

Contents


1	Introduction	2
2	Business problem and business value	4
3	Requirements	5
4	Architectural overview	5
5	Component model	9
6	Deployment	10
7	Deployment Considerations	18
8	Appendix A: Bill of Materials	19
9	Appendix B: OSD Drive and Journal Proposal Changes	21
10	Appendix C: Policy.cfg	24
11	Appendix D: OS Network Configuration	25
12	Appendix E: Network switches configuration files	28
13	Resources:	37
14		37

1 Introduction

The objective of this guide is to present a step-by-step guide on how to implement SUSE Enterprise Storage (v4) on Supermicro hardware platforms.

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 Supermicro 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).

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 \(https://www.suse.com/documentation/ses-4/\)](https://www.suse.com/documentation/ses-4/) .


There are several methods for installing a Ceph cluster with SUSE Enterprise Storage. This guide demonstrates SUSE Enterprise Storage's preferred approach, based on *Salt* technology.

1.1 Configuration

The reference architecture described was built as a joint effort between Ingram Micro, Supermicro and SUSE. The equipment was deployed to Ingram Micro's briefing center in Buffalo, NY where the SUSE Enterprise Storage software-defined solution was installed and tested.

The SUSE Enterprise Storage cluster leveraged three family of Supermicro servers. The role/functionality of each SUSE Enterprise Storage component will be explained in more detail in the architectural overview section.

For Ceph admin and monitor functions:

- One Supermicro SuperServer 6028TR-HTR (<https://www.supermicro.com/products/system/2u/6028/sys-6028tr-htr.cfm>)  system with 4-node capacity.

For RADOS (RGW) and iSCSI Gateway functions:

- Two Supermicro SuperServer 1028TP-DTR (<https://www.supermicro.com/products/system/1U/1028/SYS-1028TP-DTR.cfm>)  systems with 2-node capacity.

For the Object Store Device (OSD) function:

- Four Supermicro SuperStorage Server 6028R-E1CR24L (<https://www.supermicro.com/products/system/2U/6028/SSG-6028R-E1CR24L.cfm>) .

1.2 Switching infrastructure:

- Cisco Nexus 9000 switches.

1.3 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.

2 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.

2.1 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.

2.2 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.

3 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.

3.1 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.

4 Architectural overview

This architecture overview section complements the [SUSE Enterprise Storage Technical Overview \(https://www.suse.com/docrep/documents/1mdg7eq2kz/suse_enterprise_storage_technical_overview_wp.pdf\)](https://www.suse.com/docrep/documents/1mdg7eq2kz/suse_enterprise_storage_technical_overview_wp.pdf)  document available online which presents the concepts behind software defined storage and Ceph as well as a quick start guide (non-platform specific).

4.1 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.

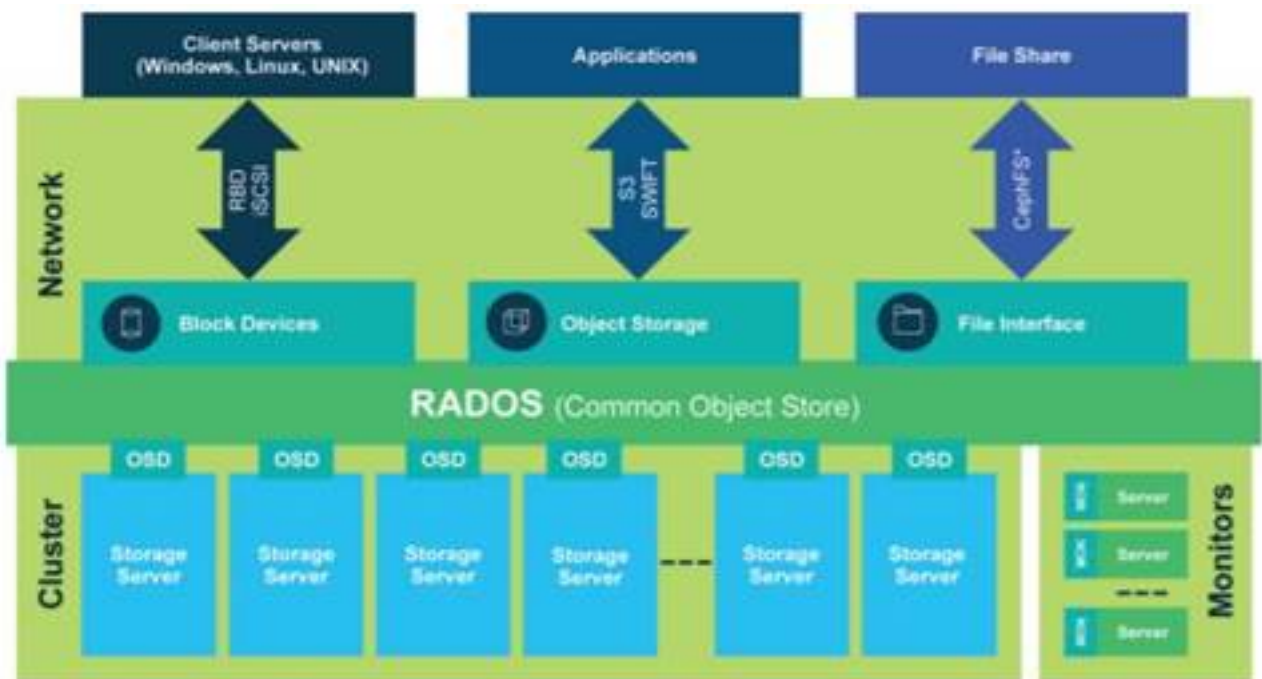



Figure 1. Ceph architecture diagram

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. Specific to our implementation:

- One of the nodes on the Supermicro SuperServer 6028TR-HTR (<https://www.supermicro.com/products/system/2u/6028/sys-6028tr-htr.cfm>)  server is deployed as our administrative physical host server. The administration server is used to deploy and configure SUSE

Enterprise Storage on the other nodes (OSDs, MONs and Object Gateways). openAttic, the central management system which supports Ceph needs to be installed on the administration server as well.

- The other three nodes on the same Supermicro SuperServer 6028TR-HTR (<https://www.supermicro.com/products/system/2u/6028/sys-6028tr-htr.cfm>)  server are deployed as monitor (MONs) nodes. Monitor nodes maintain information about the cluster health state, a map of the other monitor nodes and a CRUSH map. They also keep history of changes performed to the cluster.
- One pair of nodes on a Supermicro SuperServer 1028TP-DTR (<https://www.supermicro.com/products/system/1U/1028/SYS-1028TP-DTR.cfm>)  server acts as our iSCSI gateway. iSCSI is a storage area network (SAN) protocol that allows clients (called initiators) to send SCSI command to SCSI storage devices (targets) on remote servers. SUSE Enterprise Storage Server includes a facility that open Ceph storage management to heterogeneous clients such as Microsoft Windows and VMware vSphere through the iSCSI protocol. These systems may scale horizontally through client usage of multi-path technology.
- Another set of the same type of Supermicro server performs the duties of RADOS gateway. As the documentation states (Ceph RADOS Gateway (https://www.suse.com/documentation/ses-4/book_storage_admin/data/cha_ceph_gw.html) ) *“Ceph RADOS Gateway is an object storage interface built on top of librados to provide applications with a RESTful gateway to Ceph Storage Clusters.”*
- Data is stored on four Supermicro SuperStorage Server 6028R-E1CR24L (<https://www.supermicro.com/products/system/2U/6028/SSG-6028R-E1CR24L.cfm>)  servers categorized as storage nodes. The nodes contain multiple storage devices that are each assigned an Object Storage Daemon (OSD). The OSD daemon assigned to the OSD stores data and manages the data replication and rebalancing processes OSD daemons also communicate with the monitor (MON) nodes and provide them with the state of the other OSD daemons.

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 Cisco Nexus 9000 switches.

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

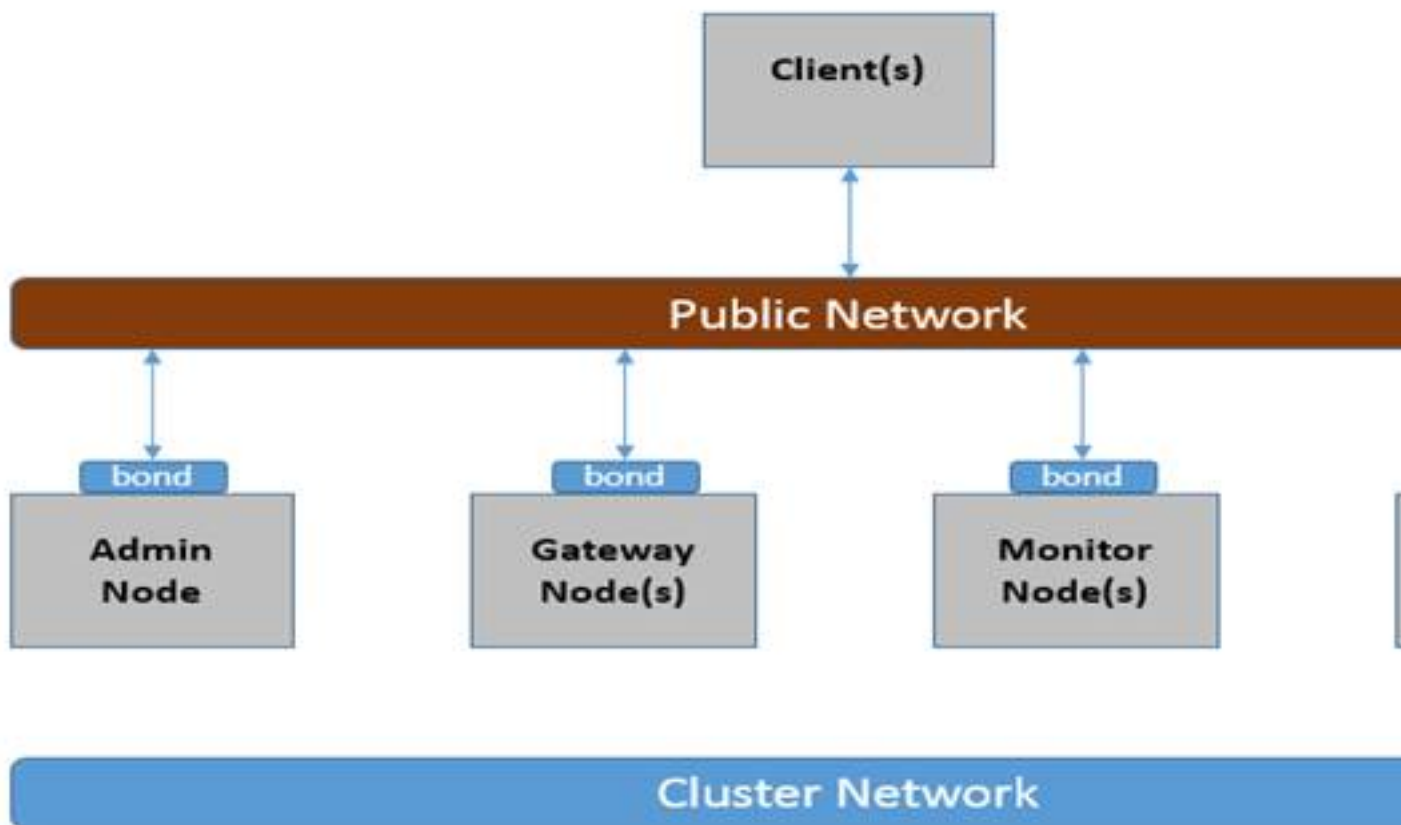


Figure 2. Sample networking diagram for Ceph cluster

4.2 Network/IP address scheme

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

Function	Hostname	Primary Network	Cluster Network	IPMI Network
<i>Admin (Host)</i>	admin.suse.imsc.int	192.168.145.10	N/A	192.168.145.110
<i>Monitor</i>	mon1.suse.imsc.int	192.168.145.11	N/A	192.168.145.111
<i>Monitor</i>	mon2.suse.imsc.int	192.168.145.12	N/A	192.168.145.112
<i>Monitor</i>	mon3.suse.imsc.int	192.168.145.13	N/A	192.168.145.113
<i>RADOS Gateway</i>	rgw1.suse.imsc.int	192.168.145.14	N/A	192.168.145.114
<i>RADOS Gateway</i>	rgw2.suse.imsc.int	192.168.145.15	N/A	192.168.145.115
<i>iSCSI Gateway</i>	igw1.suse.imsc.int	192.168.145.16	N/A	192.168.145.116
<i>iSCSI Gateway</i>	igw2.suse.imsc.int	192.168.145.17	N/A	192.168.145.117
<i>OSD Node</i>	osd1.suse.imsc.int	192.168.145.21	192.168.146.21	192.168.145.121
<i>OSD Node</i>	osd2.suse.imsc.int	192.168.145.22	192.168.146.22	192.168.145.122
<i>OSD Node</i>	osd3.suse.imsc.int	192.168.145.23	192.168.146.23	192.168.145.123
<i>OSD Node</i>	osd4.suse.imsc.int	192.168.145.24	192.168.146.24	192.168.145.124

5 Component model

The preceding sections provided significant details on the both the overall Supermicro 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 entitlement. 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.

6 Deployment

This deployment section should be seen as a supplement online [documentation](https://www.suse.com/documentation/). Specifically, the [SUSE Enterprise Storage 4 Administration and Deployment Guide](https://www.suse.com/documentation/ses-4/book_storage_admin/data/book_storage_admin.html) as well as [SUSE Linux Enterprise Server Administration Guide](https://www.suse.com/documentation/sles-12/book_sle_admin/data/book_sle_admin.html) and [Subscription Management Tool \(SMT\) for SLES 12 SP2](https://www.suse.com/documentation/sles-12/book_smt/data/book_smt.html). Thus, the emphasis is on specific design and configuration choices.

6.1 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 E for the switch-side configuration.

- 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.

6.2 HW Deployment configuration (suggested)

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

- Ensure Boot Mode is set to 'UEFI' for all the physical nodes that comprise the SUSE Enterprise Storage Cluster.

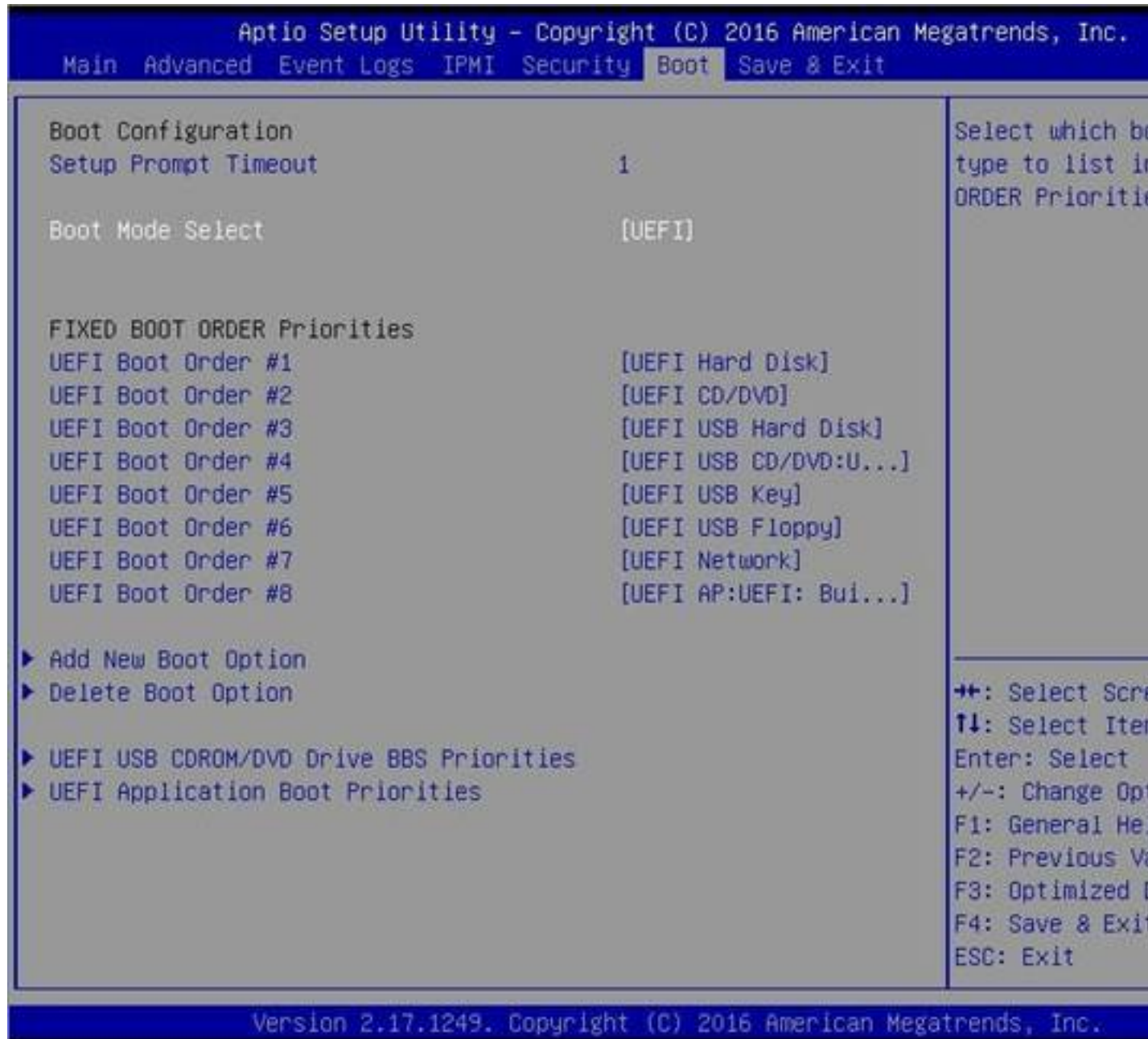


Figure 3. UEFI settings

- Follow the Supermicro SATA DOM (SuperDOM) Endurance Use Cases (https://www.supermicro.com/datasheet/datasheet_SuperDOM.pdf) [↗](#) recommendations for the OS installation:
- Verify BIOS/uEFI level on the physical servers correspond to those on the SUSE YES certification for the Supermicro platform:
 - SYS-6028TR-HTR - <https://www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=145180> [↗](#)
 - SYS-1028TP-DTR - <https://www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=145197> [↗](#)
 - 6028R-E1CR24L - <https://www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=145178> [↗](#)

6.3 Operating System Deployment and Configuration

Installation of the Operating System is completed using the Supermicro remote console. We mount the ISO image and proceed with boot to the Operating System:



Figure 4. Mounting ISO media via Remote Console

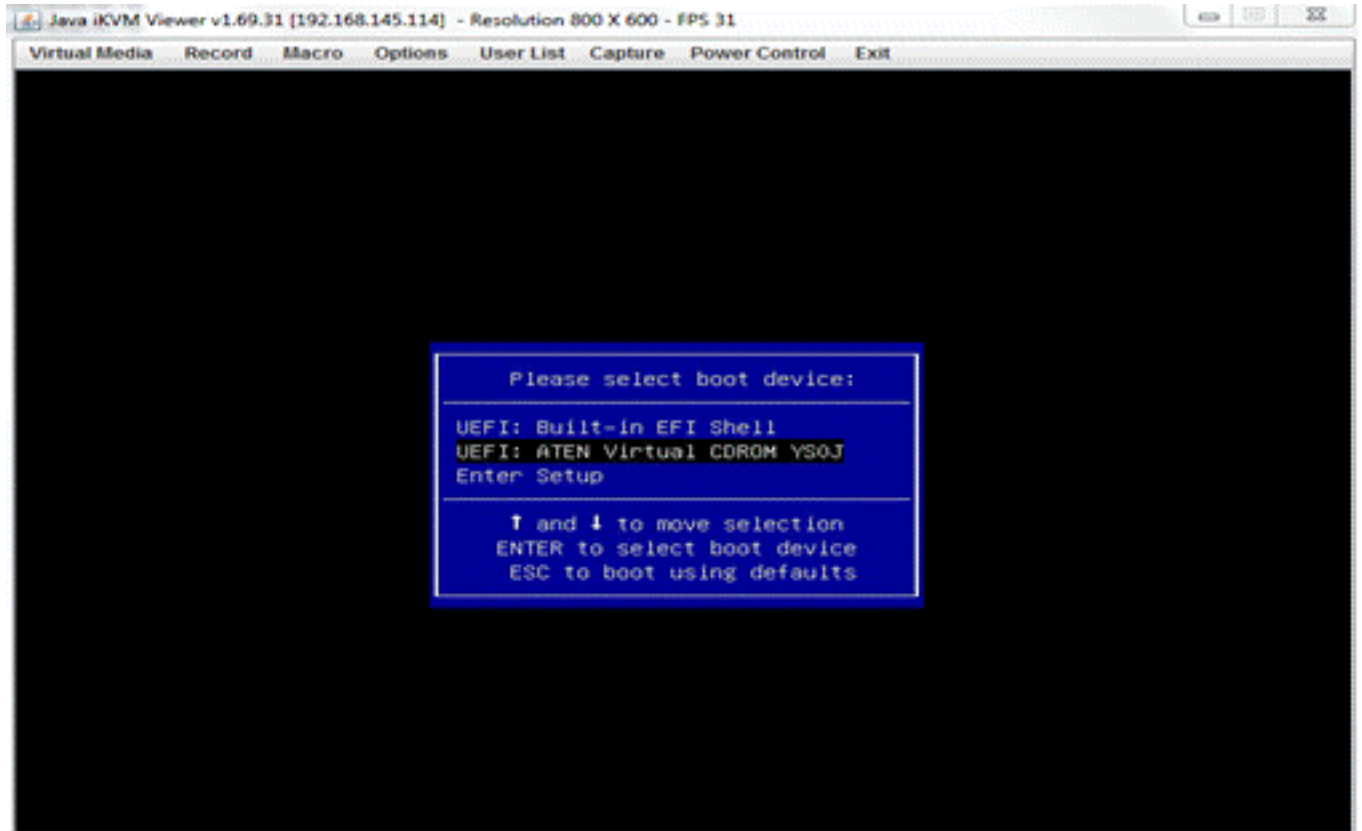


Figure 5. Booting from Virtual CDROM

The Operating System will be installed on the SATA Disk on Module (SuperDOM). Based on best practices for SuperDOM devices:

- No swap space or swap files on the SuperDOM – In general, avoid excessive writes. The nodes have three 3.5” disk drive slots where HDD/SDDs can be placed for any write-intensive activity.
- No RAID 1. Instead, copy the OS installation from the OS DOM to the “backup” DOM and have it ready for an alternate boot in the event of a failure on the main DOM. To accomplish this, follow these instructions after the SUSE Linux Enterprise Server installation is complete.
 - Boot from the SLES 12 SP2 DVD and select Rescue Mode.
 - Enter *root* at the login prompt and hit Enter.
 - Determine the device names for the Supermicro SATA DOMs using *lsscsi*.


- Copy the entire drive contents from the boot drive to the backup drive using the command `dd if= <OS drive device name> of= <backup drive device name> bs=4M`
 - Example: `dd if=/dev/sda of=/dev/sdb bs=4M`
- Reboot.
- Note: Consider setting a process in place to synchronize these devices periodically as well as after major changes (ex: patch updates).

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.

6.4 SW Deployment configuration (DeepSea and Salt)

Salt along with *DeepSea* is a stack of components that help deploy and manage server infrastructure. It is very scalable, fast, and relatively easy to get running.

There are three key *Salt* imperatives that need to be followed and are described in detail in section 4 ([Deploying with DeepSea and Salt \(https://www.suse.com/documentation/ses-4/book_storage_admin/data/ceph_install_saltstack.html\)](https://www.suse.com/documentation/ses-4/book_storage_admin/data/ceph_install_saltstack.html) ):

- The *Salt Master* is the host that controls the entire cluster deployment. Ceph itself should NOT be running on the master as all resources should be dedicated to Salt master services. In our scenario, we used the Admin server as the Salt master.
- *Salt minions* are nodes controlled by Salt master. OSD, monitor, and gateway nodes are all Salt minions in this installation.
- Salt minions need to correctly resolve the Salt master's host name over the network. This can be achieved using unique DNS names for the various interfaces or by having the unique names in `/etc/hosts` files local to each node.

DeepSea consists of series of *Salt* files to automate the deployment and management of a Ceph cluster. It consolidates the administrator's decision making in a single location around cluster assignment, role assignment and profile assignment. *Deepsea* collects each set of tasks into a goal or stage.

The following steps, performed in order will be used for this reference implementation:

Base assumption: SLES12 SP2 and SUSE Enterprise Storage 4 extension installed and register on each node of the cluster (including the Admin server):

```
Install DeepSea on the Salt master (our Admin VM):  
zypper in deepsea
```

This command installs the *salt-master* package on the Admin node as well

```
Start the salt-master service and enable:  
systemctl start salt-master.service  
systemctl enable salt-master.service
```

- Install the *salt-minion* on all cluster nodes (including the Admin):

```
zypper in salt-minion
```

- Configure all minions to connect to the Salt master: Modify the entry for *master* in the `/etc/salt/minion`

- Ex: In our case: *master: admin.suse.imsc.int*
- Start the *salt-minion* service and enable:
 - *systemctl start salt-minion.service*
 - *systemctl enable salt-minion.service*
- Verify that the Salt state file */srv/pillar/ceph/master_minion.sls* points to the Salt master – Restart the master service if changes are made (*systemctl restart salt-master.service*)
- Accept all salt keys on the Salt master: *salt-key --accept-all* and verify their acceptance (*salt-key --list-all*)
- If the OSD nodes were used in a prior installation, zap ALL the OSD disks (*ceph-disk zap <DISK>*)
- At this point, you can deploy and configure the cluster:
 - Prepare the cluster: *salt-run state.orch ceph.stage.prep*
 - Run the discover stage to collect data from all minions and create configuration fragments:
 - *salt-run state.orch ceph.stage.discovery*
 - As the discovery process completes, there are two cluster-specific changes to be performed:
 - The proposed data and journal profile proposed for the Supermicro OSD hosts does not account for the NVME and the SATA DoM devices.
 - A */srv/pillar/ceph/proposals/policy.cfg* file needs to be created to instruct Salt on the location and configuration files to use for the different components that make up the Ceph cluster (Salt master, admin, monitor, and OSDs).
 - See Appendix B and C for illustrations on data and journal profile changes as well as the *policy.cfg* file used in the installation.
 - Next, proceed with the configuration stage to parse the *policy.cfg* file and merge the included files into the final form

- *salt-run state.orch ceph.stage.configure*
- The last two steps manage the actual deployment. Deploy monitors and ODS daemons first:
 - *salt-run state.orch ceph.stage.deploy* (Note: The command can take some time to complete, depending on the size of the cluster).
 - Check for successful completion via: *ceph -s*
 - Finally, deploy the services(gateways [iSCSI, RADOS], and openATTIC to name a few): *salt-run state.orch ceph.stage.services*

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*

7 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 \(https://www.suse.com/documentation/ses-4/book_storage_admin/data/book_storage_admin.html\)](https://www.suse.com/documentation/ses-4/book_storage_admin/data/book_storage_admin.html) 

- This guide is focused on *Salt* as the preferred 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.

8 Appendix A: Bill of Materials

Component / System

Role	Qty	Component	Notes
<i>Admin/MON servers</i>	1*	SYS-6028TR-HTR	<p>*One enclosure with 4 blade nodes. Node consists of:</p> <ul style="list-style-type: none"> • 1X E5-2623V4 26G CPU • 2X 8GB DD4-2400 ECC REG DIMM • 2X SMC SATA3 DOM 64GB MLC • 1X Dual-port 10G Ethernet w SFP + W/ CDR (AOC-STGN-I2S) • Assembly and Testing

<i>Gateways</i>	2**	SYS-1028TP-DTR	<p>** One enclosure with 2 blade nodes. Node consists of:</p> <ul style="list-style-type: none"> • 1X E5-2620V4 2.1G CPU • 8X 8GB DDR4-2400 ECC REG DIMM • 1X SMC SATA3 DOM 64GB MLC • 1X Dual-port 10G Ethernet w SFP + W/ CDR (AOC-STGN-I2S) • Assembly and Testing <p>Note: 2 enclosures = 4 gateway nodes (2 iSCSI, 2 RGW)</p>
<i>OSD Hosts</i>	4	SSG-6028R-E1CR24L	<p>Each servers consists of:</p> <ul style="list-style-type: none"> • 2X E2630V4 2.2G CPU • 8X 32GB DD4-2400 ECC REG DIMM • 2X SMC SATA3 DOM 64GB MLC • 24X Toshiba 3.5" 6TB 7.2K RPM SATA 128M 512e HDD • 2X Intel DC P3600 400GB NVMe PCIe3.0 , MLC AIC SSD • 1X Dual-port 10G Ethernet w SFP + W/ CDR (AOC-STGN-I2S) • 1X `SIOM 2-port 10G SFP + , Intel 82599ES Controller Add-on Card (AOC-MTGN-I2S) • Assembly and Testing

Software	1	SUSE Enterprise Storage Subscription Base configuration	10 subscriptions provided with base configuration with the folloing configuration: <ul style="list-style-type: none"> • Up to four OSD nodes • Up to six instances for SES infrastructure nodes (MON, Gateways, Admin)
Software	2	SUSE Enterprise Storage Subscription Expansion nodes	2 additional subscriptions to cover remaining SES infrastructure nodes

Note: The computer room where the equipment is located has redundant networking equipment allowing the Ceph cluster to be configured according to SUSE Enterprise Storage best practices.

9 Appendix B: OSD Drive and Journal Proposal Changes

The proposal generated by *salt-run state.orch ceph.stage.discovery* does not accurately reflect the environment. The file is listed below highlighting needed changes. Entries with */* */* are author's comments.

storage:

data + journals: [] */*NOTE: Disks and journals should be listed under the data + journals category */*

osds: / NOTE: No drives should be under the osds entry. */*

- /dev/disk/by-id/ata-Supermicro_SSD_SMC0515D93716CAM3007 / NOTE: Should not be listed */*

- /dev/disk/by-id/scsi-3500003973bd81dee

- /dev/disk/by-id/scsi-3500003973bf81032

- /dev/disk/by-id/scsi-3500003973bd81e0a
- /dev/disk/by-id/scsi-3500003973bd81dfe
- /dev/disk/by-id/scsi-3500003973bf01991
- /dev/disk/by-id/scsi-3500003973bf81033
- /dev/disk/by-id/scsi-3500003973bf81035
- /dev/disk/by-id/scsi-3500003973b8810e9
- /dev/disk/by-id/scsi-3500003973bf81039
- /dev/disk/by-id/scsi-3500003973b701f01
- /dev/disk/by-id/scsi-3500003973b701f03
- /dev/disk/by-id/scsi-3500003973b8810ea
- /dev/disk/by-id/scsi-3500003973b78171f
- /dev/disk/by-id/scsi-3500003973bf01a20
- /dev/disk/by-id/scsi-3500003973bf01997
- /dev/disk/by-id/scsi-3500003973bf81023
- /dev/disk/by-id/scsi-3500003973bf80ef7
- /dev/disk/by-id/scsi-3500003973bd81dfd
- /dev/disk/by-id/scsi-3500003973bf0193a
- /dev/disk/by-id/scsi-3500003973bf81034
- /dev/disk/by-id/scsi-3500003973bd81de9
- /dev/disk/by-id/scsi-3500003973bf81026
- /dev/disk/by-id/scsi-3500003973bf81024
- /dev/disk/by-id/scsi-3500003973bf81025
- /dev/disk/by-id/nvme-SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN /*NVMEs should be journals */
- /dev/disk/by-id/nvme-SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN /*NVME should be journals */

We proceeded to make changes to the file so that we have a successful *configure* step. The modified file below:

storage:

data + journals:

```
- /dev/disk/by-id/scsi-3500003973bd81dee: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973bf81032: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973bd81e0a: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973bd81dfe: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973bf01991: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973bf81033: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973bf81035: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973b8810e9: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973bf81039: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973b701f01: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973b701f03: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973b8810ea: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD6351009T400AGN
- /dev/disk/by-id/scsi-3500003973b78171f: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN
- /dev/disk/by-id/scsi-3500003973bf01a20: /dev/disk/by-id/nvme-
SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN
```

-	/dev/disk/by-id/scsi-3500003973bf01997:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	
-	/dev/disk/by-id/scsi-3500003973bf81023:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	
-	/dev/disk/by-id/scsi-3500003973bf80ef7:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	
-	/dev/disk/by-id/scsi-3500003973bd81dfd:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	
-	/dev/disk/by-id/scsi-3500003973bf0193a:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	
-	/dev/disk/by-id/scsi-3500003973bf81034:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	
-	/dev/disk/by-id/scsi-3500003973bd81de9:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	
-	/dev/disk/by-id/scsi-3500003973bf81026:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	
-	/dev/disk/by-id/scsi-3500003973bf81024:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	
-	/dev/disk/by-id/scsi-3500003973bf81025:	/dev/disk/by-id/nvme-
	SNVMe_INTEL_SSDPEDME40CVMD635100DM400AGN	

osds: []

NOTE: There is ONE space between the colon separating the OSD and journal entries. Accurate spacing is important with *Salt*.

10 Appendix C: Policy.cfg

policy.cfg file

cluster-ceph/cluster.sls

role-master/cluster/adminvm01.suse.imsc.int.sls

role-admin/cluster/adminvm01.suse.imsc.int.sls

role-mon/stack/default/ceph/minions/mon*.yaml

role-mon/cluster/mon*.sls


```
role-igw/stack/default/ceph/minions/igw*.yaml
role-igw/cluster/igw*.sls
role-rgw/cluster/rgw*.sls
config/stack/default/global.yaml
config/stack/default/ceph/cluster.yaml
profile-1Supermicro59GB-2Intel372GB-24TOSHIBA5589GB-1/cluster/*.sls
profile-1Supermicro59GB-2Intel372GB-24TOSHIBA5589GB-1/stack/default/ceph/minions/
*.yaml
## End of policy.cfg file
```

11 Appendix D: OS Network Configuration

Perform the network configuration during the OS installation. The three illustrations below show the configuration of one of the OSD servers and associated bond settings.

```
adminvm01.tlp - root@192.168.145.19:22 - Bitvise xterm
YaST2 - lan @ osd1

Network Settings
Global Options—Overview—Hostname/DNS—Routing

Name                                     IP Address      Device  Note
82599ES 10-Gigabit SFI/SFP+ Network Connection  NONE           eth0    enslaved in
82599ES 10-Gigabit SFI/SFP+ Network Connection  NONE           eth1    enslaved in
82599ES 10-Gigabit SFI/SFP+ Network Connection  NONE           eth2    enslaved in
82599ES 10-Gigabit SFI/SFP+ Network Connection  NONE           eth3    enslaved in
Bond Network                             192.168.145.21  bond0
Bond Network                             192.168.146.21  bond1

82599ES 10-Gigabit SFI/SFP+ Network Connection
MAC : 0c:c4:7a:d3:76:b8
BusID : 0000:04:00.0
* Device Name: eth0
* Started automatically at boot
* Bonding master: bond0

[Add][Edit][Delete]

[Help]                                     [Cancel]

F1 Help  F3 Add  F4 Edit  F5 Delete  F9 Cancel  F10 OK
```

YaST view of all network interfaces for an OSD server. Interfaces eth0 and eth1 are bonded (bond0) and make up the primary network. Eth2 and eth3 form the second bond (bond1) for the cluster network.

```
adminvm01.tlp - root@192.168.145.19:22 - Bitvise xterm
YaST2 - lan @ osd1

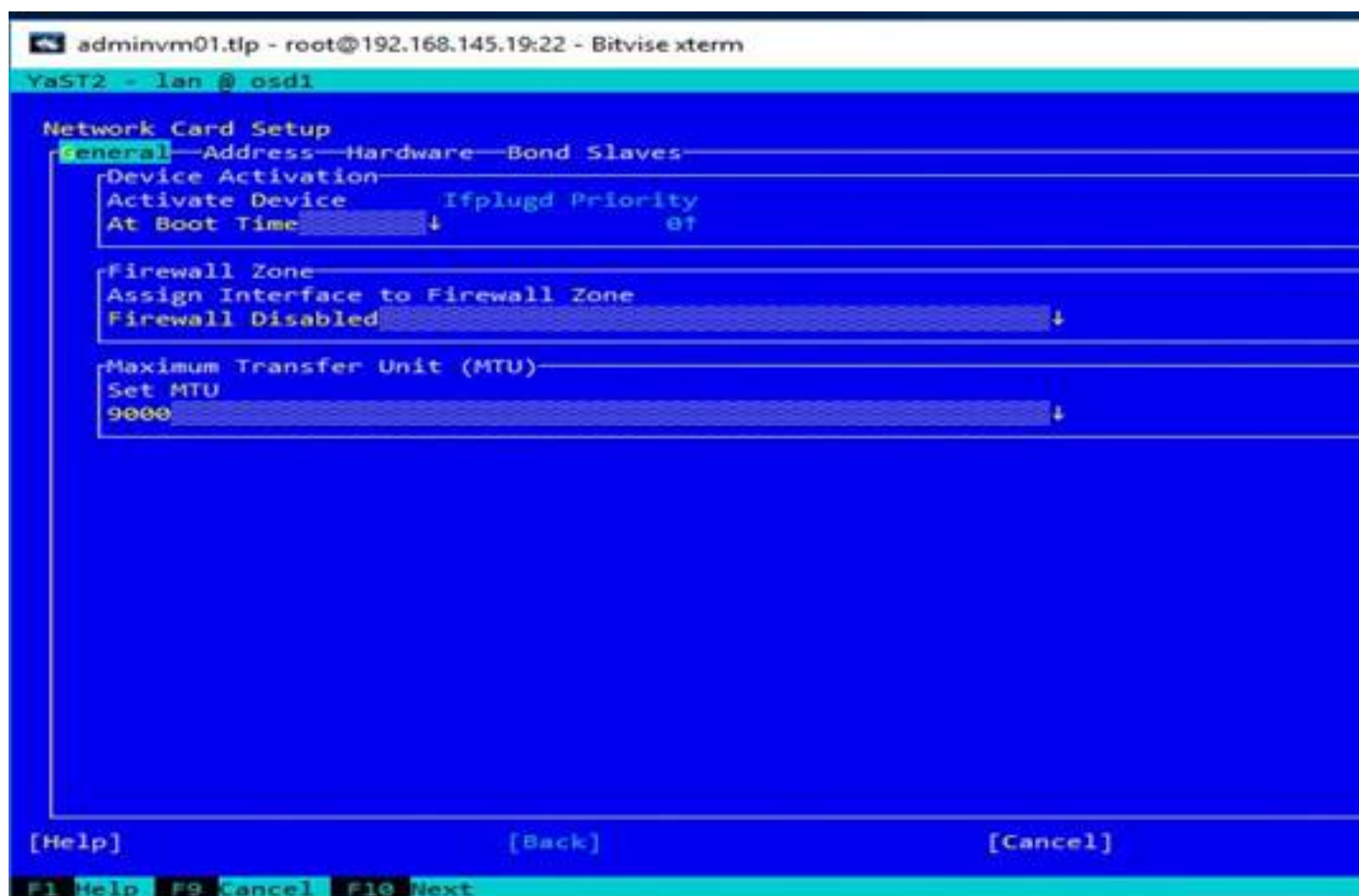
Network Card Setup
General—Address—Hardware—Bond Slaves
Bond Slaves and Order
[x] eth2 - 82599ES 10-Gigabit SFI/SFP+ Network Connection configured
[x] eth3 - 82599ES 10-Gigabit SFI/SFP+ Network Connection configured

[Up] [Down]

Bond Driver Options
mode=802.3ad miimon=100

[Help] [Back] [Cancel]
F1 Help F9 Cancel F10 Next
```

Bond drive options showing lacp.



MTU 9000 setting for the cluster network bond.

12 Appendix E: Network switches configuration files

12.1 Nexus: A-side configuration

!Command: show running-config

!Time: Wed Apr 5 15:29:39 2017

version 7.0(3)I4(5)

hostname N9K-Suse-A

vdc N9K-Suse-A id 1

limit-resource vlan minimum 16 maximum 4094 limit-resource vrf minimum 2 maximum 4096
limit-resource port-channel minimum 0 maximum 511 limit-resource u4route-mem minimum
248 maximum 248 limit-resource u6route-mem minimum 96 maximum 96 limit-resource
m4route-mem minimum 58 maximum 58 limit-resource m6route-mem minimum 8 maximum 8
feature telnet
cfs eth distribute
feature interface-vlan
feature lacp
feature vpc
no password strength-check
username admin password 5
vlan 1,145-146
vlan 145
name SuseData
vlan 146
name SuseStorage
vrf context management
ip route 0.0.0.0/0 10.128.99.1
vpc domain 100
peer-switch
role priority 10
peer-keepalive destination 10.128.99.88 source 10.128.99.87 delay restore 150 peer-gateway
auto-recovery ip arp synchronize
interface Vlan1
no shutdown
interface Vlan145
no shutdown
no ip redirects
ip address 192.168.145.145/24
interface Vlan146
no shutdown

```
mtu 9216
no ip redirects
ip address 192.168.146.145/24
interface port-channel1
switchport access vlan 145
vpc 1
interface port-channel2
switchport access vlan 145
vpc 2
interface port-channel3
switchport access vlan 145
vpc 3
interface port-channel4
switchport access vlan 145
vpc 4
interface port-channel5
switchport access vlan 145
vpc 5
interface port-channel6
switchport access vlan 145
vpc 6
interface port-channel7
switchport access vlan 145
vpc 7
interface port-channel8
switchport access vlan 145
vpc 8
interface port-channel9
switchport access vlan 145
vpc 9
```

```
interface port-channel10
switchport access vlan 145
vpc 10
interface port-channel11
switchport access vlan 145
vpc 11
interface port-channel12
switchport access vlan 145
vpc 12
interface port-channel25
switchport access vlan 146
mtu 9216
vpc 25
interface port-channel26
switchport access vlan 146
mtu 9216
vpc 26
interface port-channel27
switchport access vlan 146
mtu 9216
vpc 27
interface port-channel28
switchport access vlan 146
mtu 9216
vpc 28
interface port-channel48
switchport mode trunk
mtu 9216
vpc 48
interface port-channel53
```

```
description VPC peer
switchport mode trunk
switchport trunk allowed vlan 1,145-146
spanning-tree port type network
vpc peer-link
interface Ethernet1/1
switchport access vlan 145
channel-group 1 mode active
interface Ethernet1/2
switchport access vlan 145
channel-group 2 mode active
interface Ethernet1/3
switchport access vlan 145
channel-group 3 mode active
interface Ethernet1/4
switchport access vlan 145
channel-group 4 mode active
interface Ethernet1/5
switchport access vlan 145
channel-group 5 mode active
interface Ethernet1/6
switchport access vlan 145
channel-group 6 mode active
interface Ethernet1/7
switchport access vlan 145
channel-group 7 mode active
interface Ethernet1/8
switchport access vlan 145
channel-group 8 mode active
interface Ethernet1/9
```



```
switchport access vlan 145
channel-group 9 mode active
interface Ethernet1/10
switchport access vlan 145
channel-group 10 mode active
interface Ethernet1/11
switchport access vlan 145
channel-group 11 mode active
interface Ethernet1/12
switchport access vlan 145
channel-group 12 mode active
interface Ethernet1/13
interface Ethernet1/14
interface Ethernet1/15
interface Ethernet1/16
interface Ethernet1/17
interface Ethernet1/18
interface Ethernet1/19
interface Ethernet1/20
interface Ethernet1/21
interface Ethernet1/22
interface Ethernet1/23
interface Ethernet1/24
interface Ethernet1/25
description SUSE OSD
switchport access vlan 146
mtu 9216
channel-group 25 mode active
interface Ethernet1/26
description SUSE OSD
```

```
switchport access vlan 146
mtu 9216
channel-group 26 mode active
interface Ethernet1/27
description SUSE OSD
switchport access vlan 146
mtu 9216
channel-group 27 mode active
interface Ethernet1/28
description SUSE OSD
switchport access vlan 146
mtu 9216
channel-group 28 mode active
interface Ethernet1/29
switchport access vlan 146
interface Ethernet1/30
interface Ethernet1/31
interface Ethernet1/32
interface Ethernet1/33
interface Ethernet1/34
interface Ethernet1/35
interface Ethernet1/36
interface Ethernet1/37
interface Ethernet1/38
interface Ethernet1/39
interface Ethernet1/40
interface Ethernet1/41
interface Ethernet1/42
interface Ethernet1/43
interface Ethernet1/44
```

```
interface Ethernet1/45
interface Ethernet1/46
interface Ethernet1/47
interface Ethernet1/48
switchport mode trunk
mtu 9216
channel-group 48 mode active
interface Ethernet1/49
interface Ethernet1/50
interface Ethernet1/51
interface Ethernet1/52
interface Ethernet1/53
description VPC Peer
switchport mode trunk
switchport trunk allowed vlan 1,145-146
channel-group 53 mode active
interface Ethernet1/54
description VPC Peer
switchport mode trunk
switchport trunk allowed vlan 1,145-146
channel-group 53 mode active
interface mgmt0
vrf member management
ip address 10.128.99.87/24
line console
line vty
session-limit 16
email
smtp-host 10.128.30.12
smtp-port 25
```

reply-to buffy@imciscoexp.com

from SuseNexus@imciscoexp.com

vrf management

boot nxos bootflash:/nxos.7.0.3.I4.5.bin

show port-channel summary

Flags: D - Down P - Up in port-channel (members) I - Individual H - Hot-standby (LACP only) s
- Suspended r - Module-removed S - Switched R - Routed U - Up (port-channel) p - Up in delay-
lACP mode (member) M - Not in use. Min-links not met

Group Port- Type Protocol Member Ports
Channel

1 Po1(SU) Eth LACP Eth1/1(P)
2 Po2(SU) Eth LACP Eth1/2(P)
3 Po3(SU) Eth LACP Eth1/3(P)
4 Po4(SU) Eth LACP Eth1/4(P)
5 Po5(SU) Eth LACP Eth1/5(P)
6 Po6(SU) Eth LACP Eth1/6(P)
7 Po7(SU) Eth LACP Eth1/7(P)
8 Po8(SU) Eth LACP Eth1/8(P)
9 Po9(SU) Eth LACP Eth1/9(P)
10 Po10(SU) Eth LACP Eth1/10(P)
11 Po11(SU) Eth LACP Eth1/11(P)
12 Po12(SU) Eth LACP Eth1/12(P)
25 Po25(SU) Eth LACP Eth1/25(P)
26 Po26(SU) Eth LACP Eth1/26(P)
27 Po27(SU) Eth LACP Eth1/27(P)
28 Po28(SU) Eth LACP Eth1/28(P)
48 Po48(SU) Eth LACP Eth1/48(P)

53 Po53(SU) Eth LACP Eth1/53(P) Eth1/54(P)

13 Resources:

14

[https://www.suse.com/documentation/ses-4/book_storage_admin/data/
book_storage_admin.html](https://www.suse.com/documentation/ses-4/book_storage_admin/data/book_storage_admin.html) ↗