SUSE® Enterprise Storage on Lenovo ThinkSystem

Publication Date: 2019-06-28

Contents

1 Introduction

The objective of this guide is to present a step-by-step guide on how to implement SUSE Enterprise Storage (v5.5) on the Lenovo ThinkSystem 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 (v5.5) cluster will be operational as described in the SUSE Enterprise Storage Deployment Guide. (https://www.suse.com/documentation/suse-enterprise-storage-5/book_storage_deployment/data/book_storage_deployment.html)

2 Target Audience

This reference guide is targeted at administrators who deploy software defined storage solutions within their data centers and make that storage available to end users. 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 Business Value

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.

Lenovo ThinkSystem

Lenovo ThinkSystem provide a cost effective and scaleable platform for the deployment of SUSE Enterprise Storage.

Lenovo delivers cost-effective, reliable, and scalable solutions by combining industry-leading technology and the world's best software-defined offerings with Lenovo ThinkShield, XClarity, and TruScale Infrastructure Services to manage the life cycle of your data center needs. ThinkSystem SR650 provides support for data analytics, hybrid cloud, hyperconverged infrastructure, video surveillance, high performance computing and much more.

4 Hardware & Software

The recommended architecture for SUSE Enterprise Storage on Lenovo ThinkSystem leverages two models of Lenovo servers. The role and functionality of each type of system within the SUSE Enterprise Storage environment will be explained in more detail in the architectural overview section.

STORAGE NODES:

- Lenovo SR650 https://www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=146019

ADMIN, MONITOR, AND PROTOCOL GATEWAYS:

- Lenovo SR630 https://www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=146342

SWITCHES:

- Lenovo NE3200 100Gb

SOFTWARE:

- SUSE Enterprise Storage (v5.5)
- SUSE Linux Enterprise Server 12 SP3

TIP

Please note that limited use subscriptions are provided with SUSE Enterprise Storage as part of the subscription entitlement

5 Requirements

Enterprise storage systems require reliability, manageability, and serviceability. The legacy storage players have established a high threshold for each of these areas and now expect the software defined storage solutions to offer the same. Focusing on these areas helps SUSE make open source technology enterprise consumable. When combined with highly reliable and manageable hardware from Lenovo, the result is a solution that meets the customer's expectation.

5.1 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.

6 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) document available online which presents the concepts behind software defined storage and Ceph as well as a quick start guide (non-platform specific).

6.1 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.

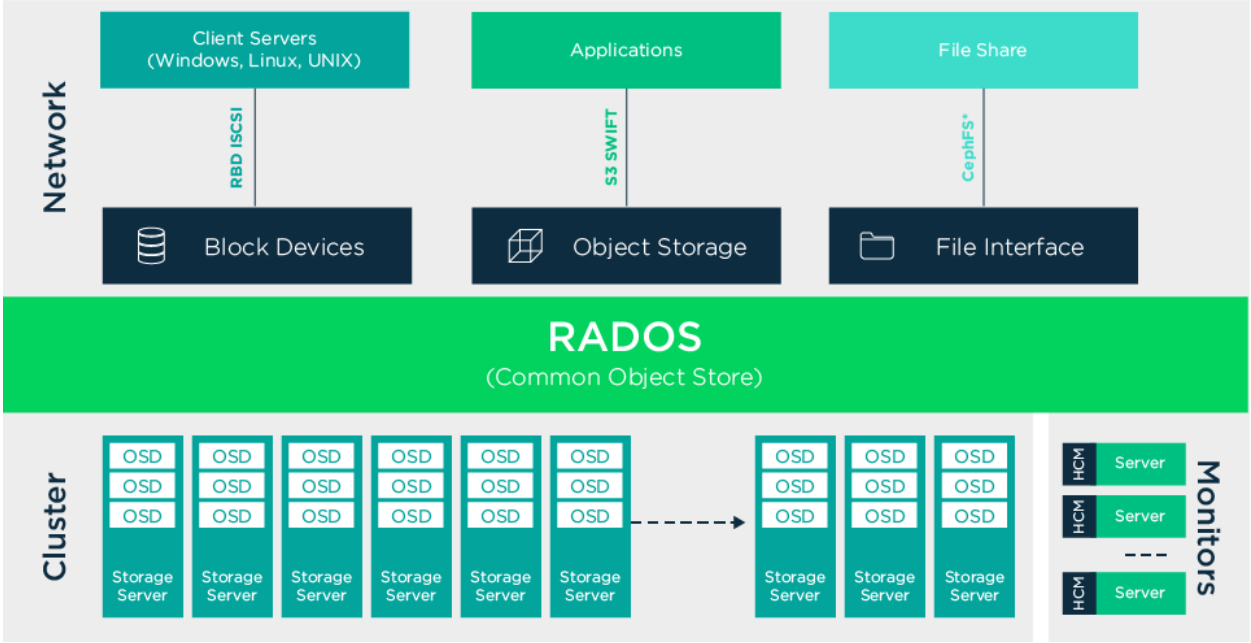


FIGURE 1: CEPH ARCHITECTURE

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

SPECIFIC TO THIS IMPLEMENTATION:

- One system is deployed as the administrative host server. The administration host is the Salt-master and hosts the SUSE Enterprise Storage Administration Interface, openATTIC, 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.

- Additional servers may be deployed as iSCSI gateway nodes. 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. This protocol is utilized for block-based connectivity to environments such as Microsoft Windows, VMware, and traditional UNIX. These systems may be scaled horizontally through client usage of multi-path technology.

- 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 assigned to the device stores data and manages the data replication and rebalancing processes. OSDs also communicate with the monitor (MON) nodes and provide them with the state of the other OSDs.

6.2 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 (backend) and client-facing (public) network traffic. This isolates Ceph OSD 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 2 (next page) shows the logical layout of the traditional Ceph cluster implementation.

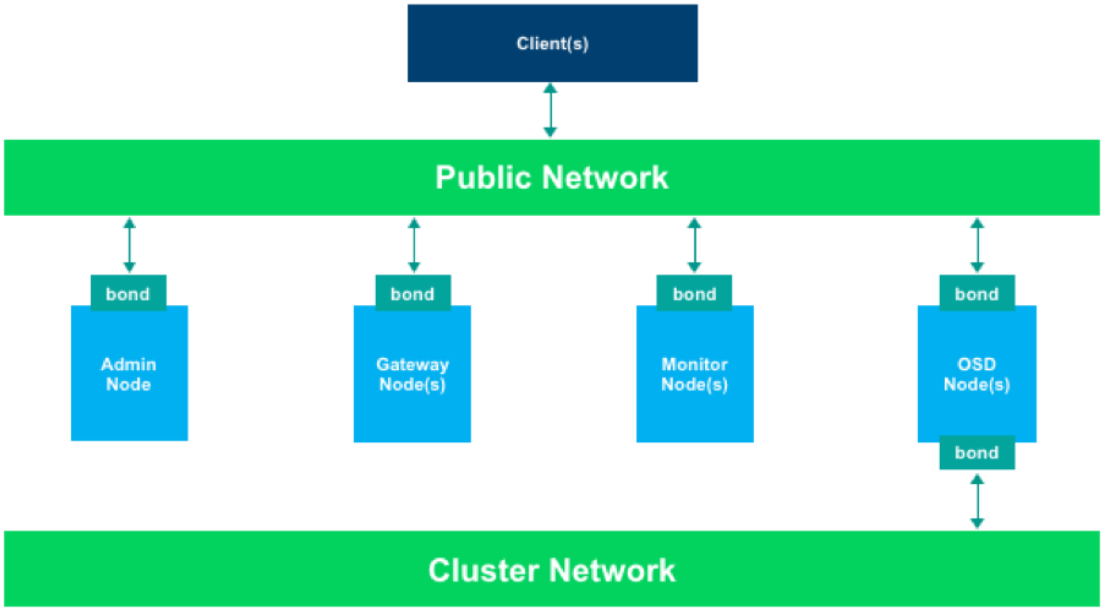


FIGURE 2: CEPH NETWORK ARCHITECTURE

7 Component Model

The preceding sections provided information 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 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 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.

8 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.

8.1 Network Deployment Overview

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

- Ensure that all network switches are updated with consistent firmware versions.

- Specific configuration for this deployment can be found in Appendix C: Network Switch Configuration & Appendix D: OS Networking 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. If not, it is recommended to use the admin node.

Function	Hostname	Primary Network	Cluster Network
Admin	sr630-1.suse.lab	172.16.250.31	N/A
Monitor	sr630-2.suse.lab	172.16.250.32	N/A
Monitor	sr630-3.suse.lab	172.16.250.33	N/A
Monitor	sr630-4.suse.lab	172.16.250.34	N/A
RGW/IGW	sr630-5.suse.lab	172.16.250.35	N/A
RGW/IGW	sr630-6.suse.lab	172.16.250.36	N/A
OSD	sr650-1.suse.lab	172.16.250.40	172.16.227.40
OSD	sr650-2.suse.lab	172.16.250.41	172.16.227.41
OSD	sr650-3.suse.lab	172.16.250.42	172.16.227.42
OSD	sr650-4.suse.lab	172.16.250.43	172.16.227.43
OSD	sr650-5.suse.lab	172.16.250.44	172.16.227.44
OSD	sr650-6.suse.lab	172.16.250.45	172.16.227.45
OSD	sr650-7.suse.lab	172.16.250.46	172.16.227.46
OSD	sr650-8.suse.lab	172.16.250.47	172.16.227.47
OSD	sr650-9.suse.lab	172.16.250.48	172.16.227.48
OSD	sr650-10.suse.lab	172.16.250.49	172.16.227.49

8.2 Hardware Recommended Actions

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.
- Verify BIOS/uEFI level on the physical servers correspond to those on the SUSE YES certification for all the platforms.
- Configure the boot media as RAID-1
- Configure all data and journal devices as individual RAID-0

8.2.1 Specific Hardware Configuration

To ensure maximum performance of the cluster, enter the bios system configuration and click UEFI Setup. Next click System Settings. Under Choose Operating Mode, change the setting to Maximum Performance

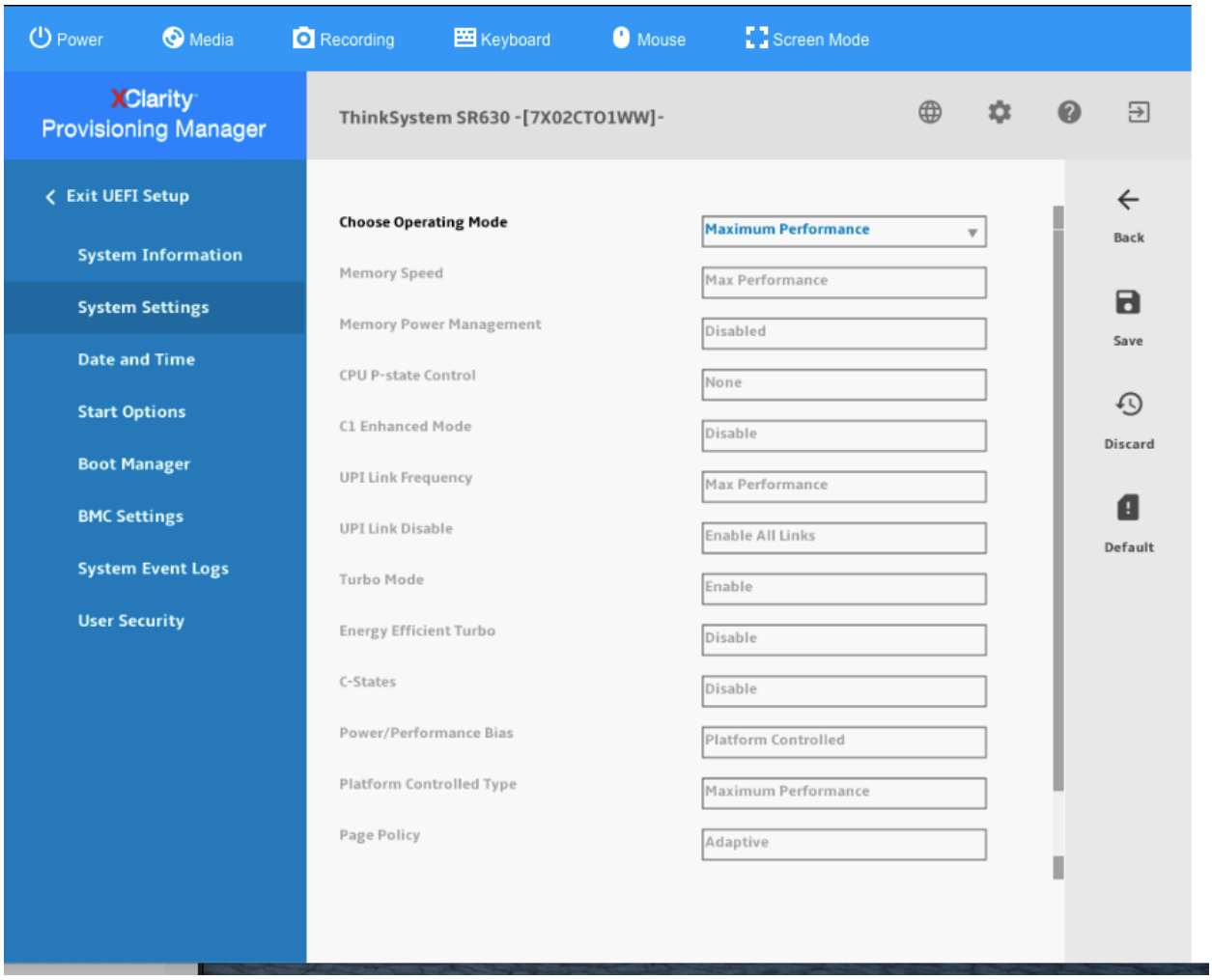


FIGURE 3: SETTING PERFORMANCE MODE

8.2.1.1 Set ConnectX-4 VPI NICS to Ethernet Mode

This configuration includes Mellanox ConnectX-4 VPI Network Interface Cards. These cards often arrive in Infiniband mode and need to be set to Ethernet mode. The way to affect this change involves following the steps outlined in the Mellanox manual for the inbox driver on SUSE Linux Enterprise 12 SP3 (http://www.mellanox.com/pdf/prod_software/SUSE_Linux_Enterprise_Server_(SLES)_12_SP3_Driver_User_Manual.pdf).

Replace the bold string with your PCI ID's

The steps required are:

```
# zypper in mstflint
# lspci |grep Mellanox
# mstconfig -d Your_PCI_ID s LINK_TYPE_P1=ETH
```

```
# mstconfig -d Your_PCI_ID s LINK_TYPE_P2=ETH
# reboot
```

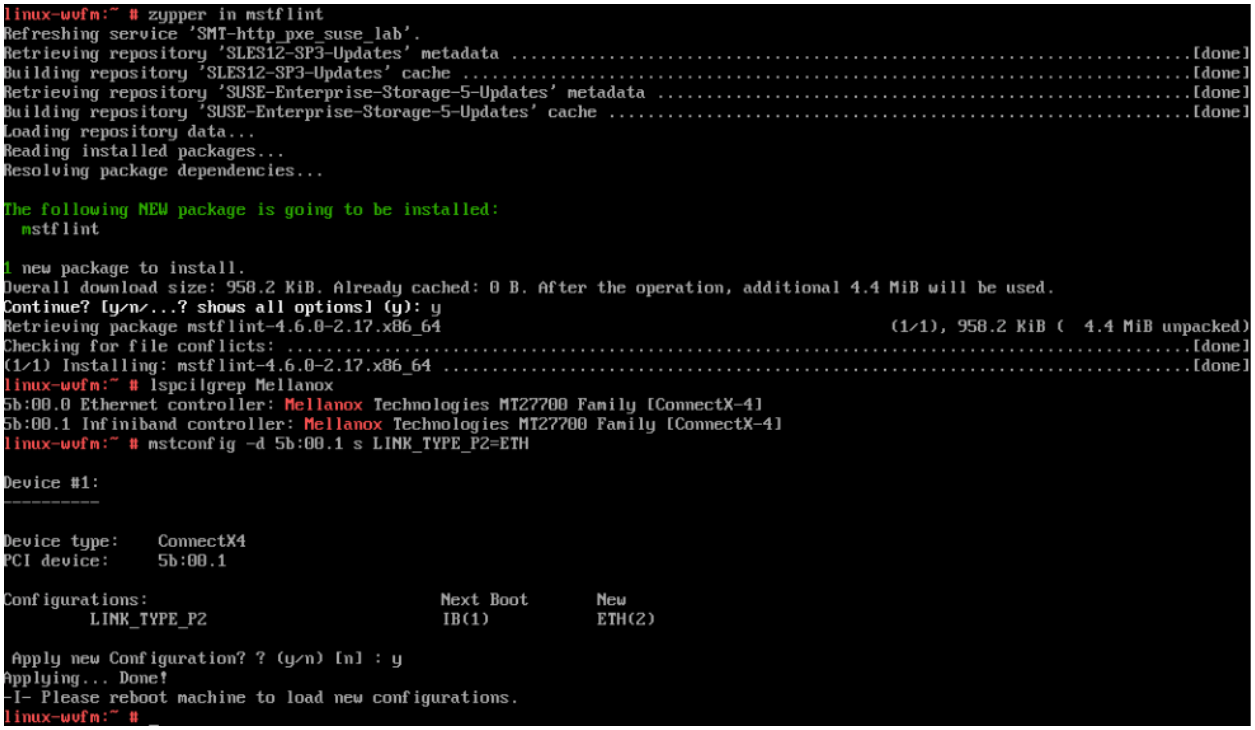


FIGURE 4: CHANGE MELLANOX CONNECTX-4 VPI NIC MODE

8.3 Operating System Installation

There are several key tasks to ensure are performed correctly during the operating system installation. During the SUSE Linux Enterprise installation, be sure and register the system with an update server. Ideally, this is a local SMT server which will reduce the time required for updates to be downloaded and applied to all nodes. By updating the nodes during installation, the system will deploy with the most up-to-date packages available, helping to ensure the best experience possible.

To speed installation, on the System Role screen, it is suggested to select KVM Virtualization Host. When the Installation Settings screen is reached, select **Software** and then un-check KVM Host Server. The resulting installation is a text mode server that is an appropriate base OS for SUSE Enterprise Server.

The next item is to ensure that the operating system is installed on the correct device. Especially on OSD nodes, the system may not choose the right drive by default. The proper way to ensure the right device is being used is to select **Create Partition Setup** on the Suggested Partitioning screen. This will then display a list of devices, allowing selection of the correct boot device. Next select **Edit Proposal Settings** and unselect the **Propose Separate Home Partition** checkbox.

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.

8.4 SUSE Enterprise Storage Installation & Configuration

8.4.1 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 over the network. This can be achieved through configuring unique host names per interface (eg osd1-cluster.suse.lab and osd1-public.suse.lab) in DNS and/or local /etc/hosts files.

Deepsea consists of a 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:

1. Install DeepSea on the Salt master which is the Admin node:

```
zypper in deepsea
```

2. Start the salt-master service and enable:

```
systemctl start salt-master.service
systemctl enable 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:
   Modify the entry for master in the */etc/salt/minion*

```
master: sesadmin.domain.com
```

5. Start the salt-minion service and enable:

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

6. List and accept all Salt keys on the Salt master: salt-key --accept-all and verify their acceptance:

```
salt-key --list-all
salt-key —-accept-all
```

7. Select the nodes to participate in the cluster:

```
salt '*' grains.append deepsea default
```

8. If the OSD nodes were used in a prior installation, zap ALL the OSD disks (ceph-disk zap <DISK>)

9. At this point, the cluster can be deployed.

a. Prepare the cluster:

```
salt-run state.orch ceph.stage.prep
```

b. Run the discover stage to collect data from all minions and create configuration fragments:

```
salt-run state.orch ceph.stage.disovery
```

c. A proposal for the storage layout needs to be generated at this time. For the hardware configuration used for this work, the following command was utilized:

```
salt-run proposal.populate name=lenovo target='sr650*'
```

The result of the above command is a deployment proposal for the disks that places the RocksDB, Write-Ahead Log (WAL), and on the same device.

d. 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 for the *policy.cfg* file used in the installation.

e. 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
```

f. 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.

g. Check for successful completion via:

```
ceph –s
```

h. Finally, deploy the services-gateways (iSCSI, RADOS, and openATTIC to name a few):

```
salt-run state.orch ceph.stage.services
```

8.4.2 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 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
```

8.5 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.

- 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 monitor 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 very 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 and 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.

9 Conclusion

The Lenovo ThinkSystem 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 Lenovo, any business can feel confident in the ability to address the exponential growth in storage they are currently faced with.

10 Appendix A: Bill of Materials

Role	Qty	Component	Notes
Admin, Monitor, Gateway, MDS Nodes	6	Lenovo ThinkSystem SR630	Configuration: • 1x Intel Xeon Silver 4116 • 32GB RAM • 2x 480GB M.2 in RAID-1 (OS) • 1x Mellanox ConnectX-4 Dual Port 100GbE
OSD Nodes	10	Lenovo ThinkSystem SR650	Configuration: • 1x Intel Xeon Gold 6142 • 96GB RAM • 2x 120GB M.2 in RAID-1 (OS) • 1x Mellanox ConnectX-4 Dual Port 100GbE • 12x 960GB SATA SSD
Network Switch	2	Lenovo ThinkSystem NE10032 Switch	Updated with latest OS image

11 Appendix B: Policy.cfg

```
cluster-ceph/cluster/*.sls
role-master/cluster/sr630-1.*.sls
role-admin/cluster/sr630-1.*.sls
role-mon/cluster/sr630-[234].*.sls
role-mgr/cluster/sr630-[234].*.sls
role-mds/cluster/sr630-[234].*.sls
role-igw/cluster/sr630-[56].*.sls
role-rgw/cluster/sr630-[56].*.sls
role-openattic/cluster/sr630-1.*.sls
config/stack/default/global.yml
config/stack/default/ceph/cluster.yml
profile-lenovo/cluster/sr650-*.sls
profile-lenovo/stack/default/ceph/minions/sr650-*.yml
```

12 Appendix C: Network Switch Configuration

The two switches are configured with a vlag and are named 100G-top and 100G-bottom. The load generation nodes are blade servers connected with 16 10Gb ethernet ports bonded in two LACP bonds, one to each switch. The cluster network carries back end and is VLAN 227. All ports are set to VLAN 127 as the default. The public network is VLAN 1.

```
show running-config
!
version "10.8.1.0"
!
hostname 100G-bottom
!
!
!
username admin role network-admin password encrypted XXXXXXXXXXXXX
feature restApi
ovsdb pki ovsdb_mgmt vrf management
ovsdb pki ovsdb_default vrf default
vlag tier-id 10
vlag priority 20
vlag isl port-channel 32
vlag hlthchk peer-ip 172.16.250.248
vlag enable
vlag instance 38 port-channel 32
vlag instance 38 enable
!
vlan 1
!
vlan 227
!
interface Ethernet1/3/1
 description "Blade Chassis left"
 switchport mode trunk
 switchport trunk allowed vlan 1,227
 mtu 9100
 channel-group 3 mode active
!
interface Ethernet1/3/2
 description "Blade Chassis left"
 switchport mode trunk
 switchport trunk allowed vlan 1,227
 mtu 9100
 channel-group 3 mode active
!
interface Ethernet1/3/3
```

```
 description "Blade Chassis left"
 switchport mode trunk
 switchport trunk allowed vlan 1,227
 mtu 9100
 channel-group 3 mode active
!
interface Ethernet1/3/4
 description "Blade Chassis left"
 switchport mode trunk
 switchport trunk allowed vlan 1,227
 mtu 9100
 channel-group 3 mode active
!
interface Ethernet1/4/1
 description "Blade Chassis left"
 switchport mode trunk
 switchport trunk allowed vlan 1,227
 mtu 9100
 channel-group 3 mode active
!
interface Ethernet1/4/2
 description "Blade Chassis left"
 switchport mode trunk
 switchport trunk allowed vlan 1,227
 mtu 9100
 channel-group 3 mode active
!
interface Ethernet1/4/3
 description "Blade Chassis left"
 switchport mode trunk
 switchport trunk allowed vlan 1,227
 mtu 9100
 channel-group 3 mode active
!
interface Ethernet1/4/4
 description "Blade Chassis left"
 switchport mode trunk
 switchport trunk allowed vlan 1,227
 mtu 9100
 channel-group 3 mode active
!
interface Ethernet1/9
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
#FIXME
```

```
!
interface Ethernet1/10
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/11
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/12
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/13
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/14
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/15
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/16
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/17
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
```

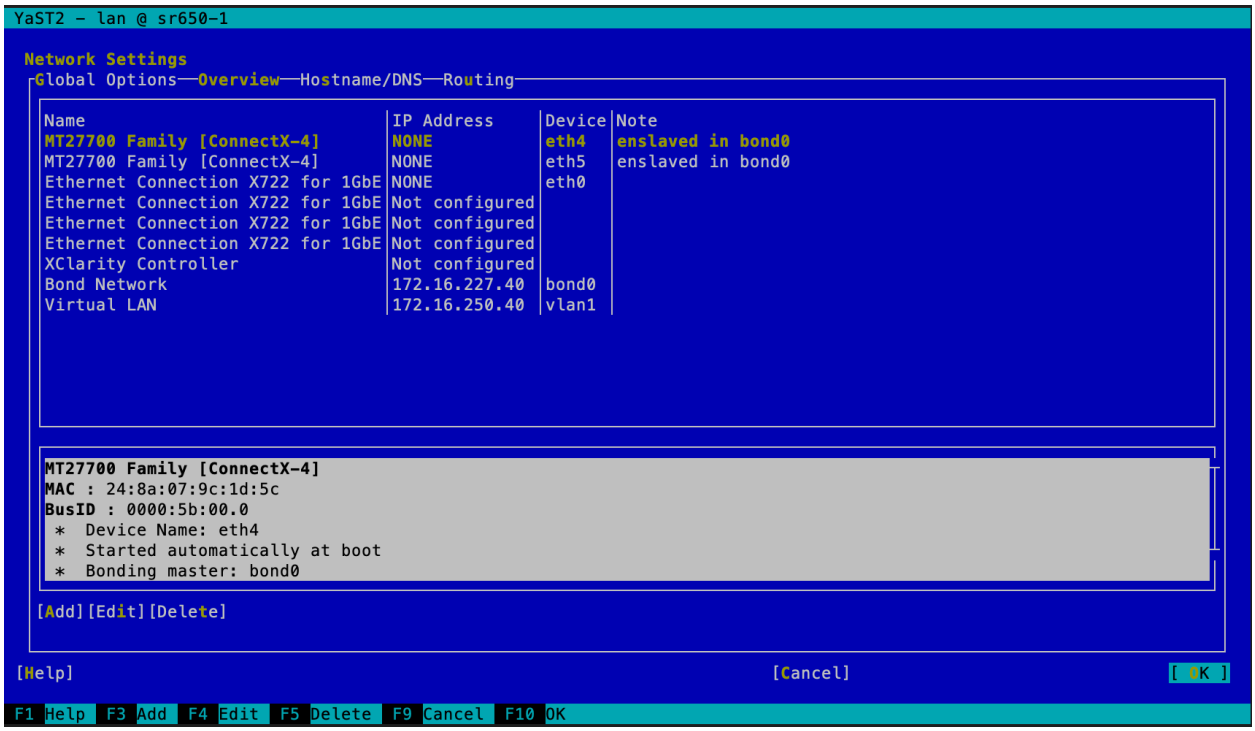
```
 mtu 9100
!
interface Ethernet1/18
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/19
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/20
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/21
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/22
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/23
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/24
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/25
 switchport mode hybrid
 switchport hybrid allowed vlan 1,227
```

```
 switchport hybrid native vlan 227
 mtu 9100
!
interface Ethernet1/31
 switchport mode trunk
 switchport trunk allowed vlan 1-2048
 mtu 9100
 channel-group 32 mode on
!
interface Ethernet1/32
 switchport mode trunk
 switchport trunk allowed vlan 1-2048
 mtu 9100
 channel-group 32 mode on
!
interface loopback0
 no switchport
!
interface Vlan1
 no switchport
 ip address 172.16.250.247/24
!
interface port-channel3
 switchport mode trunk
 switchport trunk allowed vlan 1,227
 mtu 9100
!
interface port-channel32
 switchport mode trunk
 switchport trunk allowed vlan 1-2048
 mtu 9100
!
ip route 0.0.0.0/0 Vlan1 172.16.250.1
!
line con 0
line vty 0 39
!
!
!
end
```

13 Appendix D: OS Networking Configuration

Each host is configured with an active passive bond. This alleviates the need for switch based configuration to support the bonding and still provides sufficient bandwidth for all IO requests

```
/etc/sysconfig/network # cat ifcfg-bond0
BONDING_MASTER='yes'
BONDING_MODULE_OPTS='mode=active-backup miimon=100'
BONDING_SLAVE0='eth4'
BONDING_SLAVE1='eth5'
BOOTPROTO='static'
BROADCAST=''
ETHTOOL_OPTIONS=''
IPADDR='172.16.227.40/24'
MTU='9000'
NAME=''
NETWORK=''
REMOTE_IPADDR=''
STARTMODE='auto'
#
/etc/sysconfig/network # cat ifcfg-vlan1
BOOTPROTO='static'
BROADCAST=''
ETHERDEVICE='bond0'
ETHTOOL_OPTIONS=''
IPADDR='172.16.250.40/24'
MTU=''
NAME=''
NETWORK=''
REMOTE_IPADDR=''
STARTMODE='auto'
VLAN_ID='1'
```


YaST2 - lan @ sr650-1
Network Settings
Global Options Overview Hostname/DNS Routing
Name IP Address Device Note
MT27700 Family [ConnectX-4] NONE eth4 enslaved in bond0
MT27700 Family [ConnectX-4] NONE eth5 enslaved in bond0
Ethernet Connection X722 for 1GbE NONE eth0
Ethernet Connection X722 for 1GbE Not configured
Ethernet Connection X722 for 1GbE Not configured
Ethernet Connection X722 for 1GbE Not configured
XClarity Controller Not configured
Bond Network 172.16.227.40 bond0
Virtual LAN 172.16.250.40 vlan1
MT27700 Family [ConnectX-4]
MAC : 24:8a:07:9c:1d:5c
BusID : 0000:5b:00.0
* Device Name: eth4
* Started automatically at boot
* Bonding master: bond0
[Add] [Edit] [Delete]
[Help]
[Cancel]
[OK]
F1 Help F3 Add F4 Edit F5 Delete F9 Cancel F10 OK

14 Appendix E: Performance Data

Comprehensive performance baselines are run as part of a reference build activity. This activity yields a vast amount of information that may be used to approximate the size and performance of the solution. The only tuning applied is documented in the implementation portion of this document.

The tests are comprised of a number of Flexible I/O (fio) job files run against multiple worker nodes. The job files and testing scripts may be found for review at: https://github.com/dm-byte/benchmaster. This is a personal repository and no warranties are made in regard to the fitness and safety of the scripts found there. The testing methodology involves two different types of long running tests. The types and duration of the tests have very specific purposes. There are both I/O simulation jobs and single metric jobs.

The length of the test run, in combination with the ramp-up time specified in the job file, is intended to allow the system to overrun caches. This is a worst-case scenario for a system and would indicate that it is running at or near capacity. Given that few applications can tolerate significant amounts of long tail latencies, the job files have specific latency targets assigned. These targets are intended to be in-line with expectations for the type of I/O operation being performed and set realistic expectations for the application environment.

The latency target, when combined with the latency window and latency window percentage set the minimum number of I/Os that must be within the latency target during the latency window time period. For most of the tests, the latency target is 20ms of less. The latency window is five seconds and the latency target is 99.99999%. The way that fio uses this is to ramp up the queue depth at each new latency window boundary until more than .00001% of all I/O's during a five second window are higher than 20ms. At that point, fio backs the queue depth down where the latency target is sustainable.

In the figures below, the x-axis labels indicate the block size in KiB on the top line and the data protection scheme on the bottom line. 3xrep is indicative of the Ceph standard 3 replica configuration for data protection while EC2 + 2 is Erasure Coded using the ISA plugin with k = 3 and m = 1. The Erasure Coding settings were selected to fit within the minimum cluster hardware size supported by SUSE.

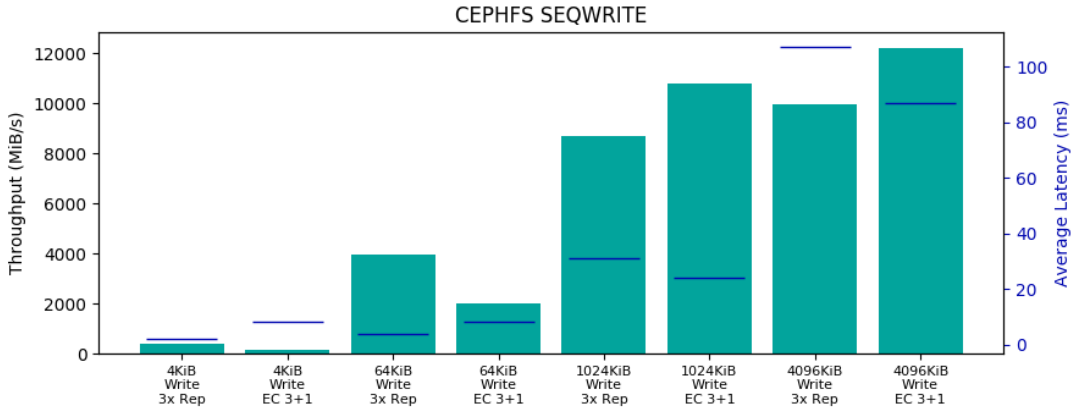
These settings, along with block size, max queue depth, jobs per node, and others, are all visible in the job files found at the repository link above.

Load testing was provided by an additional two Lenovo ThinkSystems on the same 100Gb network and 15 blade servers on a 10Gb Network.

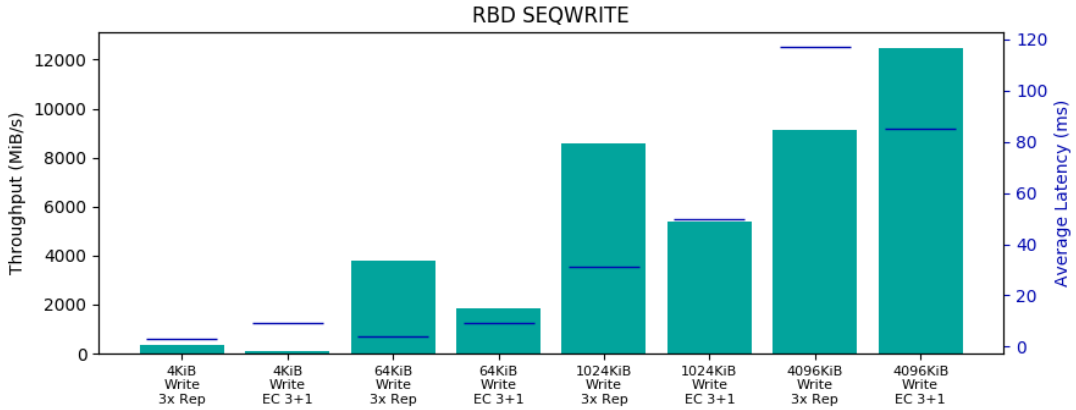
14.1 Sequential Writes

Sequential write I/O testing was performed across block sizes ranging from 4KiB to 4MiB.

These tests have associated latency targets: 4k is 10ms, 64k is 20ms, 1MiB is 100ms and 4MiB is 300ms.



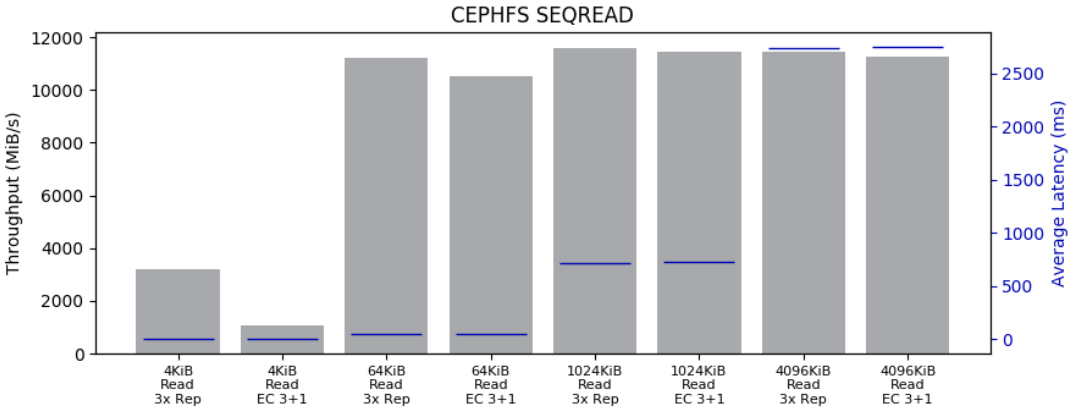
Protection	Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)
3x Rep	373	95626	2
EC 3+1	123	31516	8
3x Rep	3977	63636	4
EC 3+1	2006	32105	8
3x Rep	8672	8672	31
EC 3+1	10775	10775	24



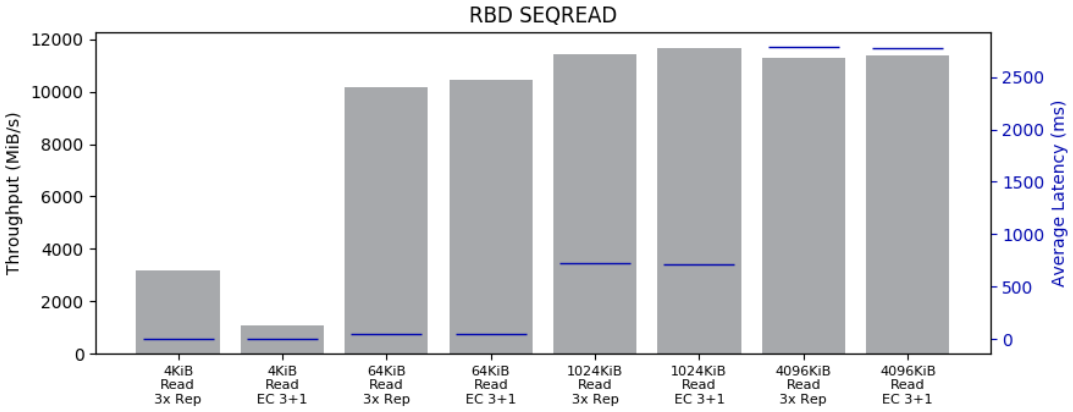
Protection	Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)
3x Rep	345	88524	3
EC 3+1	109	27909	9
3x Rep	3805	60881	4
EC 3+1	1844	29506	9
3x Rep	8604	8604	31
EC 3+1	5408	5408	50

14.2 Sequential Reads

The sequential read tests were conducted across the same range of block sizes as the write testing. The latency targets are only present for 4k sequential reads, where it is set to 10ms.



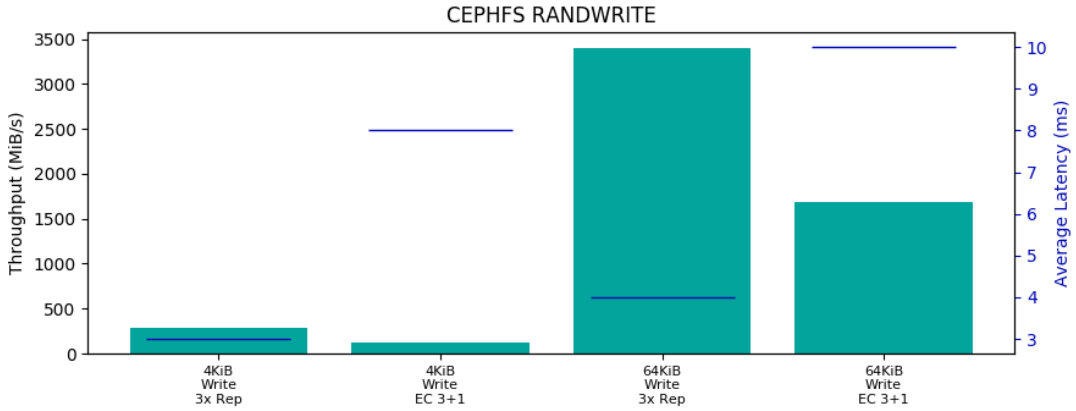
Protection	Read BW (MiB/s)	Read IOPS	Read Average Latency (ms)
3x Rep	3218	824058	0
EC 3+1	1057	270818	0
3x Rep	11193	179092	48
EC 3+1	10534	168543	51
3x Rep	11594	11589	711
EC 3+1	11456	11451	721



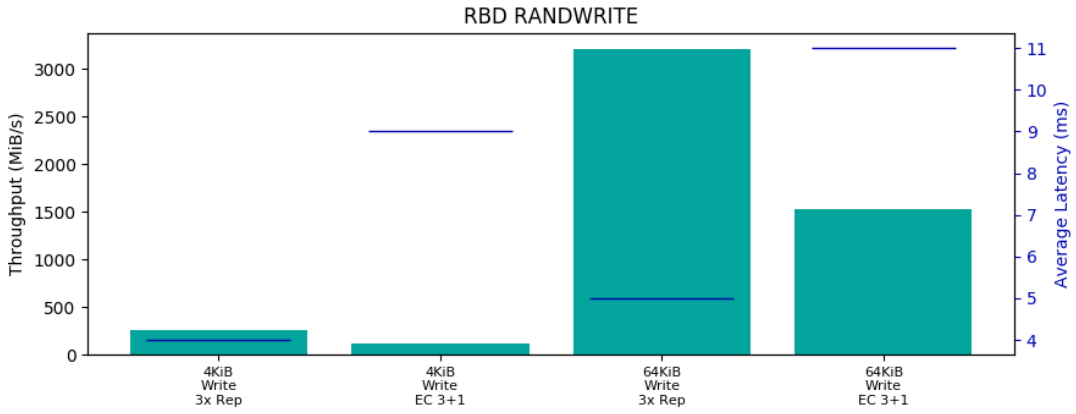
Protection	Read BW (MiB/s)	Read IOPS	Read Average Latency (ms)
3x Rep	3161	809270	0
EC 3+1	1057	270691	1
3x Rep	10185	162961	53
EC 3+1	10480	167687	51
3x Rep	11447	11442	718
EC 3+1	11678	11673	709

14.3 Random Writes

Random write tests were performed with the smaller I/O sizes of 4k and 64k. The 4k tests have a latency target of 10ms and the 64k tests have a latency target of 20ms.



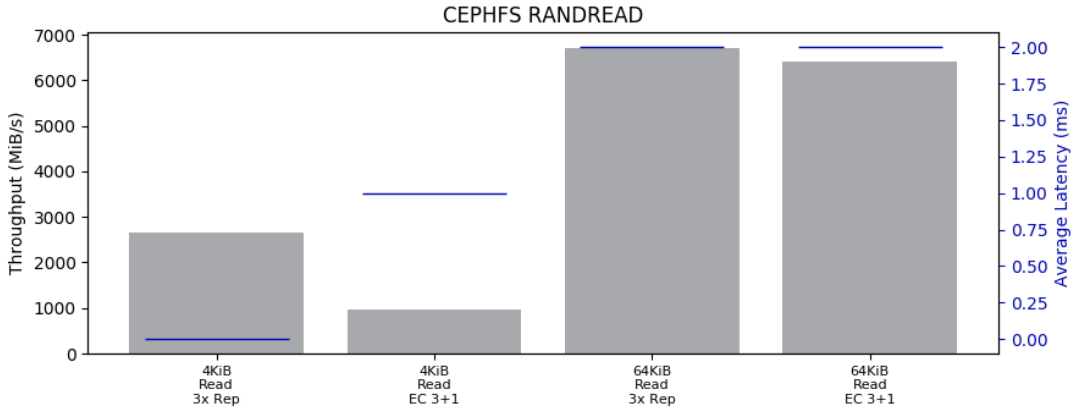
Protection	Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)
3x Rep	286	73459	3
EC 3+1	128	32835	8
3x Rep	3403	54452	4
EC 3+1	1686	26984	10



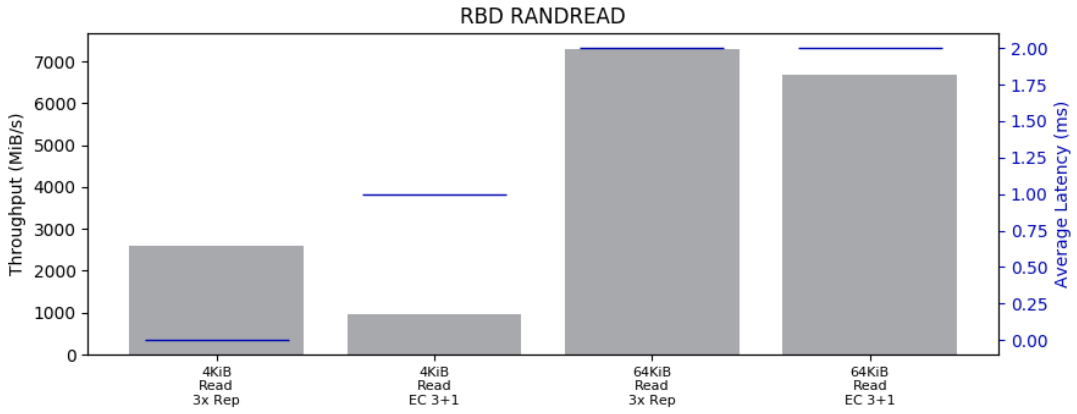
Protection	Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)
3x Rep	252	64556	4
EC 3+1	115	29471	9
3x Rep	3208	51339	5
EC 3+1	1527	24437	11

14.4 Random Reads

The random read tests were conducted on both 4k and 64k I/O sizes with latency targets of 10ms and 20ms respectively.



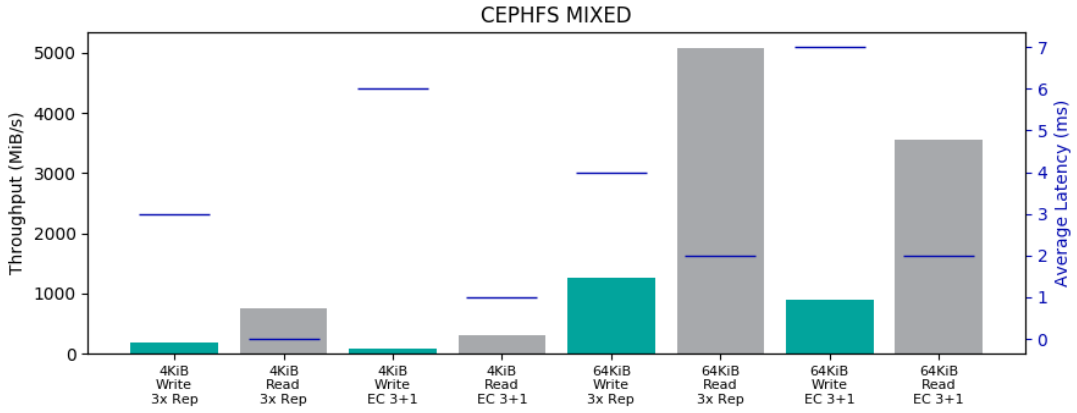
Protection	Read BW (MiB/s)	Read IOPS	Read Average Latency (ms)
3x Rep	2648	678019	0
EC 3+1	965	247211	1
3x Rep	6715	107450	2
EC 3+1	6413	102617	2



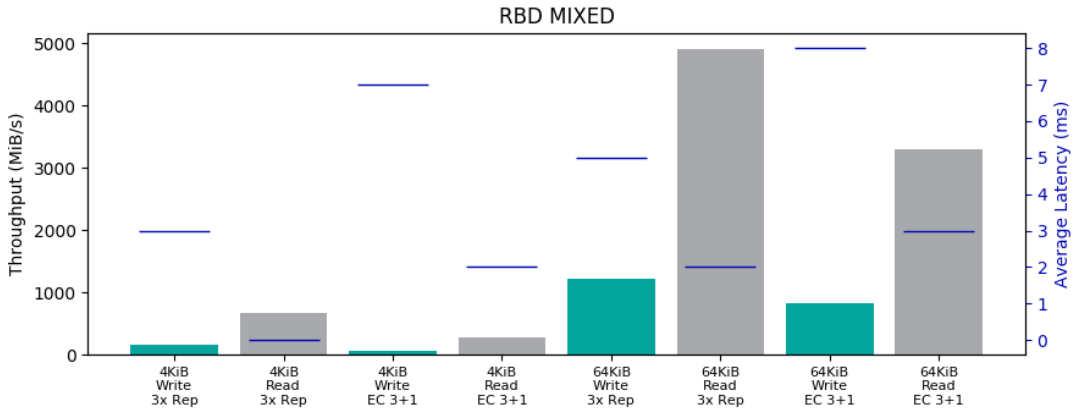
Protection	Read BW (MiB/s)	Read IOPS	Read Average Latency (ms)
3x Rep	2583	661341	0
EC 3+1	956	244898	1
3x Rep	7301	116830	2
EC 3+1	6669	106709	2

14.5 Mixed I/O

The mixed I/O tests were conducted on 4k and 64k I/O sizes with latency targets of 10ms and 20ms respectively. I/O is tested with 80% random reads and 20% random writes.



Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)	Protection	Read Average Latency (ms)	Read IOPS	Read BW (MiB/s)
188	48252	3	3x Rep	0	193187	754
75	19361	6	EC 3+1	1	77732	303
1265	20254	4	3x Rep	2	81321	5082
884	14150	7	EC 3+1	2	56926	3557



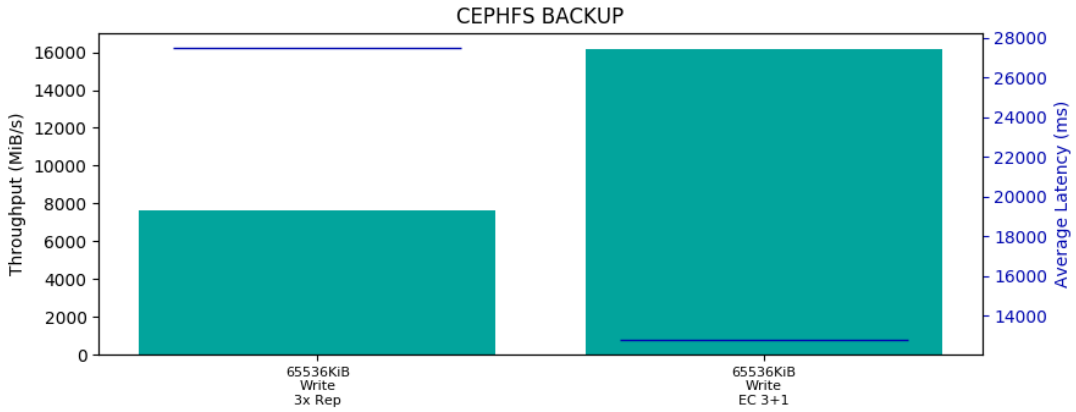
Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)	Protection	Read Average Latency (ms)	Read IOPS	Read BW (MiB/s)
166	42729	3	3x Rep	0	171270	669
67	17347	7	EC 3+1	2	69685	272
1223	19578	5	3x Rep	2	78631	4914
820	13128	8	EC 3+1	3	52882	3305

15 Workload Simulations

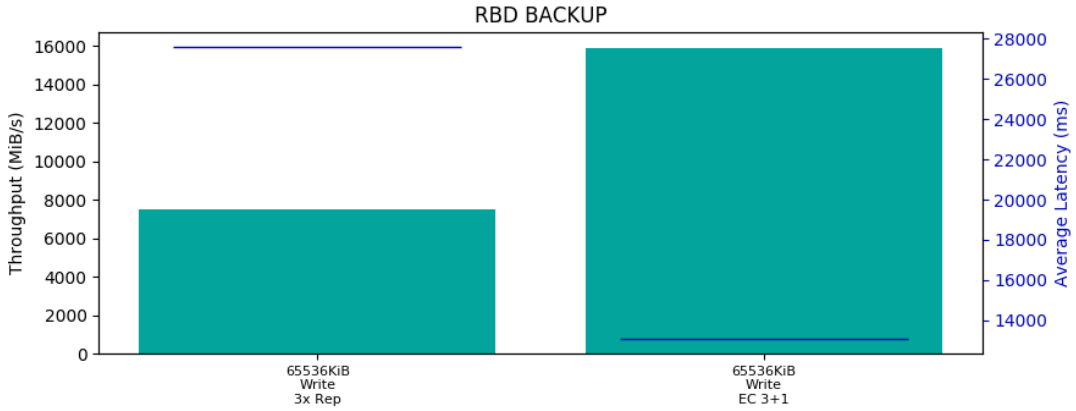
The following test results are workload oriented.

15.1 Backup Simulation

The backup simulation test attempts to simulate the SUSE Enterprise Storage cluster being used as a disk-based backup target that is either hosting file systems on RBDs or is using CephFS. The test had a latency target of 200ms at the time of the test run. The latency target has since been removed.



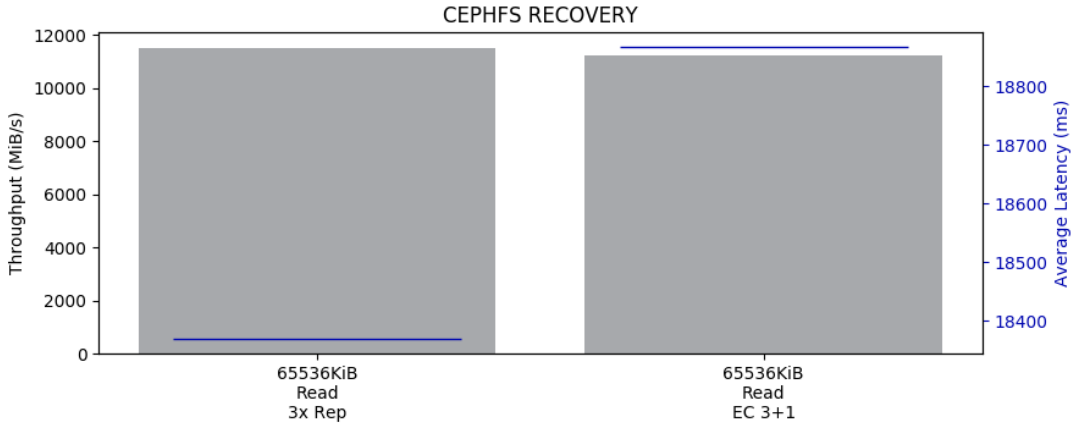
Protection	Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)
3x Rep	7618	117	27465
EC 3+1	16178	251	12794



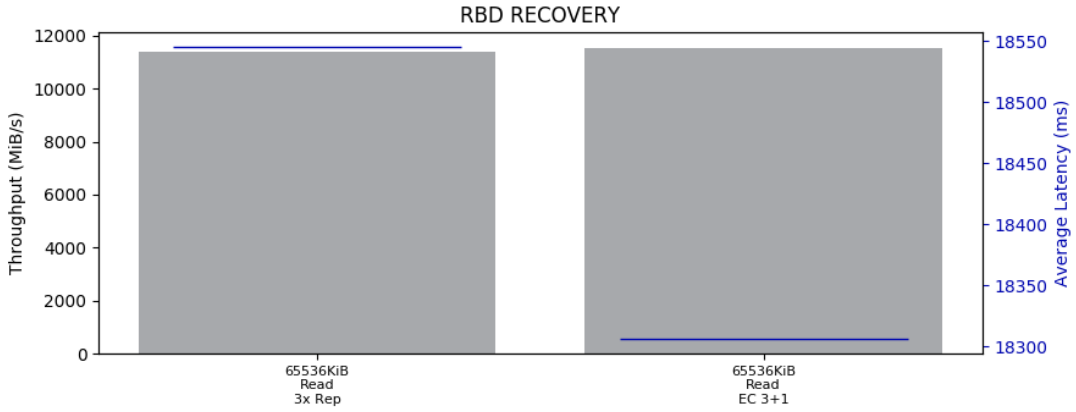
Protection	Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)
3x Rep	7522	115	27582
EC 3+1	15892	246	13074

15.2 Recovery Simulation

The recovery workload is intended to simulate recovery jobs being run from SUSE Enterprise Storage. It tests both RBD and CephFS.



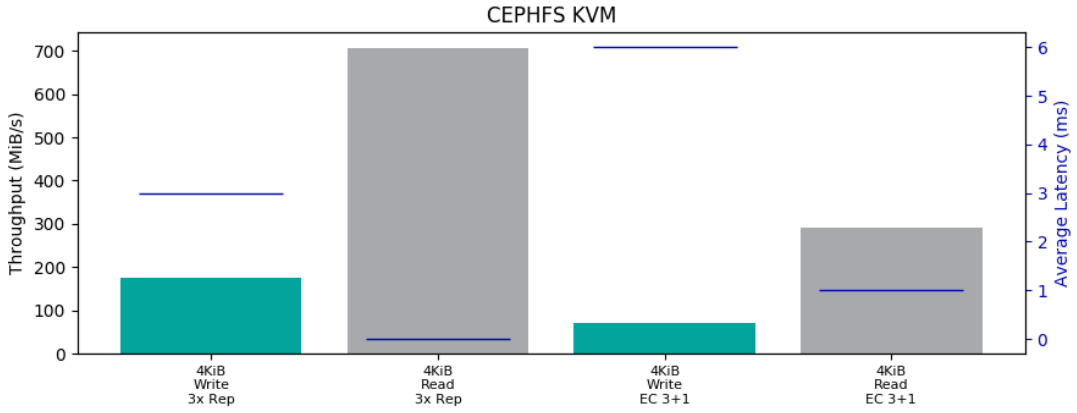
Protection	Read BW (MiB/s)	Read IOPS	Read Average Latency (ms)
3x Rep	11523	178	18369
EC 3+1	11239	173	18867



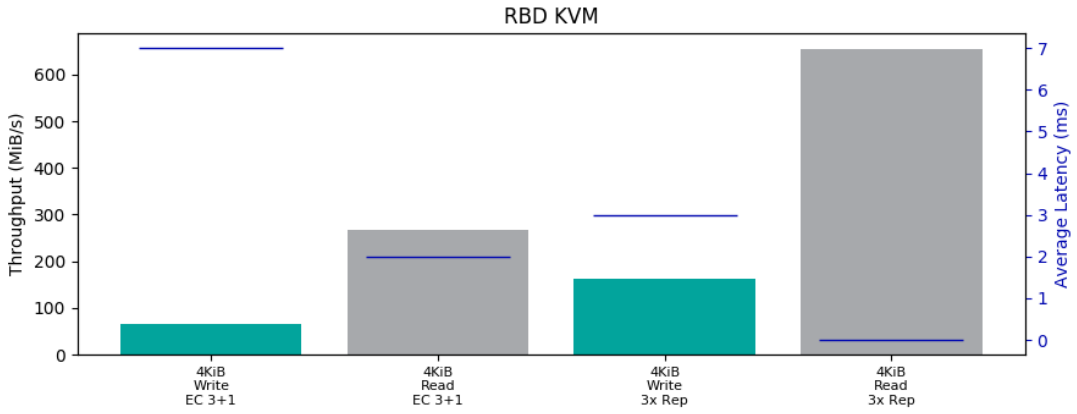
Protection	Read BW (MiB/s)	Read IOPS	Read Average Latency (ms)
3x Rep	11418	176	18545
EC 3+1	11542	178	18306

15.3 KVM Virtual Guest Simulation

The kvm-krbd test roughly simulates virtual machines running. This test has a 20ms latency target and is 80% read with both reads and writes being random.



Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)	Protection	Read Average Latency (ms)	Read IOPS	Read BW (MiB/s)
176	45224	3	3x Rep	0	181233	707
72	18615	6	EC 3+1	1	74759	292



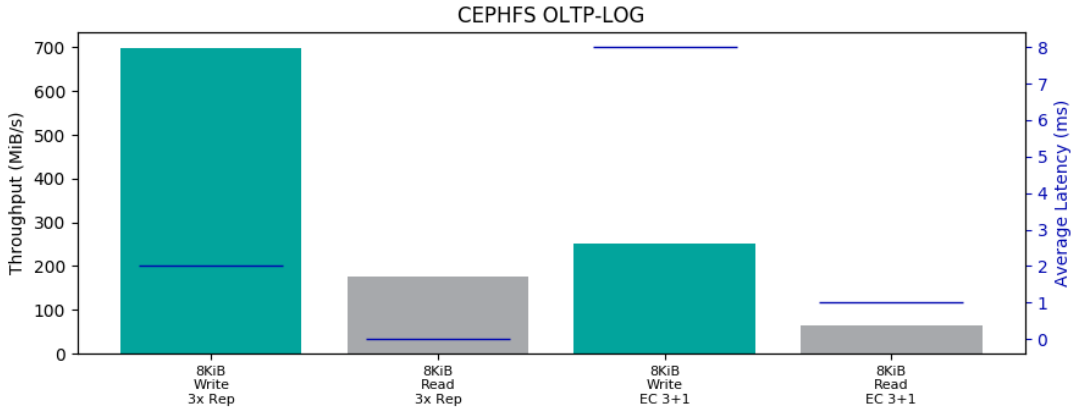
Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)	Protection	Read Average Latency (ms)	Read IOPS	Read BW (MiB/s)
66	17083	7	EC 3+1	2	68615	268
163	41876	3	3x Rep	0	167831	655

15.4 Database Simulations

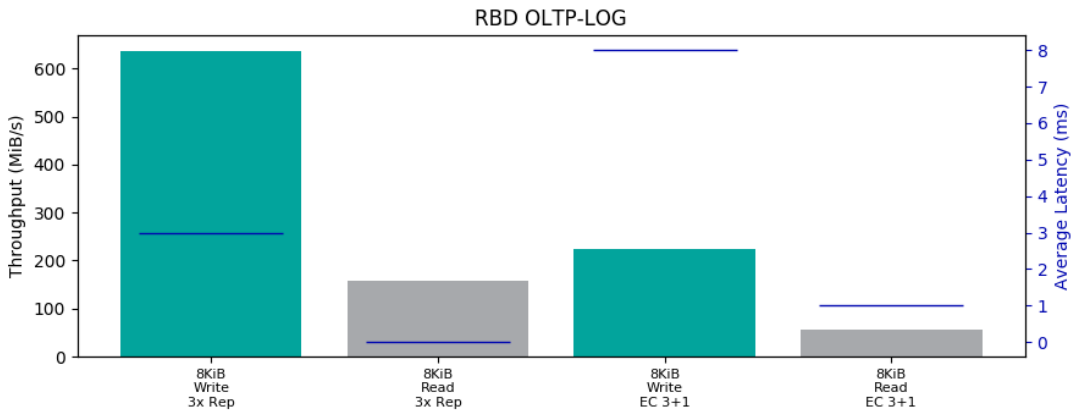
It is important to keep sight of the fact that Ceph is not designed for high performance database activity. These tests provide a baseline understanding of performance expectations should a database be deployed using SUSE Enterprise Storage.

15.5 OLTP Database Log

The database log simulation is based on documented I/O patterns from several major database vendors. The I/O profile is 80% sequential 8KB writes with a latency target of 1ms.



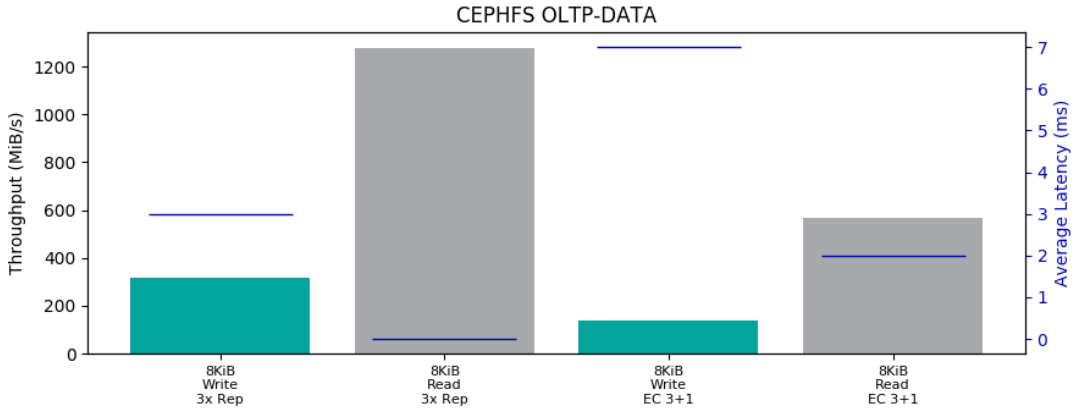
Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)	Protection	Read Average Latency (ms)	Read IOPS	Read BW (MiB/s)
699	89552	2	3x Rep	0	22442	175
251	32239	8	EC 3+1	1	8101	63



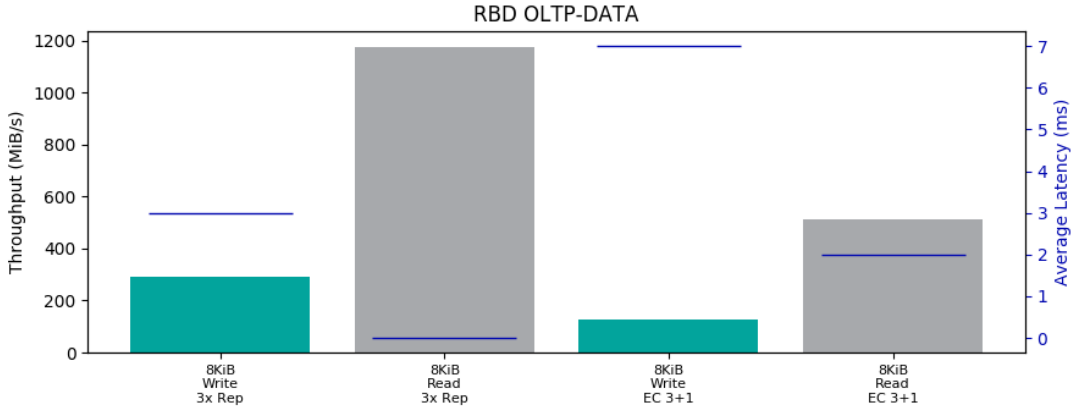
Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)	Protection	Read Average Latency (ms)	Read IOPS	Read BW (MiB/s)
637	81594	3	3x Rep	0	20435	159
224	28756	8	EC 3+1	1	7205	56

15.6 OLTP Database Datafile

The OLTP Datafile simulation is set for an 80/20 mix of random reads and writes. The latency target is 10ms.



Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)	Protection	Read Average Latency (ms)	Read IOPS	Read BW (MiB/s)
318	40813	3	3x Rep	0	163590	1278
140	18031	7	EC 3+1	2	72424	565



Write BW (MiB/s)	Write IOPS	Write Average Latency (ms)	Protection	Read Average Latency (ms)	Read IOPS	Read BW (MiB/s)
293	37593	3	3x Rep	0	150593	1176
128	16388	7	EC 3+1	2	65813	514

16 Resources

SUSE Enterprise Storage Technical Overview https://www.suse.com/docrep/documents/1mdg7e-q2kz/suse_enterprise_storage_technical_overview_wp.pdf

SUSE Enterprise Storage v5—Deployment Guide https://www.suse.com/documentation/suse-enterprise-storage-5/book_storage_deployment/data/book_storage_deployment.html

SUSE Linux Enterprise Server 12 SP3—Administration Guide https://www.suse.com/documentation/sles-12/book_sle_admin/data/book_sle_admin.html

Subscription Management Tool for SLES 12 SP3 https://www.suse.com/documentation/sles-12/book_smt/data/book_smt.html