specifies on which hosts the playbook will be run. The extra variable file specifies the configuration variables used by the playbook. For items a) to f) above, each item has its own playbook and inventory, but for facilitating DevOps' work of adjusting the configuration items per different environments, extra variable files are integrated for most of playbooks.

Table 9 shows an example of the configuration file for Ceph cluster deployment.

Table 9. Configuration file for Ceph cluster deployment

```
---
osd_scenario: collocated
devices:
  - /dev/sdb
  - /dev/sdc
  - /dev/sdd
  - /dev/sde
  - /dev/sdf
  - /dev/sdg
  - /dev/sdh
  - /dev/sdi
  - /dev/sdj
  - /dev/sdk
  - /dev/sdl
  - /dev/sdm
# This is for journal
dedicated_devices:
  - /dev/nvme0n1
  - /dev/nvme0n1
  - /dev/nvme0n1
  - /dev/nvme0n1
  - /dev/nvme0n1
  - /dev/nvme0n1
  - /dev/nvme1n1
  - /dev/nvme1n1
  - /dev/nvme1n1
```

```
  - /dev/nvme1n1
  - /dev/nvme1n1
  - /dev/nvme1n1
# normal config
ceph_origin: repository
ceph_repository: rhcs
ceph_repository_type: cdn
ceph_rhcs_version: 3
monitor_interface: eno1
public_network:  <IP>/<prefix>
cluster_network: <IP>/<prefix>

# for expanding cluster
fetch_directory: /var/tmp/ceph-ansible-keys
```

Table 10: Inventory file for deploying Ceph cluster.

Table 10: Inventory file for deploying Ceph cluster

```
[mons]
ceph-node1
ceph-node2
ceph-node3
[osds]
ceph-node1
ceph-node2
ceph-node3
```

The automation solution also supports expanding the Ceph cluster in two ways:

- Expand Ceph cluster to more nodes — Append more OSD nodes in “[osds]” section of Table 10: Inventory file for deploying Ceph cluster, and then re-run the Ceph deployment playbook to extend the Ceph cluster.
- Expand Ceph cluster by adding more drives on each node — Append more disks in sections “devices” and “dedicated_devices” of Table 9. Configuration file for Ceph cluster deployment. Then re-run the Ceph deployment playbook to extend the Ceph cluster.

7 Appendix A: Lenovo bill of materials

This appendix contains the Bill of Materials (BOMs) for the different hardware configurations supported for the Red Hat Ceph Storage deployments.

7.1 Server BOM

The following section contains the BOM for Lenovo ThinkSystem Servers.

There are three tables below. The first two are for 2.5'' disks configuration, and the second is for 3.5'' disks configuration. For setting up a Ceph cluster, choose either of the two tables.

The tables below are generated using the online Lenovo Data Center Solution Configurator. Each table represents one single Ceph node. To order N Ceph nodes to construct the Ceph cluster, just multiply by N with the value in the "Quantity" column.

Table 11. Single Ceph node in disk configuration mode-1

Code	Description	Quantity
7X06CTO1WW	Storage-development-mode-1-2.5" : ThinkSystem SR650 - 3yr Warranty	1
AUVV	ThinkSystem SR650 2.5" Chassis with 8, 16 or 24 bays	1
AWER	Intel Xeon Silver 4116 12C 85W 2.1GHZ	2
AUND	ThinkSystem 32GB TruDDR4 2666 MHz (2Rx4 1.2V) RDIMM	12
AUR5	ThinkSystem 2U/Twr 2.5" AnyBay 8-Bay Backplane	1
AURA	ThinkSystem 2U/Twr 2.5" SATA/SAS 8-Bay Backplane	2
AURZ	ThinkSystem SR590/SR650 Rear HDD Kit	1
5977	Select Storage devices - no configured RAID required	1
B31E	ThinkSystem RAID 930-16i 8GB Flash PCIe 12Gb Adapter	2
B0YS	ThinkSystem 2.5" 2.4TB 10K SAS 12Gb Hot Swap 512e HDD	22
B0ZV	ThinkSystem 3.5" Intel S4600 960GB Mainstream SATA 6Gb Hot Swap SSD	4
AUMV	ThinkSystem M.2 with Mirroring Enablement Kit	1
B11V	ThinkSystem M.2 5100 480GB SATA 6Gbps Non-Hot Swap SSD	2
AURC	ThinkSystem SR550/SR590/SR650 (x16/x8)/(x16/x16) PCIe FH Riser 2 Kit	1
AUKH	ThinkSystem 1Gb 4-port RJ45 LOM	1
AUKX	ThinkSystem Intel X710-DA2 PCIe 10Gb 2-Port SFP+ Ethernet Adapter	2
5053	SFP+ SR Transceiver	4
AVWF	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
AUPW	ThinkSystem XClarity Controller Standard to Enterprise Upgrade	1
B0MJ	Feature Enable TPM 1.2	1
AUQB	Lenovo ThinkSystem Mainstream MB - 2U	1
B31F	ThinkSystem M.2 480GB SSD Thermal Kit	1

Table 12. Single Ceph node in disk configuration mode-2

Code	Description	Quantity
7X06CTO1WW	storage development-mode-2--2.5" : ThinkSystem SR650 - 3yr Warranty	1
AUVV	ThinkSystem SR650 2.5" Chassis with 8, 16 or 24 bays	1
AWER	Intel Xeon Silver 4116 12C 85W 2.1GHZ	2
AUND	ThinkSystem 32GB TruDDR4 2666 MHz (2Rx4 1.2V) RDIMM	12
AURA	ThinkSystem 2U/Twr 2.5" SATA/SAS 8-Bay Backplane	2
AUR5	ThinkSystem 2U/Twr 2.5" AnyBay 8-Bay Backplane	1
AURZ	ThinkSystem SR590/SR650 Rear HDD Kit	1
5977	Select Storage devices - no configured RAID required	1
AUV1	ThinkSystem RAID 930-24i 4GB Flash PCIe 12Gb Adapter	1
AUNG	ThinkSystem RAID 530-8i PCIe 12Gb Adapter	1
B0YS	ThinkSystem 2.5" 2.4TB 10K SAS 12Gb Hot Swap 512e HDD	22
B11C	ThinkSystem U.2 Intel P4500 1.0TB Entry NVMe PCIe3.0 x4 Hot Swap SSD	2
AUU7	ThinkSystem 3.5" 6TB 7.2K SAS 12Gb Hot Swap 512e HDD	2
AUMV	ThinkSystem M.2 with Mirroring Enablement Kit	1
B11V	ThinkSystem M.2 5100 480GB SATA 6Gbps Non-Hot Swap SSD	2
AURC	ThinkSystem SR550/SR590/SR650 (x16/x8)/(x16/x16) PCIe FH Riser 2 Kit	1
AUKH	ThinkSystem 1Gb 4-port RJ45 LOM	1
AUKX	ThinkSystem Intel X710-DA2 PCIe 10Gb 2-Port SFP+ Ethernet Adapter	2
AVWF	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
AUPW	ThinkSystem XClarity Controller Standard to Enterprise Upgrade	1
B0MJ	Feature Enable TPM 1.2	1
AUQB	Lenovo ThinkSystem Mainstream MB - 2U	1
B31F	ThinkSystem M.2 480GB SSD Thermal Kit	1

Table 13. Single Ceph node in disk configuration mode-3

Code	Description	Quantity
7X06CTO1WW	storage development-mode-3--3.5" : ThinkSystem SR650 - 3yr Warranty	1
AUVW	ThinkSystem SR650 3.5" Chassis with 8 or 12 bays	1
AWER	Intel Xeon Silver 4116 12C 85W 2.1GHZ	2
AUND	ThinkSystem 32GB TruDDR4 2666 MHz (2Rx4 1.2V) RDIMM	16
AUR8	ThinkSystem SR650/SR590 3.5" AnyBay 12-Bay Backplane	1
AURZ	ThinkSystem SR590/SR650 Rear HDD Kit	1
5977	Select Storage devices - no configured RAID required	1
B31E	ThinkSystem RAID 930-16i 8GB Flash PCIe 12Gb Adapter	1
B496	ThinkSystem 3.5" 14TB 7.2K SAS 12Gb Hot Swap 512e HDD	12
AUUX	ThinkSystem 3.5" PM963 1.92TB Entry NVMe PCIe 3.0 x4 Hot Swap SSD	2
AUMV	ThinkSystem M.2 with Mirroring Enablement Kit	1
B11V	ThinkSystem M.2 5100 480GB SATA 6Gbps Non-Hot Swap SSD	2
AURC	ThinkSystem SR550/SR590/SR650 (x16/x8)/(x16/x16) PCIe FH Riser 2 Kit	1
AUKH	ThinkSystem 1Gb 4-port RJ45 LOM	1
AUKX	ThinkSystem Intel X710-DA2 PCIe 10Gb 2-Port SFP+ Ethernet Adapter	2
AVWF	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
AUPW	ThinkSystem XClarity Controller Standard to Enterprise Upgrade	1
B0MJ	Feature Enable TPM 1.2	1
AUQB	Lenovo ThinkSystem Mainstream MB - 2U	1
B31F	ThinkSystem M.2 480GB SSD Thermal Kit	1

7.2 Networking BOM

This section contains the BOM for different types of supported networking switches.

7.2.1 NE0152T 1GbE Switch

Table 14. BOM of NE0152T

Code	Description	Quantity
7Y810011WW	Lenovo ThinkSystem NE0152T RackSwitch (Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

7.2.2 NE1032 10GbE Switch

Table 15.BOM of NE1032

Code	Description	Quantity
AU3A	Lenovo NE1032 RackSwitch (Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

7.2.3 NE1032T 10GbE Switch

Table 16. BOM of NE1032T

Code	Description	Quantity
AU38	Lenovo NE1032T RackSwitch(Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

7.2.4 NE1072T 10GbE Switch

Table 17. BOM of NE1072T

Code	Description	Quantity
AU36	Lenovo NE1072T RackSwitch(Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

7.3 Rack BOM

Table 18. BOM of Rack

Code	Description	Quantity
93634PX	42U 1100mm Enterprise V2 Dynamic Rack	1
00YJ780	0U 20 C13/4 C19 Switched and Monitored 32A 1 Phase PDU	2

7.4 Red Hat Subscription options

Table 19. BOM of Red Hat subscriptions

Code	Description	Quantity
00YH849	Red Hat Ceph Storage, 3 Physical Nodes, to 216TB, Prem RH Support, 3 years	Variable

Resources

For more information about the topics in this document, see the following resources:

- Red Hat Ceph Storage overview:
  redhat.com/en/technologies/storage/ceph
- Product Documentation for Red Hat Ceph Storage 3:
  access.redhat.com/documentation/en-us/red_hat_ceph_storage/3/
- Lenovo Data Center Solution Configurator:
  dcsc.lenovo.com

Trademarks and special notices

© Copyright Lenovo 2019.

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo, the Lenovo logo, AnyBay, AnyRAID, BladeCenter, NeXtScale, RackSwitch, Rescue and Recovery, ThinkSystem, System x, ThinkCentre, ThinkVision, ThinkVantage, ThinkPlus and XClarity are trademarks of Lenovo.

Red Hat, Red Hat Enterprise Linux and the Shadowman logo are trademarks of Red Hat, Inc., registered in the U.S. and other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries. The OpenStack mark is either a registered trademark/service mark or trademark/service mark of the OpenStack Foundation, in the United States and other countries, and is used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel Inside (logos), and Xeon are trademarks of Intel Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.