SUSE® Enterprise Storage 6 on Huawei Taishan Implementation Guide

SUSE® Enterprise Storage 6 on Huawei Taishan Implementation Guide

Kai Liu

Publication Date: Mar 15 2020

SUSE LLC
10 Canal Park Drive
Suite 200
Cambridge MA 02141
USA
https://documentation.suse.com

Contents

1 Introduction 1

2 Target Audience 2

3 Hardware & Software 3

4 Business Problem and Business Value 4

4.1 Business Problem 4

4.2 Business Value 4

4.3 SUSE Enterprise Storage 4

4.4 Huawei Taishan 5

5 Architectural Overview 6

5.1 Solution Architecture 6

5.2 Networking Architecture 8

5.3 Network/IP Address Scheme 9

6 Component Model 11

7 Deployment 12

7.1 Network Considerations 12

7.2 Hardware Considerations 12

7.3 Operating System Considerations 13

7.4 SUSE Enterprise Storage Installation & Configuration 15

Software Deployment Configuration (Deepsea and Salt) 15 • Prepare All Nodes 15 • Prepare OSD Disks on OSD Nodes 16 • Install and Configure Deepsea on Admin Node 17 • Deploy Using Deepsea 18

7.5 Post-deployment Quick Tests 21

8 Tips 22

8.1 Use a different NTP server 22

8.2 Copy files to all cluster nodes 22

8.3 Important files 22

8.4 How to completely uninstall the cluster for reinstall 23

8.5 Erase OSD drives 23

8.6 How to get salt pillar information 23

8.7 SES built-in network benchmark 24

8.8 Ceph built-in OSD benchmark 24

8.9 Ceph built-in pool scope benchmark 24

8.10 Interface bonding 24

8.11 Offline setup 24

8.12 Change node roles 25

8.13 More tips 25

9 Conclusion 26

10 References and Resources 27

A Bill of Materials 28

B Network Switch Configuration 30

C policy.cfg example 31

D drive_groups.yml example 32

1 Introduction

The objective of this guide is to present a step-by-step guide on how to implement SUSE Enterprise Storage 6 on the Huawei Taishan platform. It is suggested that the document be read in its entirety, along with the supplemental appendix information before attempting the process.

The deployment presented in this guide aligns with architectural best practices and will support the implementation of all currently supported protocols as identified in the SUSE Enterprise Storage documentation.

Upon completion of the steps in this document, a working SUSE Enterprise Storage 6 cluster will be operational as described in the SUSE Enterprise Storage Deployment and Administration Guide (https://www.suse.com/documentation/ses-5/book_storage_admin/data/book_storage_admin.html).

2 Target Audience

This reference architecture is targeted at 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.

3 Hardware & Software

The recommended architecture for SUSE Enterprise Storage on Huawei Taishan leverages two models of Huawei servers. The role and functionality of each type of system within the SUSE Enterprise Storage environment will be explained in more detail in the *Chapter 5, Architectural Overview* section.

STORAGE NODES:

- Huawei 5280 (https://www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=147070)

ADMIN, MONITOR, AND PROTOCOL GATEWAYS:

- Huawei 2280 (https://www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=146997)

SWITCHES:

- Huawei CE6851-48S6Q-HI 10Gb

SOFTWARE:

- SUSE Enterprise Storage 6
- SUSE Linux Enterprise Server 15 SP1

4 Business Problem and Business Value

4.1 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 took place driven by the need to access and leverage new data sources (e.g. external sources such as social media) and the ability to "manage" new data types (e.g. 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 (e.g. RAID). These solutions struggle when scaling from Terabytes to Petabytes at reasonable cost and performance levels.

4.2 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 files, blocks, or objects depending on the 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.

4.3 SUSE Enterprise Storage

SUSE Enterprise Storage delivers a highly scalable, resilient, self-healing storage system 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.

4.4 Huawei Taishan

Huawei Taishan servers provide a cost effective and scalable platform for the deployment of SUSE Enterprise Storage. These platforms unlocks the full potential of the Kunpeng CPU, raising the bar of SPECint benchmark by 25%, with up to 128 cores, 32 DDR4 DIMMs, PCIe 4.0, CCIX, and 100 GE LOM.

Featuring models tailored for computing, storage, or balanced needs, Taishan is perfect for demanding workloads such as big data analytics, database acceleration, high-performance computing, and cloud services. Taishan servers empower data centers with the ultimate efficiency.

5 Architectural Overview

This architecture overview section complements the SUSE Enterprise Storage Technical Overview (https://www.suse.com/media/white-paper/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).

5.1 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 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 over the public network, resulting in horizontally scaling performance. Non-native protocols, such as iSCSI, S3, and NFS require the use of gateways. While these gateways may be thought of as a limiting factor, the iSCSI and S3 gateways can scale horizontally using load balancing techniques.

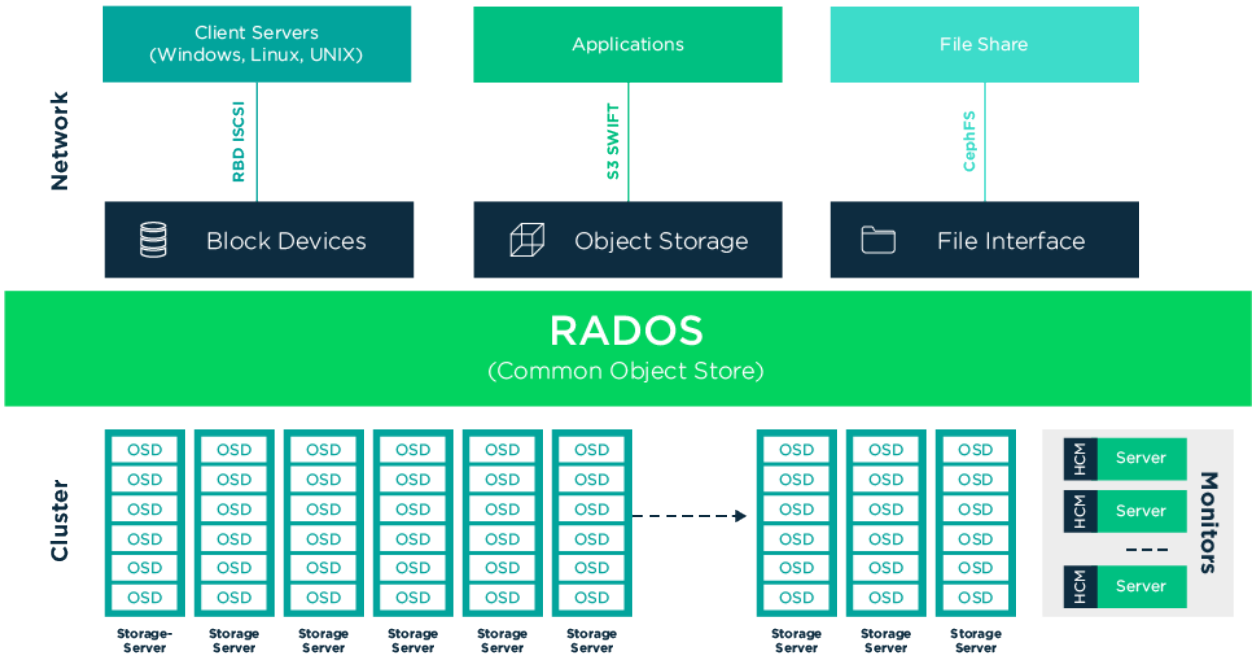

FIGURE 5.1: SES ARCHITECTURE

In addition to the required network infrastructure, the minimum SUSE Enterprise Storage cluster comprises of a minimum of one administration server (physical or virtual), four object storage device nodes (OSDs), and three monitor nodes (MONs).

Please refer to the SES 6 Deployment Guide (https://documentation.suse.com/en-us/ses/6/single-html/ses-deployment/#storage-bp-hwreq) for more details on the hardware requirement.

SPECIFIC TO THIS IMPLEMENTATION:

- One system is deployed as the administrative server. It is the Salt Master and hosts the SUSE Enterprise Storage Administration Interface, dashboard, which is the central management system which supports the cluster.
- Three systems 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.
- It is strongly recommended to deploy monitors and other services on dedicated nodes. However it is also possible to deploy the monitors on the OSD nodes if there are enough hardware resources. This is the case in this specific reference setup.

- The RADOS gateway provides S3 and Swift based access methods to the cluster. These nodes are generally situated behind a load balancer infrastructure to provide redundancy and scalability. It is important to note that the load generated by the RADOS gateway can consume a significant amount of compute and memory resources making the minimum recommended configuration contain 6-8 CPU cores and 32GB of RAM.

- SUSE Enterprise Storage requires a minimum of four systems as storage nodes. The storage nodes contain individual storage devices that are each assigned an Object Storage Daemon (OSD). The OSD daemon assigned to the device 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.

5.2 Networking Architecture

A software-defined solution is only as reliable 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 internal and client-facing public network traffic. This isolates Ceph OSD daemon replication activities from Ceph clients. This may be achieved through separate physical networks or through use of VLANs.

- Redundancy and capacity in the form of bonded network interfaces connected to switches.

Figure 5.2, "Ceph Network Architecture" shows the logical layout of the Ceph cluster implementation.

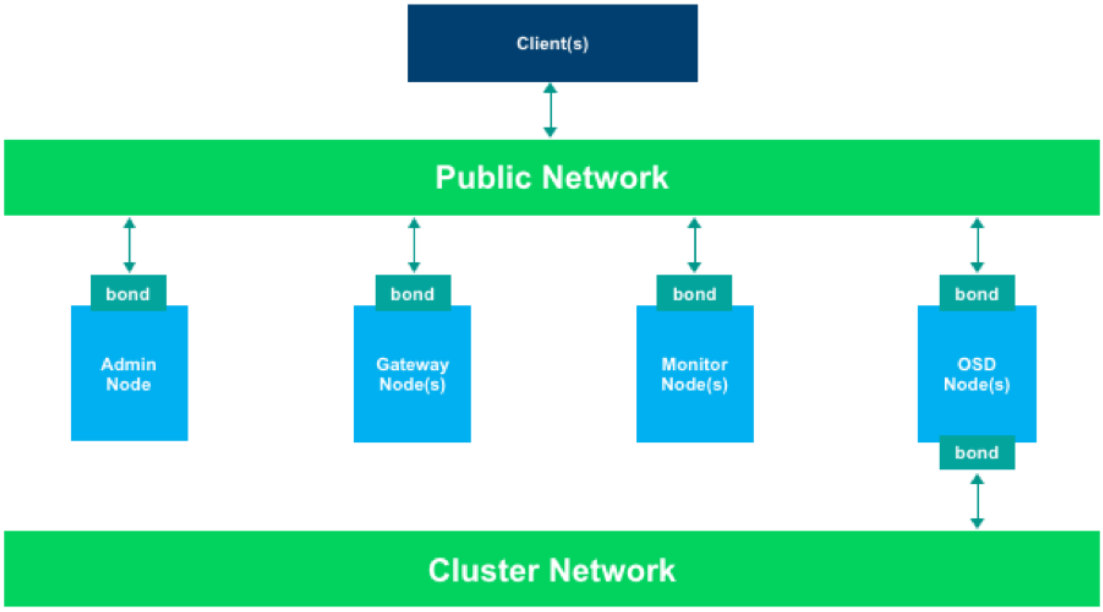

FIGURE 5.2: CEPH NETWORK ARCHITECTURE

5.3 Network/IP Address Scheme

Specific to this implementation, the following naming and addressing scheme were utilized.

TABLE 5.1: NODE ROLES AND NETWORK ADDRESSES

Role	Hostname	Public Network	Cluster Network
Admin	admin.example.com	10.1.1.3	N/A
Monitor	ceph1.example.com	10.1.1.4	N/A
Monitor	ceph2.example.com	10.1.1.5	N/A
Monitor	ceph3.example.com	10.1.1.6	N/A
OSD Node	ceph1.example.com	10.1.1.4	10.2.1.4
OSD Node	ceph2.example.com	10.1.1.5	10.2.1.5
OSD Node	ceph3.example.com	10.1.1.6	10.2.1.6

Role	Hostname	Public Network	Cluster Network
OSD Node	ceph4.example.com	10.1.1.7	10.2.1.7

6 Component Model

The preceding sections provided information on the both the overall Huawei 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 3 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 SLES - Allows enterprise customers to optimize the management of SUSE Linux Enterprise (and product 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 product 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.

7 Deployment

This deployment section should be seen as a supplement online documentation (https://www.suse.com/documentation/). Specifically, the SUSE Enterprise Storage 5 Deployment Guide (https://www.suse.com/documentation/suse-enterprise-storage-5/book_storage_deployment/data/book_storage_deployment.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). It is assumed that a Subscription Management Tool server exists within the environment. If not, please follow the information in Subscription Management Tool (SMT) for SLES (https://www.suse.com/documentation/sles-12/book_smt/data/book_smt.html) to make one available. The emphasis is on specific design and configuration choices.

In this document we use `example.com` as the domain name for the nodes, replace it with your real domain name in your own installation.

7.1 Network Considerations

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

- Ensure that all network switches are updated with consistent firmware versions.
- Network IP addressing and IP ranges need proper planning. In optimal environments, a dedicated storage subnet should be used for all SUSE Enterprise Storage nodes on the primary network, with a separate, dedicated 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.

7.2 Hardware Considerations

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

- Configure the system to performance mode if you prefer performance over power efficiency. To change that, reboot the system, press `Del` when prompted during system initializing to boot into the BIOS setup menu. Then select *Advanced* › *Performance Config* › *Power Policy*, select *Performance* for best performance or *Efficiency* for power efficiency.

Note

The Huawei Taishan server platform currently does not support OS controlled processor frequency scaling. The only way to change the system power policy is to reboot and change it in the system BIOS.

- In case you can't see the SUSE Installer screen after booting up the SUSE installation medium, check the BIOS option *Advanced* › *MISC Config* › *Support SPCR* and set it to *Disabled*.
- A RAID-1 volume consists of two 600GB SAS hard drive is enough for the OS disk.
- If hard drives are connected to hardware RAID controller(s) with hardware write cache, configure each of them as individual RAID-0 volume and make sure hardware caching is enabled.
- Try to balance the drives across controllers, ports, and enclosures. Avoid making one part of the I/O subsystem busy while leaving other parts idle.
- If SAS/SATA SSDs are installed, make sure to attach then to a dedicated HBA or RAID controller rather than to the controller that already has many HDDs attached.

7.3 Operating System Considerations

The following considerations for the Operating System should be attended to:

- The underlying OS for SES 6 is SUSE Linux Enterprise Server 15 SP1. Other OS versions are not supported. During installation, make sure below addon modules are selected.
  - Base system module
  - Server Applications module
  - SUSE Enterprise Storage 6
- During installation, don't select any GUI components such as X-Window system, GNOME or KDE, as they are not needed to run the storage service.

- It is highly recommended to register the systems to an update server to install the latest updates available, helping to ensure the best experience possible. The systems could be registered directly to SUSE Customer Center if it is a small cluster, or could be registered to a local SMT or RMT server when the cluster is large. Installing updates from a local SMT/RMT server will dramatically reduce the time required for updates to be downloaded to all nodes.

Tip

Refer to Repository Mirroring Tool Guide (https://documentation.suse.com/en-us/sles/15-SP1/single-html/SLES-rmt/) for how to setup a RMT server.

- Ensure that the operating system is installed on the correct device. Especially on OSD nodes, the installer may not choose the right one from many available drives.

- Hostnames of all nodes should be properly configured. Full hostname (i.e. with domain name) should always be assigned for each node or else the deployment may fail. Make sure `hostname -s`, `hostname -f` and `hostname -i` commands return proper results for short hostname (without dots), full hostname and IP addresses. Each node must also be able to resolve hostname of all nodes, including its own name.

  - For a rather small cluster, hosts files can be used for name resolution. Also see *Section 8.2, "Copy files to all cluster nodes"* for how to conveniently keep the hosts on all nodes in sync.

  - Having a DNS server is recommended for a larger cluster. See the SUSE Linux Enterprise Server Administration Guide (https://documentation.suse.com/sles/15-SP1/single-html/SLES-admin/#cha-dns) for how to setup a DNS server.

- Do ensure that NTP is configured to point to a valid, physical NTP server. This is critical for SUSE Enterprise Storage to function properly, and failure to do so can result in an unhealthy or non-functional cluster. And keep in mind that the NTP service is not designed to be run on an virtualized environment, so make sure the NTP server been used is an physical machine or it may cause strange clock drifting problem.

7.4 SUSE Enterprise Storage Installation & Configuration

7.4.1 Software 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:

- 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 host 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.

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 layout, node role assignment and drive assignment. Deepsea collects each set of tasks into a goal or stage.

The following steps, performed in order, were used for this reference implementation. All commands were run by root user.

7.4.2 Prepare All Nodes

1. Install salt master on the Admin node:

```
zypper in salt-master
```

2. Start the salt-master service and enable start on boot:

```
systemctl enable --now salt-master.service
```

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

```
zypper in salt-minion
```

4. Configure all minions to connect to the Salt master:
   Create a new file `/etc/salt/minion.d/master.conf` with the following content:

   ```
   master: admin.example.com
   ```

5. Restart the salt-minion service and enable it:

   ```
   systemctl restart salt-minion.service
   systemctl enable salt-minion.service
   ```

6. List Salt fingerprints on all the minions:

   ```
   salt-call --local key.finger
   ```

7. List all incoming minion fingerprints on the Salt master, verify them against the fingerprints on each minions to make sure they all match. If they do, accept all Salt keys on the Salt master:

   ```
   salt-key -F
   salt-key --accept-all
   salt-key --list-all
   ```

8. Verify if Salt works properly by "ping" each minions from the Salt master. They should all return True on success:

   ```
   salt '*' test.ping
   ```

9. Now check and make sure the time on all nodes are the same. In later stage Deepsea will setup all nodes to synchronize time from the admin node, but before that is done, strange errors may occur if time on each node are largely out of sync. So it's better to set all nodes to the same time manually first. For example, run below command on your admin node:

   ```
   salt '*' cmd.run 'date -s "2020-03-19 17:30:00"'
   ```

   Use your actual time in the same format when running the command. It doesn't have to be super accurate, as later all nodes will be synchronized by the chrony time service.

7.4.3 Prepare OSD Disks on OSD Nodes

If the OSD nodes were used in a prior installation, or the disks are used by other applications before, zap ALL the OSD disks first.

Important

This must be done on all the OSD disks that were used before, or else the deployment may fail when activating OSDs.

Warning

Below commands should not be copied and executed on your installation blindly. The device names used below are just examples, you need to change them to match only the OSD disks in your own installation. Failed to use the correct device name may erase your OS disk or other disks that may hold valuable data.

1. Wipe the beginning of each partition:

```
for partition in /dev/sdX[0-9]*
do
  dd if=/dev/zero of=$partition bs=4096 count=1 oflag=direct
done
```

2. Wipe the beginning of the drive:

```
dd if=/dev/zero of=/dev/sdX bs=512 count=34 oflag=direct
```

3. Wipe the end of the drive:

```
dd if=/dev/zero of=/dev/sdX bs=512 count=33 \
  seek=$((`blockdev --getsz /dev/sdX` - 33)) oflag=direct
```

7.4.4 Install and Configure Deepsea on Admin Node

1. Install deepsea package on Admin node:

```
# zypper in deepsea
```

2. Check */srv/pillar/ceph/master_minion.sls* for correctness.

3. Check */srv/pillar/ceph/deepsea_minions.sls* file, make sure the deepsea_minions option targets the correct nodes. In the usual case, it can simply be put like below to match all Salt minions in the cluster:

```
deepsea_minions: '*'
```

4. Create */srv/pillar/ceph/stack/ceph/cluster.yml* with below options:

```
cluster_network: <net/mask of cluster network>
public_network: <net/mask of public network>
time_server: <Address of NTP server, if this line is omitted admin node will be
 used>
```

7.4.5 Deploy Using Deepsea

At this point Deepsea commands can be run on the admin node to deploy the cluster.

Tip

Each command can be run either as:

```
salt-run state.orch ceph.stage.<stage name>
```

Or:

```
deepsea stage run ceph.stage.<stage name>
```

The latter form is preferred as it outputs real time progress.

7.4.5.1 Stage 0: Prepare

During this stage, all required updates are applied and your system may be rebooted.

```
deepsea stage run ceph.stage.0
```

Important

If the Salt master reboots during Stage 0, you need to run Stage 0 again after it boots up.

Optionally, create the */var/lib/ceph* btrfs subvolume:

```
salt-run state.orch ceph.migrate.subvolume
```

7.4.5.2 Stage 1: Discovery

During this stage, all hardware in your cluster is detected and necessary information are collected for the Ceph configuration.

```
deepsea stage run ceph.stage.1
```

Note

Configure cluster and public network in */srv/pillar/ceph/stack/ceph/cluster.yml* if not yet done as described in *Create cluster.yml*.

Now a */srv/pillar/ceph/proposals/policy.cfg* file needs to be created to instruct Deepsea on the location and configuration files to use for the different components that make up the Ceph cluster (Salt master, admin, monitor, OSD and other roles).

To do so, copy the example file to the right location then edit it to match your installation:

```
cp /usr/share/doc/packages/deepsea/examples/policy.cfg-rolebased /srv/pillar/ceph/
proposals/policy.cfg
```

Tip

See *Appendix C, policy.cfg example* for the one used when installing the cluster described in this document.

7.4.5.3 Stage 2: Configure

During this stage necessary configuration data are prepared in particular format.

```
deepsea stage run ceph.stage.2
```

Tip

Use below command to check the attributes of each node:

```
salt '*' pillar.items
```

7.4.5.4 Define drive groups

DriveGroups information are defined in the file */srv/salt/ceph/configuration/files/drive_groups.yml*. It specifies what drives should be used for data device, DB device, or WAL device, and other parameters for setting up the OSDs.

1. First take a look of all the disks on all OSD nodes:

   ```
   salt-run disks.details
   ```

   It lists the vendor, model, size and type of the disks. Those information can be used to match a group of drives and assign them to different uses.

2. Now define drive groups in the drive_groups.yml file.

   

   ## Tip

   See *Appendix D, drive_groups.yml example* for the drive group definition used in this example cluster. For complete information refers to the Deployment Guide (https://documentation.suse.com/en-us/ses/6/single-html/ses-deployment/#ds-drive-groups)

3. After finished editing *drive_groups.yml*, run below commands to see the result definition. Exam it carefully and make sure it meets your expectation before moving on to next step.

   ```
   salt-run disks.list
   salt-run disks.report
   ```

7.4.5.5 Stage 3: Deploy

A basic Ceph cluster with mandatory Ceph services is created in this stage.

```
deepsea stage run ceph.stage.3
```

Note

It may take quite some time for above command to finish if your cluster is large, your have a lot of disks, or your Internet bandwidth is limited while you didn't register the nodes to local SMT server.

After the above command is finished successfully, check whether the cluster is up by running:

```
ceph -s
```

7.4.5.6 Stage 4: Services

Additional features of Ceph like iSCSI, Object Gateway and CephFS can be installed in this stage. Each is optional and up to your situation.

```
deepsea stage run ceph.stage.4
```

7.5 Post-deployment Quick Tests

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

```
ceph status
ceph osd pool create test 1024
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 tell mon.* injectargs --mon-allow-pool-delete=true
ceph osd pool delete test test --yes-i-really-really-mean-it
ceph tell mon.* injectargs --mon-allow-pool-delete=false
```

8 Tips

8.1 Use a different NTP server

The default time server is the admin node. To change it, add

```
time_server: <server address>
```

in cluster.yml

8.2 Copy files to all cluster nodes

`salt-cp` command can be used to copy files from the salt master node to minion nodes. This can be very convenient, for example, to keep /etc/hosts file in sync on all nodes.

```
salt-cp '*' /etc/hosts /etc/hosts
```

8.3 Important files

/etc/salt/minion

Salt minion configuration file

/etc/salt/minion_id

Salt minion name. Useful if changed host name and need to change minion name accordingly.

/srv/pillar/ceph/deepsea_minions.sls

Deepsea minion targets

/srv/pillar/ceph/stack/ceph/cluster.yml

Deepsea cluster configuration for the cluster "ceph" (the default cluster name). Need to run stage 2 configure after modification. Or refresh pillar and check:

```
# salt <target> saltutil.pillar_refresh
# salt <target> pillar.items
```

CLUSTER CONFIGURATION FILES

/srv/pillar/ceph/stack/global.yml

Affects all minions in the Salt cluster.

/srv/pillar/ceph/stack/ceph/cluster.yml

Affects all minions in the cluster named "ceph".

/srv/pillar/ceph/stack/ceph/roles/role.yml

Affects all minions that are assigned the specific role in the ceph cluster.

/srv/pillar/ceph/stack/cephminions/<minion ID>/yml

Affects the individual minion.

8.4 How to completely uninstall the cluster for reinstall

In case you did something wrong and would like to start over without re-installing the whole OS.

```
# salt-run disengage.safety
# salt-run state.orch ceph.purge
```

8.5 Erase OSD drives

Sometimes it's necessary to completely erase the information left from other programs or previous installtions.

Besides of removing the partition table, the first few hundred MBs must also be cleared because that's the Bluestore small partition that has a filesystem, which holds some key information of that Bluestore disk. If it is not erased, the filesystem will be intact and it still holds the old information so reinstall will fail.

```
salt '<osd target>' cmd.run 'for d in {<start letter>..<end letter>}; do ceph-disk zap /
dev/sd$d; dd if=/dev/zero of=/dev/sd$d bs=1M count=500; done'
```

8.6 How to get salt pillar information

```
# salt '*' pillar.items
```

This can only give information after running stage 1 AKA the discovery stage

8.7 SES built-in network benchmark

https://www.suse.com/documentation/suse-enterprise-storage-5/singlehtml/book_storage_admin/book_storage_admin.html#storage.bp.performance.net_issues

```
# salt-run net.iperf cluster=ceph output=full
```

8.8 Ceph built-in OSD benchmark

https://www.suse.com/documentation/suse-enterprise-storage-5/singlehtml/book_storage_admin/book_storage_admin.html#storage.bp.performance.slowosd

```
# ceph tell osd.<id> bench
```

8.9 Ceph built-in pool scope benchmark

```
# rados -p <pool name> bench 60 write
```

8.10 Interface bonding

Use following parameters for bonding in 802.3ad mode (need switch support). mode = 802.3ad miimon = 100 lacp_rate = fast xmit_hash_policy = layer3 + 4

Recommended Size for the BlueStore's WAL and DB Device https://www.suse.com/documentation/suse-enterprise-storage-5/singlehtml/book_storage_deployment/book_storage_deployment.html#about.bluestore

https://www.suse.com/documentation/suse-enterprise-storage-5/singlehtml/book_storage_deployment/book_storage_deployment.html#rec.waldb.size

8.11 Offline setup

Setup a SMT or RMT server, and mirror below repositories from SCC.

- SLE-Product-SLES15-SP1-Pool
- SLE-Product-SLES15-SP1-Updates
- SLE-Module-Server-Applications15-SP1-Pool
- SLE-Module-Server-Applications15-SP1-Updates
- SLE-Module-Basesystem15-SP1-Pool
- SLE-Module-Basesystem15-SP1-Updates
- SUSE-Enterprise-Storage-6-Pool
- SUSE-Enterprise-Storage-6-Updates

Then point all nodes to the SMT/RMT server.

8.12 Change node roles

After change of node roles by editing policy.cfg, need to run Stage 2 Configure to refresh configuration files.

```
# deepsea stage run ceph.stage.2
```

8.13 More tips

Check the SES 6 Administration Guide (https://documentation.suse.com/ses/6/single-html/ses-admin/#part-troubleshooting) for more hints & tips, FAQ, and troubleshooting techniques.

9 Conclusion

The Huawei Taishan series represents a strong capacity-oriented platform. When combined with the access flexibility and reliability of SUSE Enterprise Storage and the industry leading support from Huawei, any business can feel confident in the ability to address the exponential growth in storage they are currently faced with.

10 References and Resources

SUSE Enterprise Storage Technical Overview

https://www.suse.com/media/white-paper/suse_enterprise_storage_technical_overview_wp.pdf

SUSE Enterprise Storage Tech Specs

https://www.suse.com/products/suse-enterprise-storage/#tech-specs

SUSE Enterprise Storage 6 - Release Notes

https://www.suse.com/releasenotes/x86_64/SUSE-Enterprise-Storage/6/

SUSE Enterprise Storage 6 - Deployment Guide

https://documentation.suse.com/ses/6/single-html/ses-deployment/#book-storage-deployment

SUSE Enterprise Storage 6 - Administration Guide

https://documentation.suse.com/ses/6/single-html/ses-admin/#book-storage-admin

SUSE Linux Enterprise Server 15 SP1 - Deployment Guide

https://documentation.suse.com/sles/15-SP1/single-html/SLES-deployment/#book-sle-deployment

SUSE Linux Enterprise Server 15 SP1 - Administration Guide

https://documentation.suse.com/sles/15-SP1/single-html/SLES-admin/#book-sle-admin

SUSE Linux Enterprise Server 15 SP1 - Storage Administration Guide

https://documentation.suse.com/sles/15-SP1/single-html/SLES-storage/#book-storage

SUSE Linux Enterprise Server 15 SP1 - Repository Mirroring Tool Guide

https://documentation.suse.com/sles/15-SP1/single-html/SLES-rmt/#book-rmt

A Bill of Materials

Role	Qty	Component	Notes
Admin Node	1	A VM on a Huawei x86 machine	The node consists of: • 8x vCPU cores • 64GB RAM • 100GB virtual disk for OS • 2x virtual NIC port to public and cluster network
OSD nodes. MON, MGR shared the OSD hosts	4	Huawei Taishan 5280	Each node consists of: • 2x Kunpeng 920 • 256GB • 1x Kunpeng 920 Integrated SAS Controller • 2x 300GB SAS HDD for OS • 34x 4TB 7.2k SATA HDD for OSD • 2x Huawei 1TB NVMe SSD for db and journal • 1x Kunpeng 920 Embedded Network Controller - 1GbE Quad Port • 1x Dual Port Intel 82599 10Gb Ethernet adapter
Software	1	SUSE Enterprise Storage 6 Subscription Base configuration	Allows for 4 storage nodes and 6 infrastructure nodes. Expansion subscriptions need to be purchased if more nodes need to be added to the cluster.

Role	Qty	Component	Notes
Switches	2	Huawei CE6851-48S6Q-HI	32 Ports of 100GbE

B Network Switch Configuration

Below are the commands used to configure the Huawei CE6851-48S6Q-HI switch for 802.3ad link aggregation.

```
port link-type trunk
mode lacp-static
load-balance src-dst-ip
local-preference disable
```

In the OS bonding configuration, make sure below options are set:

```
BONDING_MODULE_OPTS='mode=802.3ad miimon=100 lacp_rate=fast xmit_hash_policy=layer3+4'
```

C policy.cfg example

```
## Cluster Assignment
cluster-ceph/cluster/*.sls

## Roles
# ADMIN
role-master/cluster/admin*.sls
role-admin/cluster/admin*.sls

# Monitoring
role-prometheus/cluster/admin*.sls
role-grafana/cluster/admin*.sls

# MON
role-mon/cluster/ceph[123]*.sls

# MGR (mgrs are usually colocated with mons)
role-mgr/cluster/ceph[123]*.sls

# MDS
role-mds/cluster/ceph2*.sls

# IGW
role-igw/cluster/ceph3*.sls

# RGW
role-rgw/cluster/ceph4*.sls

# NFS
# role-ganesha/cluster/ganesha*.sls

# COMMON
config/stack/default/global.yml
config/stack/default/ceph/cluster.yml

# Storage
role-storage/cluster/ceph[1234]*.sls
```

D drive_groups.yml example

```
default:
  target: 'I@roles:storage'
  data_devices:
    # Use all hard disks as data device
    rotational: 1
  db_devices:
    # Use solid state drives as db device
    rotational: 0
```