SUSE Enterprise Storage v4 Implementation Guide

Lenovo platform

Written by:
José Betancourt, SUSE

We adapt. You succeed.

Contents

Introduction

The objective of this guide is to present a step-by-step guide on how to implement SUSE Enterprise Storage (v4) on Lenovo hardware platforms. This testing was completed on the ThinkServer RD650 model specifically.

It is suggested that the document be read in its entirety, along with the supplemental appendix information before attempting the process.

The platform is built and deployed to show customers the ability to deploy a robust SUSE Enterprise Storage cluster on the Lenovo ThinkServer platform. Its goal is to show architectural best practices and how to build a Ceph-based cluster that will support the implementation of two key gateways: iSCSI and RADOS (RGW). Although our deployment took place on the RD650 platform, the implementation practices are equally applicable to other platforms such as the x3000 family.

Upon completion of the steps in this document, a working SUSE Enterprise Storage (v4) deployment will be operational as described in the SUSE Enterprise Storage 4 Deployment and Administration Guide.

There are several methods for installing a Ceph cluster with SUSE Enterprise Storage. This guide demonstrates the *ceph-deploy* methodology for cluster build and deployment.

Configuration

The equipment is deployed at Lenovo's Executive Briefing Center in Stuttgart, Germany where the SUSE Enterprise Storage software-defined solution was installed and tested.

Server infrastructure:

- Four Lenovo ThinkServer RD650 for OSD and Monitor Nodes
  - 2X E5-2630 v3 – 2.4Ghz processors (8 cores/16 threads each)
  - 256GB RAM
  - 2X 400GB SSD drives (ST400FM0053)
  - 7X1TB HDD drives (ST1000NM0023)
  - 1X 32GB SD card
  - 2X ThinkServer Intel X250-2 Converged Network Adapters (10GbE)
- One Lenovo ThinkServer RD650 as Administrative Node
  - 2X E5-2630 v3 – 2.4Ghz processors (8 cores/16 threads each)
  - 128GB RAM
  - 1 X1TB HDD drives (ST1000NM0023)
  - 1X 32GB SD card
  - 2X ThinkServer Intel X250-2 Converged Network Adapters (10GbE)

Switching infrastructure:

- Two Lenovo G8272 Top of Rack (TOR) switches. Switch details available at: http://www.lenovo.com/images/products/system-x/pdfs/datasheets/rackswitch_g8272_ds.pdf

Software:

- SUSE Enterprise Storage 4. (Please note: The SUSE Enterprise Storage subscription includes a limited use [for SUSE Enterprise Storage] entitlement for SUSE Linux Enterprise Server as well.

Target Audience

This reference architecture is focused on administrators who deploy software defined storage solutions within their data centers and making the different storage services accessible to their own customer base. By following this document as well as those referenced herein, the administrator should have a full view of the SUSE Enterprise Storage architecture, deployment and administrative tasks, with a specific set of recommendations for deployment of the hardware and networking platform.

Business problem and business value

SUSE Enterprise Storage delivers a highly scalable, resilient, self-healing storage environment designed for large scale environments ranging from hundreds of terabytes to petabytes. This software defined storage product can reduce IT costs by leveraging industry standard servers to present unified storage servicing block, file, and object protocols. Having storage that can meet the current needs and requirements of the data center while supporting topologies and protocols demanded by new web-scale applications, enables administrators to support the ever-increasing storage requirements of the enterprise with ease.

Business problem

Customers of all sizes face a major storage challenge: While the overall cost per terabyte of physical storage has gone down over the years, a data growth explosion is taking place driven by the need to access and leverage new data sources (ex: external sources such as social media) and the ability to 'manage' new data types (ex: unstructured or object data). These ever increasing "data lakes" need different access methods: File, block, or object.

Addressing these challenges with legacy storage solutions would require either a number of specialized products (usually driven by access method) with traditional protection schemes (ex: RAID). These solutions struggle when scaling from terabytes to petabytes at reasonable cost and performance levels.

Business value

This software defined storage solution enables transformation of the enterprise infrastructure by providing a unified platform where structured and unstructured data can co-exist and be accessed as file, block, or object depending on application requirements. The combination of open-source software (Ceph) and industry standard servers reduce cost while providing the on-ramp to unlimited scalability needed to keep up with future demands.

Requirements

Legacy enterprise storage systems established a high threshold of reliability, availability, and serviceability (RAS) that customers now demand from software defined storage solutions. Focusing on these capabilities help SUSE make open source technologies consumable by the enterprise. When combined with the Supermicro platform, the result is a solution that meets customer's expectation.

Functional requirements

A SUSE Enterprise Storage solution is:

- Simple to setup and deploy, within the documented guidelines of system hardware, networking and environmental prerequisites.
- Adaptable to the physical and logical constraints needed by the business, both initially and as needed over time for performance, security, and scalability concerns.
- Resilient to changes in physical infrastructure components, caused by failure or required maintenance.
- Capable of providing optimized object and block services to client access nodes, either directly or through gateway services.

Architectural overview

This architecture overview section complements the SUSE Enterprise Storage Technical Overview document available online which presents the concepts behind software defined storage and Ceph as well as a quick start guide (non-platform specific).

Solution architecture

SUSE Enterprise Storage provides unified block, file and object access based on Ceph. Ceph is a distributed storage solution designed for scalability, reliability and performance. A critical component of Ceph is the RADOS object storage. RADOS enables a number of object storage nodes to function together to store and retrieve data from the cluster using object storage techniques. The result is a storage solution that is abstracted from the hardware.

Ceph supports both native and traditional client access. The native clients are aware of the storage topology and communicate directly with the storage daemons, resulting in horizontally scaling performance. Non-native protocols, such as iSCSI, S3, and NFS require the use of gateways. These gateways can scale horizontally using load balancing techniques.

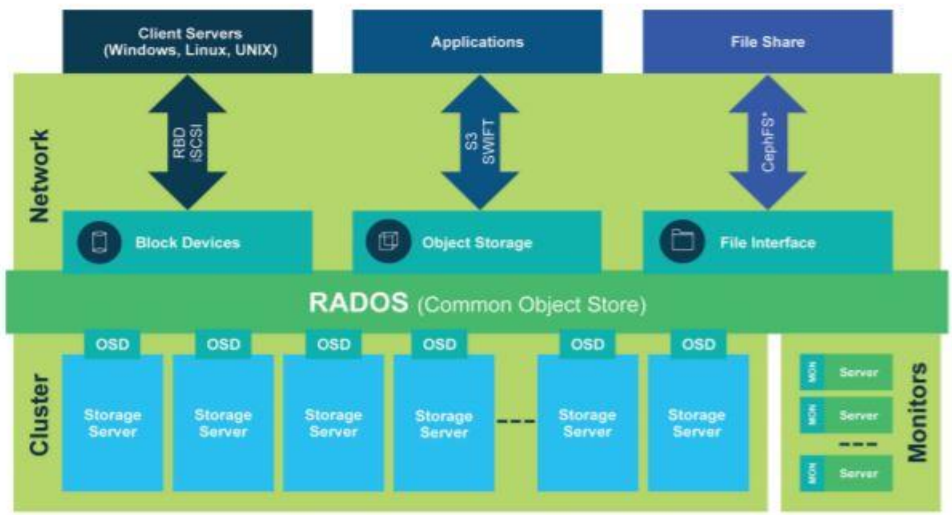


Figure 1. Ceph architecture diagram

In addition to the required network interfaces, the minimum SUSE Enterprise Storage cluster comprises of a minimum of one administration server (physical or virtual), four object storage device nodes (OSDs), three monitor nodes (MONs), and one or more Ceph Object Gateway. On installations with less than seven nodes (like our setup), four physical servers perform OSD functions with 3 of the same acting as monitors. In these instances, the minimal recommendations for a monitor node (ex: 2GB or RAM, separate non-OSD disk) need to be added to the OSD requirements when sizing the environment. For larger production implementations, monitor, gateway, and OSD workloads should be running on separate hardware, each meeting or exceeding the requirements for the size of the workload being managed. More information is available in the Systems Requirements section of the SUSE Enterprise Storage Administration and Deployment Guide.

Networking architecture

A software-defined storage solution is as reliable and performant as its slowest and least redundant component. This makes it important to design and implement a robust, high performance storage network infrastructure. From a network perspective for Ceph, this translates into:

- Separation of cluster (backend) and client-facing network traffic and isolate Ceph OSD daemon replication activities from Ceph client to storage cluster access.
- Redundancy and capacity in the form of bonded network interfaces connected to the two Lenovo G8272 Top of Rack (TOR) switches.

Figure 2 (next page) shows the logical layout of the traditional Ceph cluster implementation.

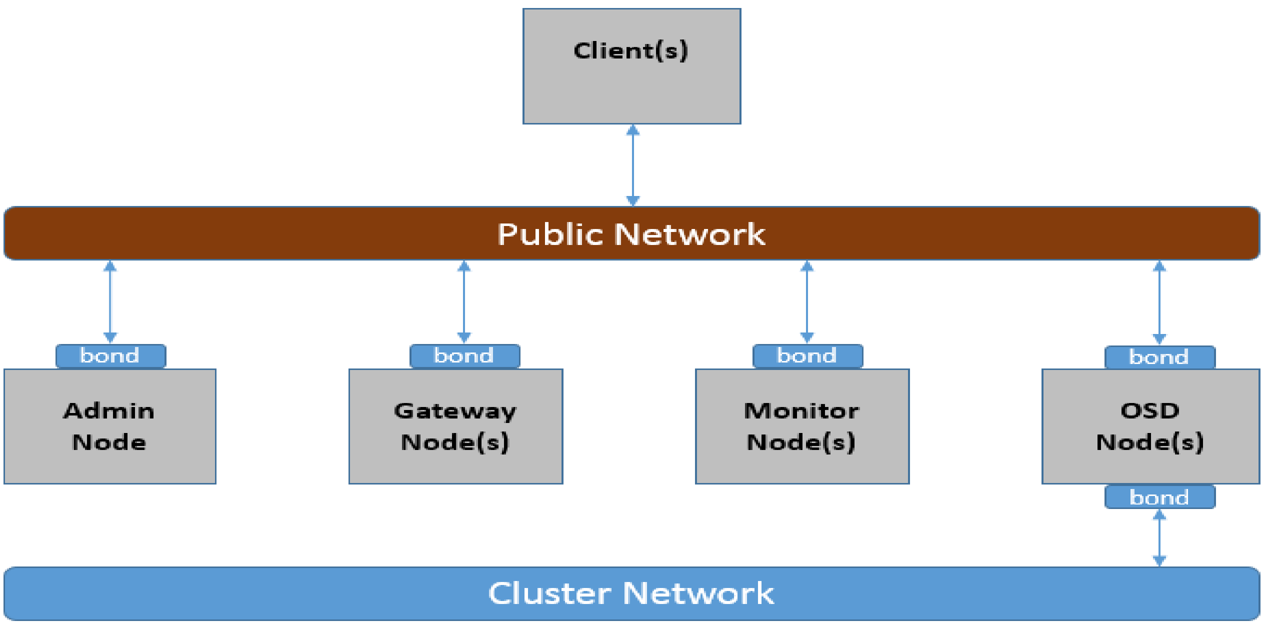


Figure 2. Sample networking diagram for Ceph cluster

Network/IP address scheme

Specific to our installation, we implemented the following naming and addressing scheme.

Function	Hostname	Primary Network	Cluster Network	Console Access
Admin Node	at-rd650a-lnx.at.eu.lenovo.com	192.168.23.61	172.16.23.61	192.168.22.61
OSD Node	at-rd650a-lnx.at.eu.lenovo.com	192.168.23.62	172.16.23.62	192.168.23.66
OSD Node	at-rd650a-lnx.at.eu.lenovo.com	192.168.23.63	172.16.23.63	192.168.23.63
OSD Node	at-rd650a-lnx.at.eu.lenovo.com	192.168.23.64	172.16.23.64	192.168.23.64
OSD Node	at-rd650a-lnx.at.eu.lenovo.com	192.168.23.65	172.16.23.65	192.168.23.65

Component model

The preceding sections provided significant details on the both the overall Lenovo hardware as well as an introduction to the Ceph software architecture. In this section, the focus is on the SUSE components: SUSE Linux Enterprise Server (SLES), SUSE Enterprise Storage (SES), and the Subscription Management Tool (SMT).

Component overview (SUSE)

- SUSE Linux Enterprise Server – A world class secure, open source server operating system, equally adept at powering physical, virtual, or cloud-based mission-critical workloads. Service Pack 2 further raises the bar in helping organizations to accelerate innovation, enhance system reliability, meet tough security requirements and adapt to new technologies.
- Subscription Management Tool for SLES12 SP2 – allows enterprise customers to optimize the management of SUSE Linux Enterprise (and extensions such as SUSE Enterprise Storage) software updates and subscription entitlements. It establishes a proxy system for SUSE Customer Center with repository and registration targets.
- SUSE Enterprise Storage – Provided as an extension on top of SUSE Linux Enterprise Server, this intelligent software-defined storage solution, powered by Ceph technology with enterprise engineering and support from SUSE enables customers to transform enterprise infrastructure to reduce costs while providing unlimited scalability.

Deployment

This deployment section should be seen as a supplement online documentation.   Specifically, the SUSE Enterprise Storage 4 Administration and Deployment Guide as well as SUSE Linux Enterprise Server Administration Guide and Subscription Management Tool (SMT) for SLES 12 SP2. Thus, the emphasis is on specific design and configuration choices.

Network Deployment overview/outline

The following considerations for the network configuration should be attended to:

- Ensure that all network switches are updated with consistent firmware versions.
- Configure 802.3ad for system port bonding and vLAG between the switches, plus enable jumbo frames on cluster network interfaces. See Appendix C for switch-side configuration key details.
- Network IP addressing and IP ranges need proper planning. In optimal environments, a single storage subnet should be used for all SUSE Enterprise Storage nodes on the primary network, with a separate, single subnet for the cluster network. Depending on the size of the installation, ranges larger than /24 may be required. When planning the network, current as well as future growth should be taken into consideration.
- Setup DNS A records for all nodes. Decide on subnets and VLANs and configure the switch ports accordingly.
- Ensure that you have access to a valid, reliable NTP service, as this is a critical requirement for all nodes. It is recommended to enable NTP on the admin node and point other nodes to it for time synchronization.

HW Deployment configuration (suggested)

The following considerations for the hardware platforms should be attended to:

- Install latest BIOS and firmware versions including the management board firmware (TSM) in the server.
- Reset BIOS to default settings (Load Optimized Defaults).
- Set BIOS Date and Time (UTC).

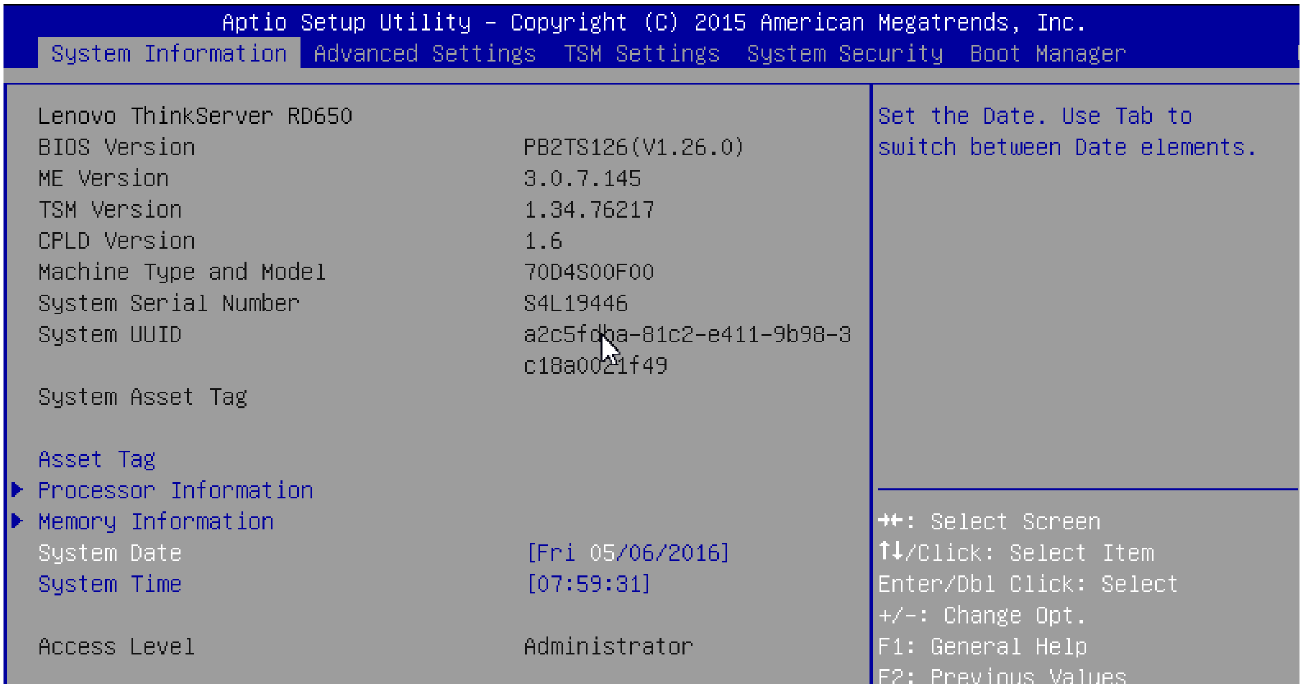


Figure 3. System Information Settings.

- Change the power settings in the Advanced Settings session from “Energy Efficiency” to “Max Performance”.

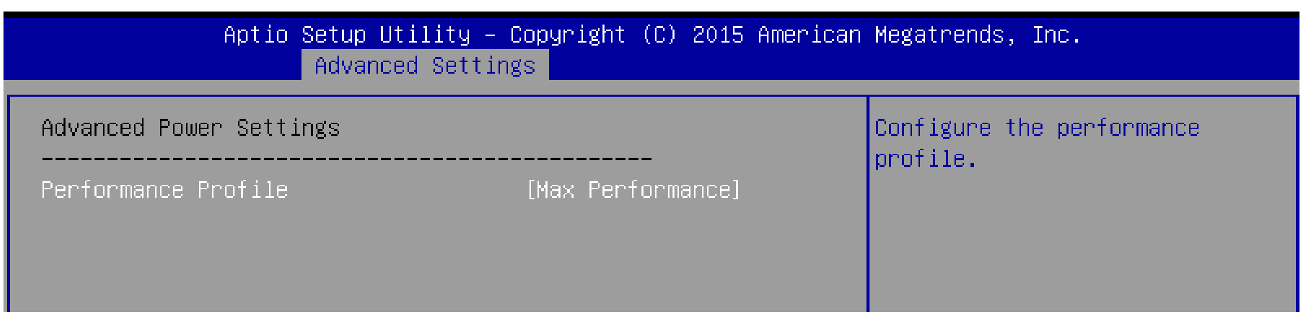


Figure 4. Advanced Settings – Performance Profile.

Operating System Deployment and Configuration

Installation of the Operating System is completed using the TSM board on the RD650 server.

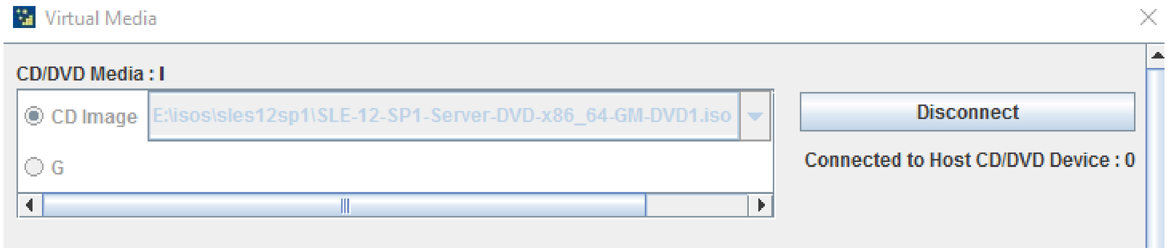


Figure 5. Mounting ISO media via Remote Console

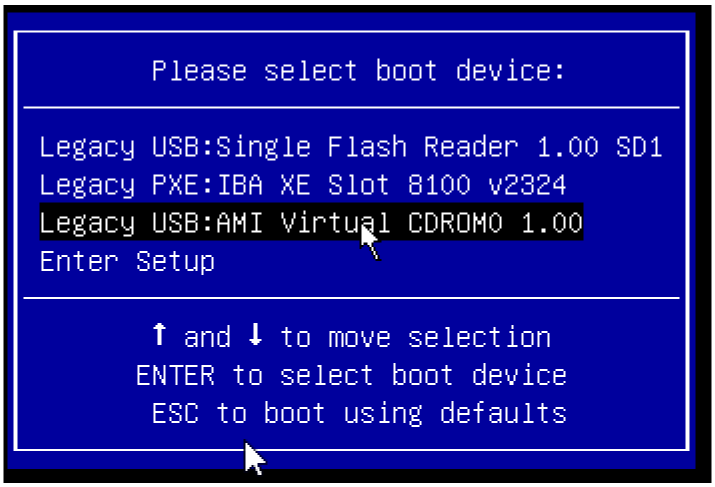


Figure 6. Booting from Virtual CDROM

Once the Operating System installation starts, perform a minimal installation and ensure that the following actions take place:

- Configure bonded interfaces. See Appendix D for OS network configuration.
- Register the system against your SMT server
- De-select AppArmor pattern from the minimal installation.
- When creating the filesystem structure for the root disk, de-select the option to have a separate */home* directory.
- Disable the firewall.
- After installation is complete, run *zypper up* to ensure all current updates are applied.

SSH and NTP Configuration

With all the nodes installed and properly configured in the network, two additional configuration steps must take place.

- SSH keys for proper communications from the admin node (at-rd650a-lnx in our instance) to all other nodes in the cluster:

```
ad-rd650a-lnx:~ # ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/root/.ssh/id_rsa):
Created directory '/root/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /root/.ssh/id_rsa.
Your public key has been saved in /root/.ssh/id_rsa.pub.
The key fingerprint is:
60:d1:21:c3:af:79:82:7b:3c:1a:15:a2:0d:b6:33:18 [MD5] root@ad-rd650a-lnx
The key's randomart image is:
+--[ RSA 2048]----+
|      .+...      |
|       .+.       |
|E o . +.         |
| + = o o.        |
|. = ...oS        |
|   o..+ .        |
|    .o o         |
|    ..+          |
|    .o .         |
+--[MD5]----------+
ad-rd650a-lnx:~ #
```

Figure 7. SSH Key generation

  - Add the public key of the server and administrators to */root/.ssh/authorized_keys*
  - Use *ssh-copy-id* to copy the key across all nodes (ex: *ssh-copy-id root@at-rd650b-lnx*). See: *man ssh-copy-id* for more information.

- NTP – Ensure time synchronization between all nodes.

SW Deployment configuration (ceph-deploy)

There are several methods to deploy a Ceph cluster with SUSE Enterprise Storage 4. This document focuses on the ***ceph-deploy*** method. Detailed information on *ceph-deploy* is available in the [SUSE Enterprise Storage Administration and Deployment Guide - Ceph-Deploy section.]

There are some preliminaries that need to be completed on all nodes:

- Create a *cephadm* user on **all nodes** (with uid 1000)
  - *useradd –m cephadm –u 1000*
  - *passwd cephadm*
  - *chown –R cephadm:users /home/cephadm*
- Modify the /etc/sudoers file and enable the user *cephadm.* Using *visudo* (as the *root* user). Ensure this is completed **across all nodes**.
  - Inside */etc/sudoers* add: *cephadm ALL=(root) NOPASSWD:ALL*

```
##
## User privilege specification
##
root ALL=(ALL) ALL
cephadm ALL=(root) NOPASSWD:ALL
```

Figure 8. *cephadm* entry in */etc/sudoers* file

- Generate ssh keys for the *cephadm* user on the admin node:
  - su – *cephadm*
  - *ssh-keygen*
  - Use *ssh-copy-id* to copy the keys across all nodes (ex: *ssh-copy-id cephadm@at-rd650b-lnx*). See man *ssh-copy-id* for more information.

SUSE Enterprise Storage installation

The next step of the software deployment is the installation of binaries across the different nodes.

To ease operations across all the nodes, create a file (ex: all.txt) with the hostnames of all the nodes that make up the cluster (including the admin node).

Before proceeding with the actual installation, ensure the *cephadm* account can ssh with no password prompts and that it can execute privileged instructions:

As user *cephadm* on admin node:

- Check for *cephadm's* ability to ssh and run *sudo* across all nodes:
  - *for SERVER in $(cat all.txt); do echo $SERVER; ssh $SERVER 'echo $HOSTNAME; sudo ls'; done*
  - Address any errors that arise (ex: unable to ssh without password or failure to execute sudo)
- Check time synchronization:
  - *for SERVER in $(cat all.txt); do echo $SERVER; ssh $SERVER 'echo $HOSTNAME; sudo ntpq –p'; done*
- Create *ceph* directory under *cephadm*'s home directory on admin node to hold keyring and configuration files:
  - *mkdir –p /home/cephadm/ceph*
- Ensure SUSE Enterprise Storage repositories are available and connected on all nodes. Once complete, proceed with the installation of the binaries:
  - On the admin node: *sudo zyper –n in ceph ceph-deploy*
  - On all other nodes: *sudo zypper –n in ceph*
- Run *ceph-deploy install* for all nodes. Use the public name of the nodes (hostname associated to the primary network.
  - *ceph-deploy install at-rd650b-lnx at-rd650c-lnx at-rd650d-lnx at-rd650e-lnx*
- Setup the monitor nodes:
  - *ceph-deploy new at-rd650c-lnx at-rd650d-lnx at-rd650e-lnx*

SUSE Enterprise Storage monitor configuration

When the steps in the SUSE Enterprise Storage installation are completed, we can proceed with the configuration of the monitor nodes. This is a two-step process:

- Adjust the *ceph.conf* file and add public and cluster networking:
  - *vi /home/cephadm/ceph/ceph.conf*
  - Set *public_network = 192.168.23.0/24* (IP Range of Admin/Primary Network)
  - Set *cluster_network = 172.16.23.0/24* (IP Range of Cluster Network)
- Configure and start the monitor agents:
  - *ceph-deploy mon create-initial*
  - To check the status of the monitors, run: *ceph –s –k /home/cephadm/ceph/ceph.client.admin.keyring*

Note: To ensure that *ceph* commands always point to the correct admin keyring file, create an alias in *cephadm*'s *.bashrc* file:

- *alias ceph='ceph –k /home/cephadm/ceph/ceph.client.admin.keyring –c /home/cephadm/ceph/ceph.conf'*

SUSE Enterprise Storage – Object Storage Daemons

The final step in the basic Ceph cluster setup is the setup of the Object Storage Daemons (OSDs). This guide focuses on the more traditional replication setup. The Enterprise SAS HDDs on each physical object storage node will be designated as an Object Storage Daemon (OSD) while the Enterprise SSDs are leveraged as journal devices for the OSDs.

Before using any of the disks as an OSD, ensure that no prior GPT and/or MBR data structures exists so that the disk is ready for partitioning via *ceph-deploy*.

Complete the following steps as user *cephadm* at */home/cephadm/ceph* on the admin node.

- Zap all the drives using *ceph-deploy* from the admin server:
  - Ex: *ceph-deploy disk zap <HOSTNAME>:<DISK>*
  - Ex: *ceph-deploy disk zap at-rd650b-lnx:sda*
  - Repeat for all drives to be used as OSD or journal

- Prepare OSD drives and journals on every OSD node:
  - *ceph-deploy osd --overwrite-conf osd prepare <HOSTNAME>:<OSD>:<JOURNAL>*
  - Example of six OSD (HDDs) with two journals:
    - *ceph-deploy -- overwrite-conf osd prepare <HOSTNAME>:sdc:/dev/sda*
    - *ceph-deploy --overwrite-conf osd prepare <HOSTNAME>:sdd:/dev/sda*
    - *ceph-deploy --overwrite-conf osd prepare <HOSTNAME>:sde:/dev/sda*
    - *ceph-deploy --overwrite-conf osd prepare <HOSTNAME>:sdf:/dev/sdb*
    - *ceph-deploy --overwrite-conf osd prepare <HOSTNAME>:sdg:/dev/sdb*
    - *ceph-deploy --overwrite-conf osd prepare <HOSTNAME>:sdh:/dev/sdb*

  - Repeat for every OSD device on every OSD node server. Our implementation has two SSDs acting as journals for six HDDs (each SSD journaling for three HDDs).

  Optional: Run the command *cat /proc/partitions* after executing the *ceph-deploy osd prepare* command to see the partitions created on the OSD drives and journals. The newly-created partitions will be used during the activation stage.

- Activate the OSD drives:
  - *ceph-deploy --overwrite-conf osd activate <HOSTNAME>:<OSD partition>:<JOURNAL partition>*
  - Example of the same six OSD devices and two journals activation:
    - *ceph-deploy --overwrite-conf osd activate <HOSTNAME>:sdc1:/dev/sda1*
    - *ceph-deploy --overwrite-conf osd activate <HOSTNAME>:sdd1:/dev/sda2*
    - *ceph-deploy --overwrite-conf osd activate <HOSTNAME>:sde1:/dev/sda3*
    - *ceph-deploy --overwrite-conf osd activate <HOSTNAME>:sdf1:/dev/sdb1*
    - *ceph-deploy --overwrite-conf osd activate <HOSTNAME>:sdg1:/dev/sdb2*
    - *ceph-deploy --overwrite-conf osd activate <HOSTNAME>:sdh1:/dev/sdb3*

  - Repeat for every OSD device on every OSD node server.

Post-deployment quick test

The steps below can be used (regardless of the deployment method) to validate the overall cluster health:

- *ceph status*
- *ceph osd pool create test 4096*
- *rados bench –p test 300 write --no-cleanup*
- *rados bench –p test 300 seq*

Once the tests are complete, you can remove the test pool via:

- *ceph osd pool delete test --yes-i-really-really-mean-it*

Deployment Considerations

Some final considerations before deploying your own version of a SUSE Enterprise Storage cluster, based on Ceph. As previously stated, please refer to the Administration and Deployment Guide

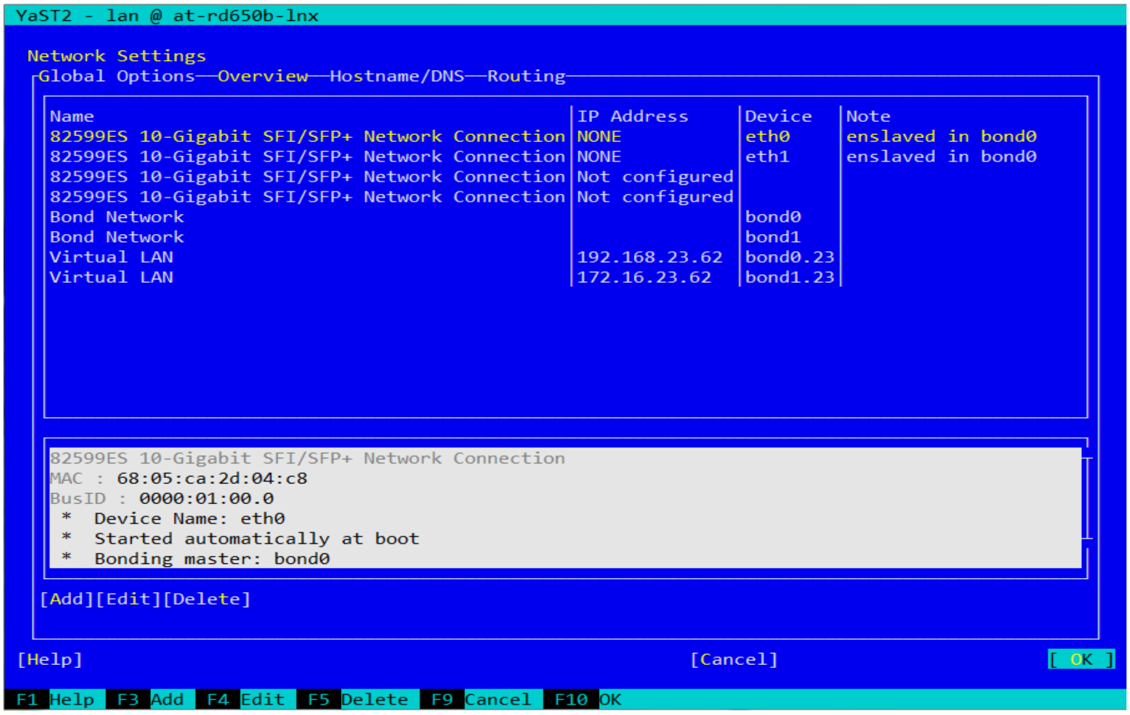
- This guide is focused on *ceph-deploy* the deployment mechanism. Do not mix deployment methods within a cluster.
- With the default replication setting of 3, remember that the client-facing network will have about half or less of the traffic of the backend network. This is especially true when component failures occur or rebalancing happens on the OSD nodes. For this reason, it is important not to under provision this critical cluster and service resource.
- It is important to maintain the minimum number of MON nodes at three. As the cluster increases in size, it is best to increment in pairs, keeping the total number of Mon nodes as an odd number. However, only really large or very distributed clusters would likely need beyond the 3 Mon nodes cited in this reference implementation. For performance reasons, it is recommended to use distinct nodes for the MON roles, so that the OSD nodes can be scaled as capacity requirements dictate.
- As described in this implementation guide as well as the SUSE Enterprise Storage documentation, a minimum of four OSD nodes is recommended, with the default replication setting of 3. This will ensure cluster operation, even with the loss of a complete OSD node. Generally speaking, performance of the overall cluster increases as more properly configured OSD nodes are added.

Appendix A: Bill of Materials

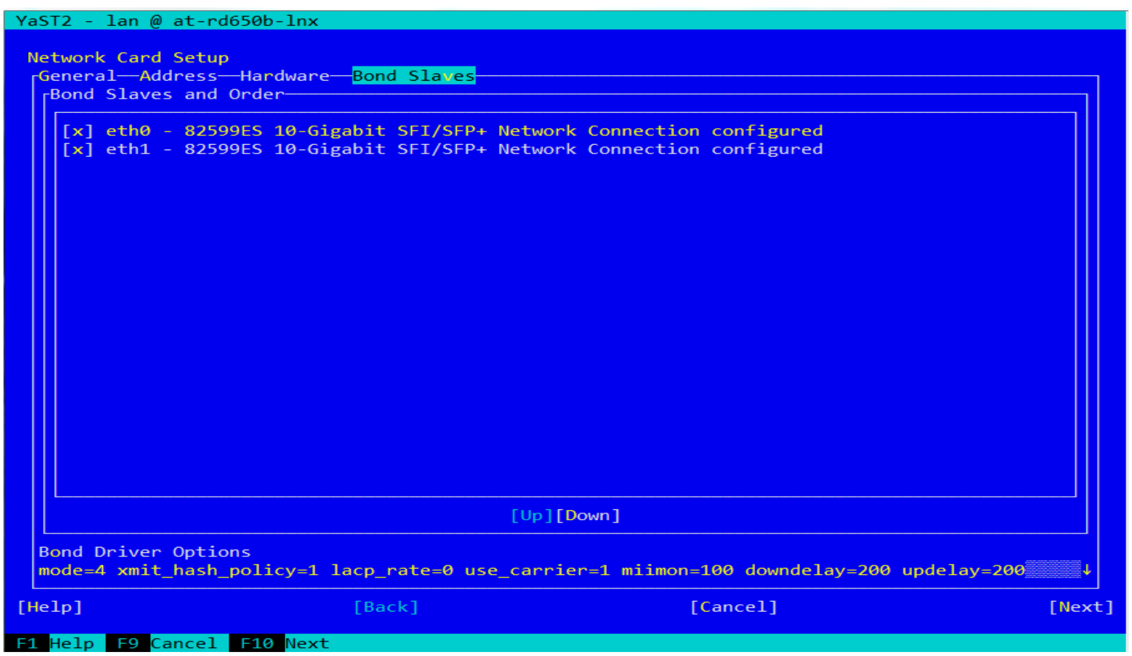
Component / System

Role	Qty	Component	Notes
Admin server	1	ThinkServer RD650	- 2X E5-2630 v3 2.4Ghz (8 cores/16 threads) - 128 GB RAM - 1X 32GB SD card - 2X 400GB SSD drives (ST400FM0053) - 7X 1TB HDD drives (ST1000NM0023) - 2X ThinkServer Intel X250-2 CAN (10GbE)
OSD/MON Hosts	4	ThinkServer RD650	Each servers consists of: - 2X E5-2630 v3 2.4Ghz (8 cores/16 threads) - 256 GB RAM - 1X 32GB SD card - 2X 400GB SSD drives (ST400FM0053) - 7X 1TB HDD drives (ST1000NM0023) - 2X ThinkServer Intel X250-2 CAN (10GbE)
Network	2	Lenovo G8272 Switch	- 48X SFP+ 10GbE - 6X QSFP+ 40GbE
Software	1	SUSE Enterprise Storage Subscription Base configuratio	10 subscriptions provided with base configuration with the folloging configuration: - Up to four OSD nodes Up to six instances for SES infrastructure nodes (MON, Gateways, Admin)

Appendix B: OS Network Configuration



YaST view of network interfaces for OSD node.



Network bond sample (bond0 – public network)

Appendix C: Network switches configuration files

Maximum frame size on the ToR switches was adjusted from1518 to 9216. The switches were also configured for bonding (802.3ad).

vLAG setup between switches.

```
vLAG already configured:

!

vlag enable

vlag tier-id 10

vlag hlthchk peer-ip 192.168.22.250

vlag isl adminkey 300

vlag adminkey 80 enable
```

LACP setup on switch ports for servers.

```
interface port 1

lacp mode active

lacp key 101

!

interface port 7

lacp mode active

lacp key 101

!

interface port 2

lacp mode active

lacp key 102

!

interface port 8

lacp mode active

lacp key 102

!

interface port 3

lacp mode active

lacp key 103
```

```
!
interface port 9
lacp mode active
lacp key 103
!
interface port 4
lacp mode active
lacp key 104
!
interface port 10
lacp mode active
lacp key 104
!
interface port 5
lacp mode active
lacp key 105
!
interface port 11
lacp mode active
lacp key 105
!
interface port 6
lacp mode active
lacp key 106
!
interface port 12
lacp mode active
lacp key 106
vlag adminkey 101 enable
vlag adminkey 102 enable
vlag adminkey 103 enable
vlag adminkey 104 enable
```

```
vlag adminkey 105 enable

vlag adminkey 106 enable
```

Resources:

[SUSE Enterprise Storage v4 - Administration and Deployment Guide]

[SUSE Linux Enterprise Server 12 SP2 - Administration Guide]

[Subscription Management Tool for SLES 12 SP2]

[Ceph - Open Source Project Documentation]

[Lenovo ThinkServer RD650 (E5-2600 v3) Product Guide]

[ThinkServer RD650 User Guide and Hardware Maintenance Manual]