Table of Contents

Super Micro Computer, Inc.
980 Rock Avenue
San Jose, CA 95131 USA
www.supermicro.com

Application Notes

SUSE Enterprise Storage v5 Implementation Guide For Supermicro SuperServer Platforms

Introduction

The objective of this guide is to present a step-by-step guide on how to implement SUSE Enterprise Storage (v5) on the Supermicro server 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 quickly deploy a robust SUSE Enterprise Storage cluster on the Supermicro server platform. Its goal is to show architectural best practices and how to build a Ceph-based cluster that will support the implementation of any of the currently supported protocols.

Upon completion of the steps in this document, a working SUSE Enterprise Storage (v5) will be operational as described in the SUSE Enterprise Storage 5 Deployment Guide

Target Audience

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

Business problem and business value

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

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, 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) has taken place. These ever increasing "data lakes" need different access methods: file, block, or object.

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

Business value

This software defined storage solution enables transformation of the enterprise infrastructure by providing a unified platform where structured and unstructured data can co-exist and be accessed as file, block, or object, depending on 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.

Requirements

Enterprise storage systems require reliability, manageability, and serviceability, which together are known as RAS. The legacy 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 affordable hardware from Supermicro, the result is a solution that meets the customer's expectation(s).

Functional requirements

A SUSE Enterprise Storage solution is:

- Simple to setup and deploy, within the documented guidelines of system hardware, networking and environmental prerequisites.

Ceph admin, monitor, and protocol gateway functions:

2 Supermicro 1U SYS-6019U-DCCP

- 128GB RAM
- 2 2TB SATA using RAID-1
- 2x Intel® Xeon® Gold 6130 16C/32T @2.1 GHz
- Intel Ethernet Controller XL710 for 40GbE QSFP+

Ceph admin, monitor, and protocol gateway functions:

1 Supermicro 2U SYS-6029TP-DCFN with 4 nodes

- Per Node Configuration
- 128GB RAM
- 2 2TB SATA using RAID-1
- 2x Intel® Xeon® 4110 8C/16T @2.1 GHz
- Intel Ethernet Controller XL710 for 40GbE QSFP+

Storage Nodes:

4 Supermicro SYS-6029U-DCSO

- 128GB RAM
- 2x Intel® Xeon® 4110 8C/16T @2.1 GHz
- 2x 2TB SATA in RAID-1
- 10x 8TB SATA
- INTEL Non-Volatile Memory Enterprise (NVMe) Solid State Drive SSDPEDKE020T7

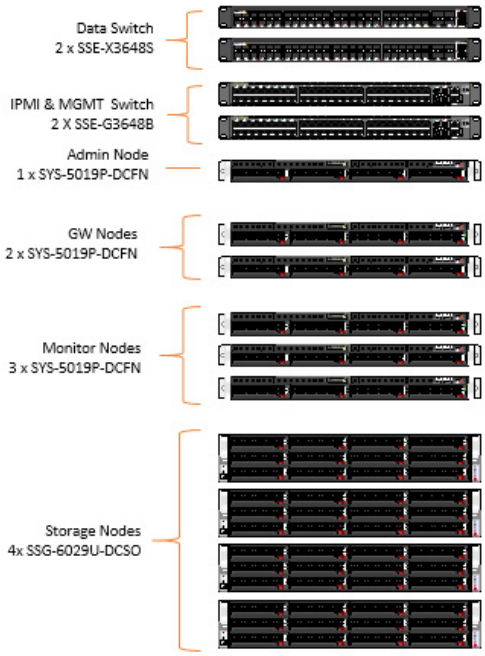
Software:

SUSE Enterprise Storage 5

- 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.

Configuration

The SUSE Enterprise Storage cluster leveraged three models of Supermicro servers. The role/functionality of each SUSE Enterprise Storage component will be explained in more detail in the architectural overview section. There were two unique server types tested for this RA that can provide the admin, monitor, and protocol gateway functions.



Switching infrastructure:

1 Supermicro 40GbE/100GbE SDN SuperSwitch

- AIC Intel® SSD DC P4600 Series (2.0TB, 1/2 Height PCIe 3.1 x4, 3D1, TLC)
- Intel Ethernet Controller XL710 for 40GbE QSFP+

Please note: The SUSE Enterprise Storage subscription includes a limited use [for SUSE Enterprise Storage] entitlement for SUSE Linux Enterprise Server

Key Benefits

1. An SSD optimized for cloud storage architectures
2. Optimized for caching across a range of workloads
3. Manageability to maximize IT efficiency
4. Industry-leading reliability and security
5. Designed for today's modern data centers

Performance

1. Sequential Read (up to) 3200 MB/s
2. Sequential Write (up to) 1575 MB/s
3. Random Read (100% Span) 610000 IOPS
4. Random Write (100% Span) 196650 IOPS
5. Latency - Read 85 µs
6. Latency - Write 15 µs
7. Power - Active Sequential Avg. 17W (Write), 9.4W (Read)
8. Power - Idle <5 W

Reliability

1. Vibration - Operating 2.17 GRMS
2. Vibration - Non-Operating 3.13 GRMS
3. Shock (Operating and Non-Operating) 50 G Trapezoidal, 170 in/s
4. Operating Temperature Range 0°C to 35°C
5. Endurance Rating (Lifetime Writes) 11.08 PBW
6. Mean Time Between Failures (MTBF) 2 million hours
7. Uncorrectable Bit Error Rate (UBER) <1 sector per 10^17 bits read
8. Warranty Period 5 yrs

P4600 NVMe Drives

Optimized for Caching Across a Range of Workloads

This cloud-inspired SSD is built with an entirely new NVMe controller that is optimized for mixed workloads commonly found in data caching and is architected to maximize CPU utilization.

With controller support for up to 128 queues, the DC P4600 helps minimize the risk of idle CPU cores and performs most effectively on Intel platforms with Intel® Xeon® processors. The queue pair-to-CPU core mapping supports high drive count and also supports multiple SSDs scaling on Intel platforms.

With the DC P4600, data centers can accelerate caching to enable more users, add more services, and perform more workloads per server. Now you can cache faster and respond faster.

Intel has built industry-leading end-to-end data protection into the DC P4600. This includes protection from silent data corruption, which can cause catastrophic downtime and errors in major businesses.

Power Loss Imminent (PLI) provides protection from unplanned power loss, and is obtained through a propriety combination of power management chips, capacitors, firmware algorithms, and a built-in PLI self-test. Intel's PLI feature provides data centers with high confidence of preventing data loss during unplanned power interruptions.

Designed for Today's Modern Data Centers

The DC P4600 is Intel's new 3D NAND SSD for mixed workloads that are common to the data caching needs of cloud-driven data centers. The mix of performance, capacity, endurance, manageability, and reliability make it the ideal solution for data caching in software-defined and converged infrastructures.

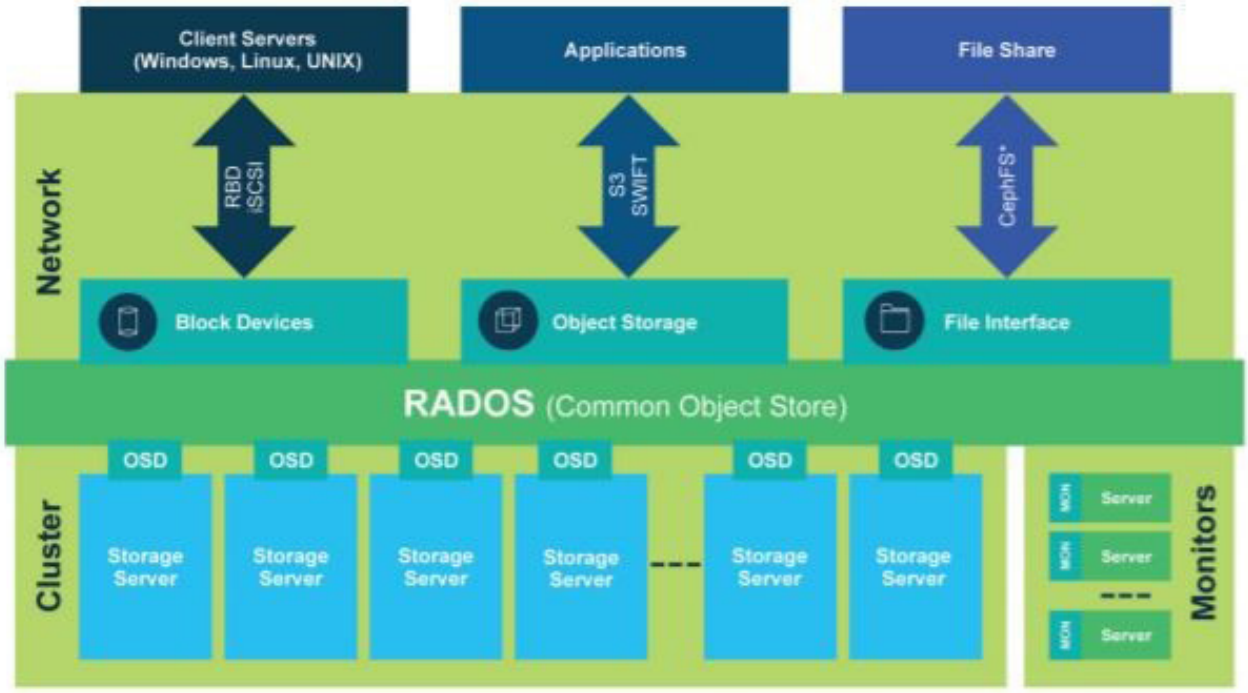
Architectural overview

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

Solution architecture

SUSE Enterprise Storage provides unified block, file, and object access based on Ceph. Ceph is a distributed storage solution designed for scalability, reliability and performance. A critical component of Ceph is the RADOS object storage. RADOS enables a number of 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. While these gateways may be thought of as a limiting factor, the ISCSI and S3 gateways can scale horizontally using load balancing techniques.



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), and three monitor nodes (MONs). Specific to this implementation:

- One TwinPro node is deployed as the administrative host server. The administration host is the salt-master and hosts openATTIC, the central management system which supports the cluster.
- Three additional TwinPro nodes are deployed as (MONs). Monitor nodes maintain information about the cluster health state, a map of the other monitor nodes and a CRUSH map. They also keep a history of changes performed to the cluster.
- Additional 1U or TwinPro nodes 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 may also be deployed on TwinPro or 1U nodes. 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.
- This configuration uses Supermicro SYS-6029U-DCSO 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 OSD stores data and manages the data replication and rebalancing processes. OSD daemons also communicate with the MONs and provide them with the state of the other OSD daemons.
- One particular focus of the OSD node design was to accelerate Write Ahead Log (WAL) and RocksDB performance by placing them on an Intel P4600 2TB NVME PCIe Solid State Drive. This provides a dedicated high throughput, low latency storage location for these critical services. Using the NVME helps reduce long tail latencies and ensures more consistent performance for the solution.

Networking architecture

A software-defined solution is 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 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 switches.

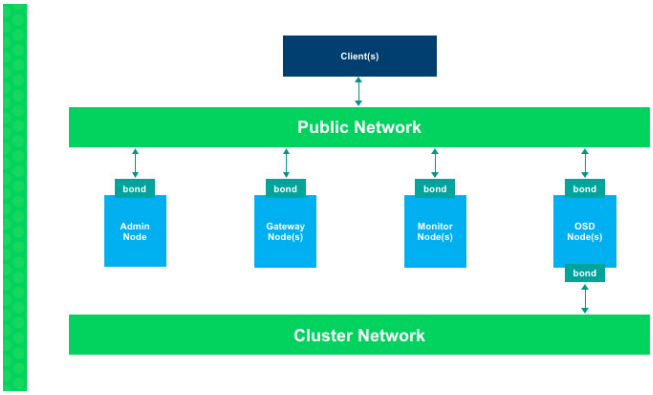


Figure 2 shows the logical layout of the traditional Ceph cluster implementation.

In this particular configuration, two VLANs were utilized to segment the traffic. The default VLAN of 1 was untagged and association with the 192.168.101.0/24 network. The cluster network of 192.168.100.0/24 was tagged to VLAN 100 and only configured on OSD nodes. Separate GbE interfaces provided admin network interfaces.

Network/IP address scheme

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

Function	Hostname	Public Network	Cluster Network	Admin Network
Admin (Host)	salt.supermicro.lab	192.168.101.90	N/A	192.168.124.90
Monitor	mon1.supermicro.lab	192.168.101.101	N/A	192.168.124.101
Monitor	mon2.supermicro.lab	192.168.101.102	N/A	192.168.124.102
Monitor	mon3.supermicro.lab	192.168.101.103	N/A	192.168.124.103
Object/ISCSI Gateway	gw1.supermicro.lab	192.168.101.104	N/A	192.168.124.104
Object/ISCSI Gateway	gw2.supermicro.lab	192.168.101.105	N/A	192.168.124.105
OSD Node	osd1.supermicro.lab	192.168.101.111	192.168.100.111	192.168.124.111
OSD Node	osd2.supermicro.lab	192.168.101.112	192.168.100.112	192.168.124.112
OSD Node	osd3.supermicro.lab	192.168.101.113	192.168.100.113	192.168.124.113
OSD Node	osd4.supermicro.lab	192.168.101.114	192.168.100.114	192.168.124.114

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 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 SLES12 SP3 – 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. The most recent release brings a new underlying storage technology, Bluestore, to the product. Bluestore significantly reduces long tail latencies and significantly improves performance of some use cases. SUSE Enterprise Storage 5 also brings the distributed file system, CephFS to production with multiple meta-data server support, allowing for broad usage of this highly-performant, scale-out technology across many use cases.

Deployment

This deployment section should be seen as a supplement online documentation - specifically, the SUSE Enterprise Storage 5 Deployment Guide as well as SUSE Linux Enterprise Server Administration Guide. 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 12 SP3 to make one available. The emphasis is on specific design and configuration choices.

Network Deployment overview/outline

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

- Ensure that all network switches are updated with consistent firmware versions.
- Configure 802.3ad for system port bonding and IRF between the switches, plus enable jumbo frames.
- Specific configuration for this deployment can be found in Appendix D: Network Switch 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.

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.
- Verify BIOS/UEFI level on the physical servers correspond to those on the SUSE YES certification for the Supermicro platforms.
- Configure a mirrored pair of drives for the operating system
- Configure all data and journal devices as individual RAID-0

Operating System Deployment and Configuration

When deploying the operating system, be sure to utilize only the correct device. This is the RAID-1 created during hardware configuration.

- Properly configure the network devices during installation. This is illustrated in Appendix E: OS Networking Configuration.
- Register the system against the SMT server.
- When prompted for extensions, select SUSE Enterprise Storage
- On the Suggested Partitioning selection, select Edit Proposal Settings and uncheck Propose Separate Home Partition
- After installation is complete, run zypper up to ensure all current updates are applied.

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):

1. 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.

2. 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 (osd1-cluster.supermicro.lab and osd1-public.supermicro.lab) in DNS and/or local /etc/hosts files.
3. 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:

- Install the salt-master packages on the admin node:

```
zypper in salt-master
```

- 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
  - In this case: master: sesadmin.domain.com
- Start the salt-minion service and enable:

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

- Clear all non-OS drives on the OSD nodes, reset the labels, and reboot the nodes:

```
dd if=/dev/zero of=/dev/sda bs=1M count=1024 oflag=direct
sgdisk -Z --clear -g /dev/sda
reboot
```

- 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
```

- Install DeepSea on the Salt master which is the Admin node:

```
zypper in DeepSea
```

- At this point, you can deploy and configure the cluster:
  - Prepare the cluster: deepsea stage run ceph.stage.prep
  - Run the discover stage to collect data from all minions and create configuration fragments:

```
deepsea stage run ceph.stage.discovery
```

  - A default proposal is generated by the previous stage, however, we desire to use a custom proposal generated by:

```
salt-run proposal.populate name=nvme ratio=9 wal=1700-2000
db=1700-2000 target='osd*' db-size=45g wal-size=5g data=1800-2200
```

  - 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 C for 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

```
deepsea stage run ceph.stage.configure
```

  - The last two steps manage the actual deployment. Deploy monitors and ODS daemons first:

```
deepsea stage run 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): deepsea stage run 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 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
```

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 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 MINs as an odd number. However, only very large or very distributed clusters would likely need beyond the three monitor nodes cited in this reference implementation. For performance reasons, it is recommended to use distinct nodes for the Mon roles, so that the OSD nodes can be scaled as capacity requirements dictate.

As described in this implementation guide as well as the SUSE Enterprise Storage documentation, a minimum of four OSD nodes is recommended, with the default replication setting of 3. This will ensure cluster operation, even with the loss of a complete OSD node. Generally speaking, performance of the overall cluster increases as more properly configured OSD nodes are added.

Appendix A: Bill of Materials

Component / System

Role	Qty	Component	Notes
Admin/MON/ Gateway servers	1	SYS-6029TP-DCFN	Each node consists of: • 2x Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz • 128GB RAM • 2x 2TB SATA in RAID-1 • 1 Intel XL710 w/40GbE QSFP+
Admin/MON/ Gateway servers	2	SYS-6029U-DCFN	Each node consists of: • 2 2TB SATA using RAID-1 • 128GB RAM • 2x Intel(R) Xeon(R) 2630V4 CPU @ 2.20GHz • Intel Ethernet Controller XL710 for 40GbE QSFP+
OSD Hosts	4	SYS-6029U-DCFN	Each node consists of: • 2x Intel(R) Xeon(R) CPU E5-2630L v4 @ 1.80GHz • 128GB RAM • 2x 2TB SATA drives in RAID-1 • 10x SATA drives • 1 Intel XL710 w/40GbE QSFP+
Software	1	SUSE Enterprise Storage Subscription Base configuration	Allows for 4 storage nodes and 6 infrastructure nodes

Appendix B: OSD Drive and Journal Proposal Changes

The proposal generated by salt-run proposal.populate name=nvme ratio=9 wal=1700-2000 db=1700-2000 target='osd*' db-size=55g wal-size=2g and selected for use is named: profile-nvme. The contents of the proposal are below.

```
ceph:
  storage:
    osds:
      /dev/disk/by-id/ata-HGST _ HUS724040ALE640 _ PK2331PAG66ART:
        db: /dev/disk/by-id/nvme-INTEL _ SSDPEDKE020T7 _
PHLE725100252P0IGN
        db _ size: 50g
        format: bluestore
        wal: /dev/disk/by-id/nvme-INTEL _ SSDPEDKE020T7 _
PHLE725100252P0IGN
        wal _ size: 2g
      /dev/disk/by-id/ata-HGST _ HUS724040ALE640 _ PK2331PAJ93B6T:
        db: /dev/disk/by-id/nvme-INTEL _ SSDPEDKE020T7 _
PHLE725100252P0IGN
```

```
        db_size: 50g
        format: bluestore
        wal: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7
PHLE725100252P0IGN
        wal_size: 2g
      /dev/disk/by-id/ata-HGST_HUS724040ALE640_PK2331PAJ94G0T:
        db: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        db_size: 50g
        format: bluestore
        wal: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        wal_size: 2g
      /dev/disk/by-id/ata-ST4000NM115-1YZ107_ZC11VNZ1:
        db: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        db_size: 50g
        format: bluestore
        wal: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        wal_size: 2g
      /dev/disk/by-id/ata-ST4000NM115-1YZ107_ZC11VP04:
        db: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        db_size: 50g
        format: bluestore
        wal: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        wal_size: 2g
      /dev/disk/by-id/ata-WDC_WD4000FYYZ-01UL1B0_WD-
WCC130368354:
        db: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        db_size: 50g
        format: bluestore
        wal: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        wal_size: 2g
      /dev/disk/by-id/scsi-35000cca22bddefac:
        db: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        db_size: 50g
        format: bluestore
        wal: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        wal_size: 2g
      /dev/disk/by-id/scsi-35000cca22be08431:
        db: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        db_size: 50g
        format: bluestore
        wal: /dev/disk/by-id/nvme-INTEL_SSDPEDKE020T7_
PHLE725100252P0IGN
        wal_size: 2g
```

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

Appendix C: Policy.cfg

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

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

# MON
role-mon/cluster/mon*.sls

# MGR (mgrs are usually colocated with mons)
role-mgr/cluster/mon*.sls

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

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

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

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

# openATTIC
role-openattic/cluster/salt*.sls

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

## Profiles
profile-nvme/cluster/*.sls
profile-nvme/stack/default/ceph/minions/*.yml
```

Appendix D: Network Switch Configuration

First, properly cable and configure each node on the switches. Ensuring proper switch configuration at the outset will prevent networking issues later. The key configuration items include ensuring the switches are properly stacked with an MLAG, creating LACP bonds, VLANS, and enabling jumbo frames. Each aggregation group needs a unique number, planned ahead of time. It is also recommended that you disable the spanning tree on the ports utilized for storage.

Configure the MLAG for stacking the switches as found here:

https://docs.cumulusnetworks.com/display/DOCS/Multi-Chassis+Link+Aggregation+-+MLAG

If enabling jumbo frame packets, follow the directions under MTU found here:

https://docs.cumulusnetworks.com/display/DOCS/Layer+1+and+Switch+Port+Attributes

The next step is to create the LACP bonds. Information on this step is found here:

https://docs.cumulusnetworks.com/display/DOCS/Bonding+-+Link+Aggregation

There are special requirements for using LACP over the stacked switches as described in the MLAG section under LACP and Dual-Connectedness

The final step is to add the interfaces to a bridge and enable VLANs. This can be found here:

https://docs.cumulusnetworks.com/display/DOCS/VLAN-aware+Bridge+Mode+for+Large-scale+Layer+2+Environments

SPECIAL NOTE:

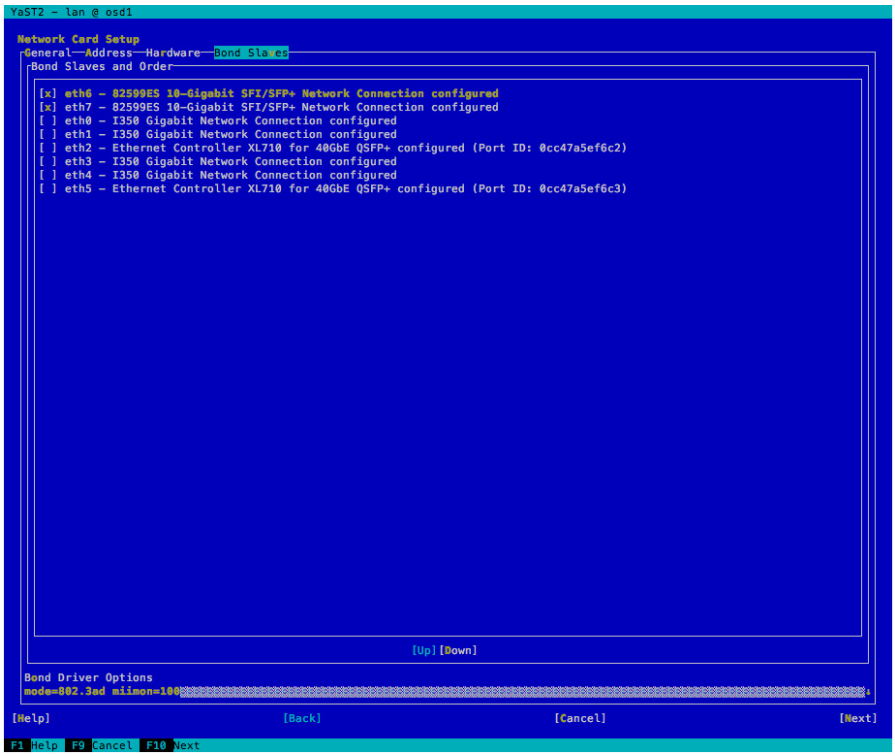
In some situations where the switch is operating at less than 100Gb, it may be necessary to force the ports on the switch to the appropriate speed. In the testing performed for this document, 40GbE adapters were used and thus the following command was issued for each interface in use:

```
link-speed 40000
```

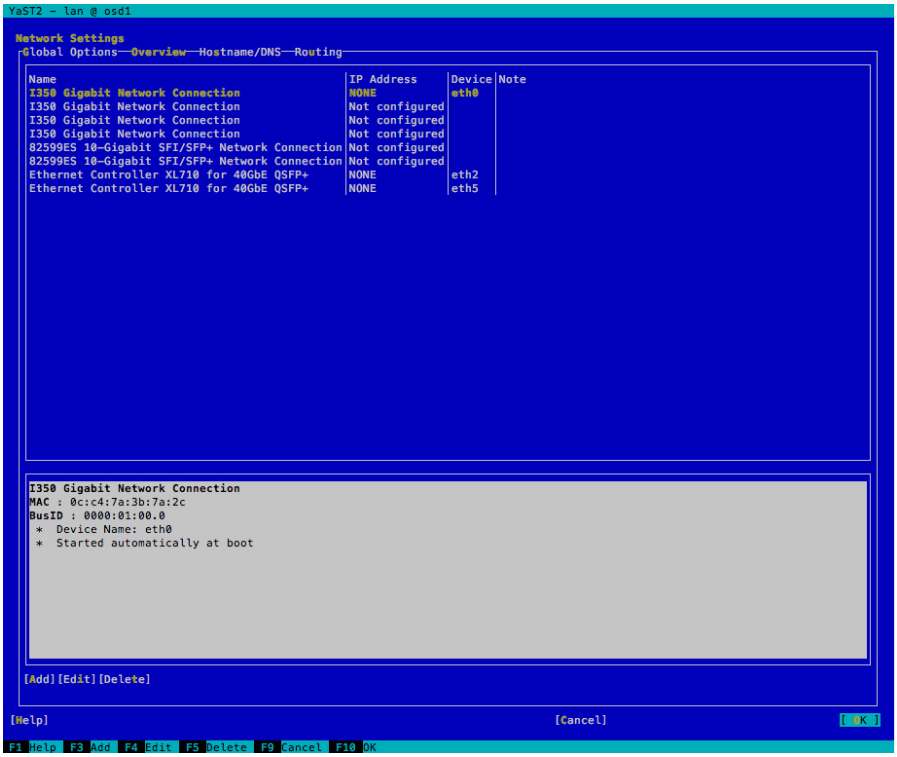
Appendix E: OS Networking Configuration

Perform the network configuration during installation.

```
Set the Intel XL710 to No Link
```

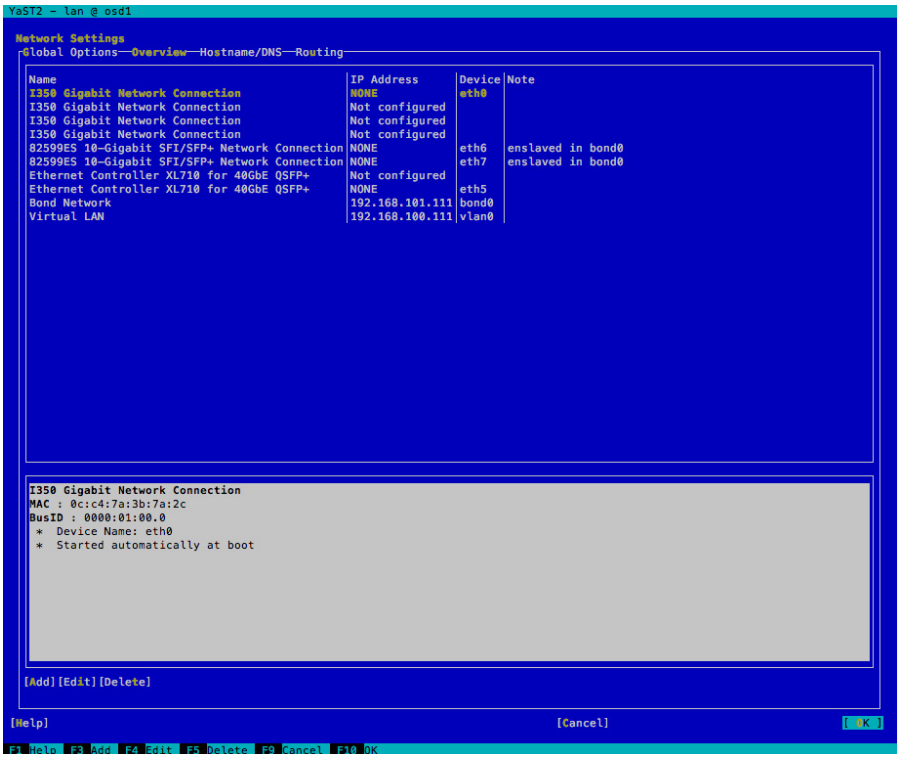


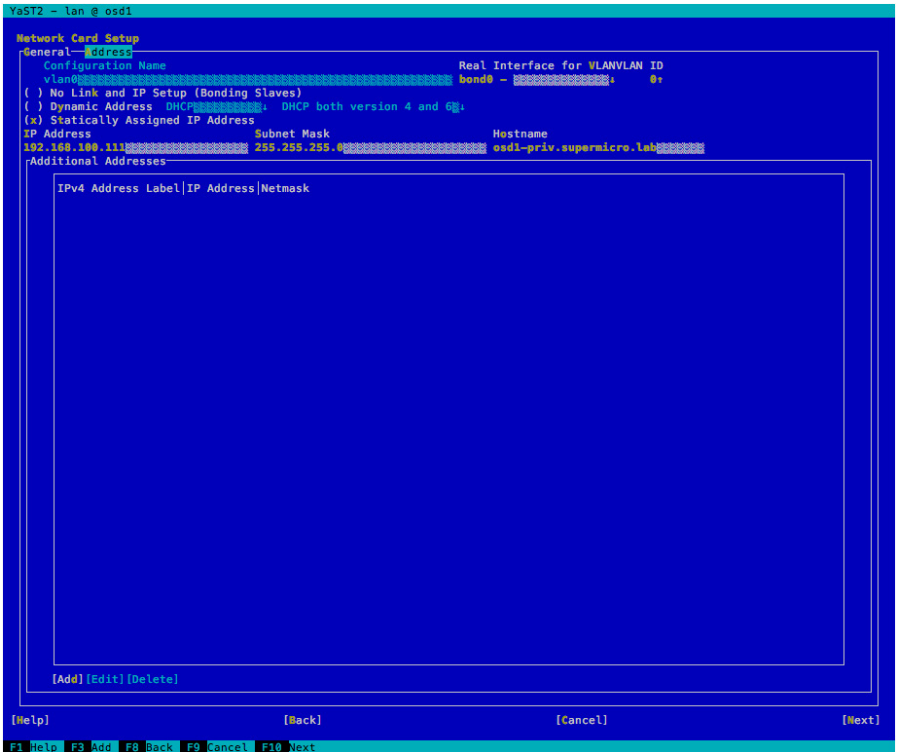
Add an interface of type bond, set it to the proper IP address for the untagged VLAN, and proceed to the Bond slaves page where the Intel XL710 interfaces should be selected and the mode set to 802.3ad



Add the VLAN interfaces, making sure to select the correct VLAN ID and setting the IP and host information.

This figure represents the proper network configuration for osdnode1 as configured in this paper.





Appendix F: 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/dmbyte/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.

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

Sequential Writes

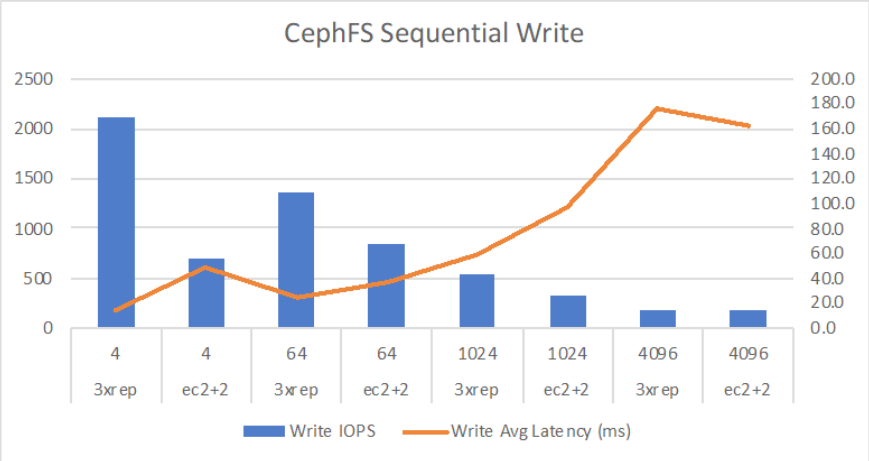
Sequential write I/O testing was performed across block sizes ranging from 4KiB to 4MiB. In the figures below, the x-axis labels indicate the block size in KiB on the top line and the data protection scheme on the bottom line. 3xrep is indicative of the Ceph standard 3 replica configuration for data protection while ec2+2 is Erasure Coded using the ISA plugin with k=2 and m=2. The Erasure Coding settings were selected to fit within the minimum cluster hardware size supported by SUSE.

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

CephFS

CephFS Sequential Write IOPS

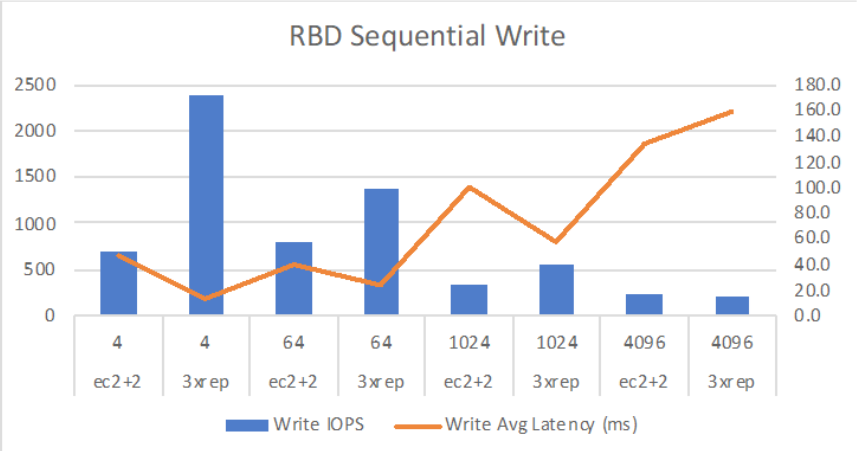
Data Protection	I/O size (KiB)	Write bandwidth (MiB/s)	Write IOPS	Write Avg Latency (ms)
3xrep	4	8	2125	15.0
ec2+2	4	2	693	48.1
3xrep	64	84	1351	23.9
ec2+2	64	52	841	37.2
3xrep	1024	543	544	58.6
ec2+2	1024	331	332	96.7
3xrep	4096	724	181	175.2
ec2+2	4096	751	188	162.9



RBD

RBD Sequential Writes

Data Protection	I/O size (KiB)	Write bandwidth (MiB/s)	Write IOPS	Write Avg Latency (ms)
ec2+2	4	2	688	47.4
3xrep	4	9	2377	13.6
ec2+2	64	50	803	40.0
3xrep	64	84	1355	23.9
ec2+2	1024	318	318	100.1
3xrep	1024	553	554	57.4
ec2+2	4096	951	238	133.0
3xrep	4096	793	198	159.3



Sequential Reads

The sequential read tests were conducted across the same range of block sizes as the write testing. Results were not published due to a bug in the testing process that generated incorrect results. Any future revision of the document will contain this data.

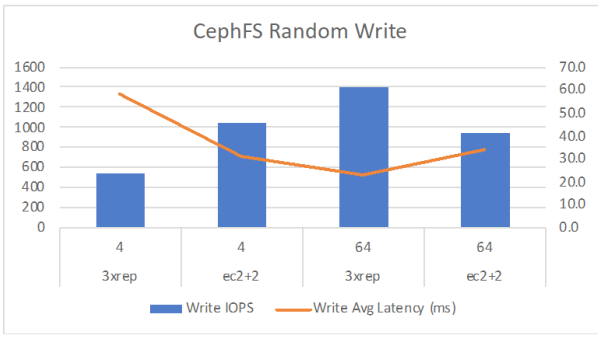
Random Writes

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.

CephFS

CephFS Random Write

Data Protection	I/O size (KiB)	Write bandwidth (MiB/s)	Write IOPS	Write Avg Latency (ms)
3xrep	4	2	547	58.4
ec2+2	4	4	1039	30.9
3xrep	64	87	1396	23.0
ec2+2	64	59	945	33.8

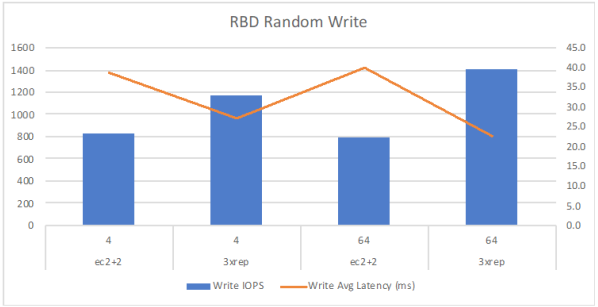


RBD

RBD Random Write

Data Protection	I/O size (KiB)	Write bandwidth (MiB/s)	Write IOPS	Write Avg Latency (ms)
ec2+2	4	3	830	38.8
3xrep	4	4	1168	27.3
ec2+2	64	49	798	40.2
3xrep	64	88	1409	22.8

RBD



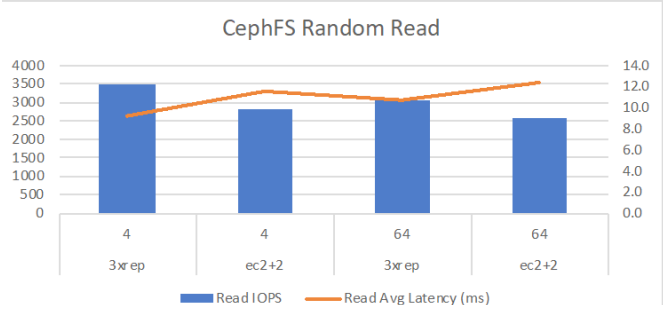
Random Reads

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

CephFS

CephFS Random Read

Data Protection	I/O size (KiB)	Read bandwidth (MiB/s)	Read IOPS	Read Avg Latency (ms)
3xrep	4	13	3471	9.2
ec2+2	4	10	2796	11.5
3xrep	64	189	3037	10.6
ec2+2	64	161	2590	12.4

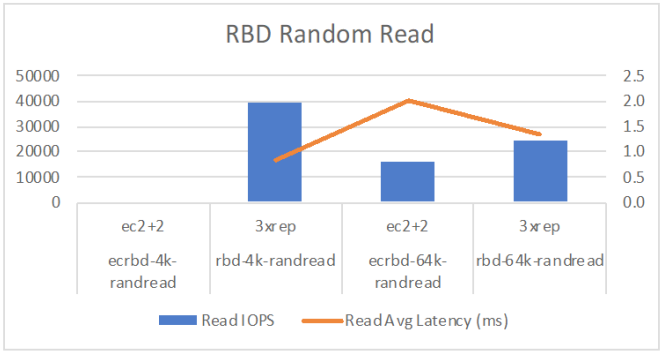


RBD

The RBD random read results are missing for the 4K erasure coded target due to an error during testing.

RBD Random Read

Data Protection	I/O size (KiB)	Read bandwidth (MiB/s)	Read IOPS	Read Avg Latency (ms)
ec2+2	4			
3xrep	4	152	39093	0.8
ec2+2	64	1011	16178	2.0
3xrep	64	1555	24890	1.3



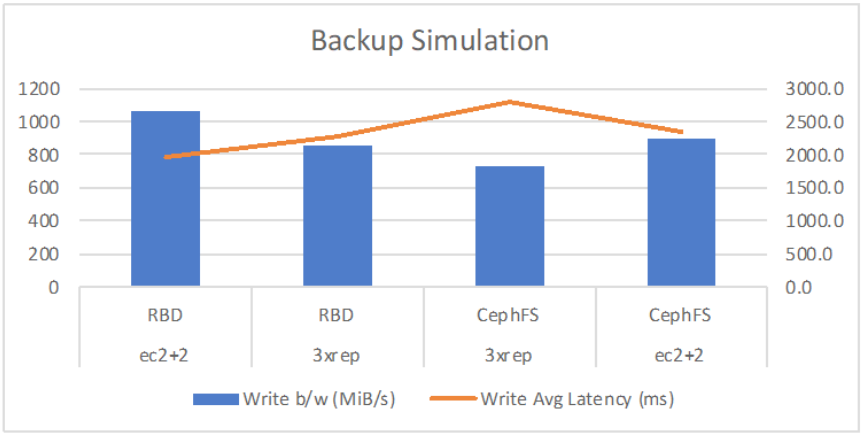
Workload Simulations

The following test results are workload oriented.

Backup

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.

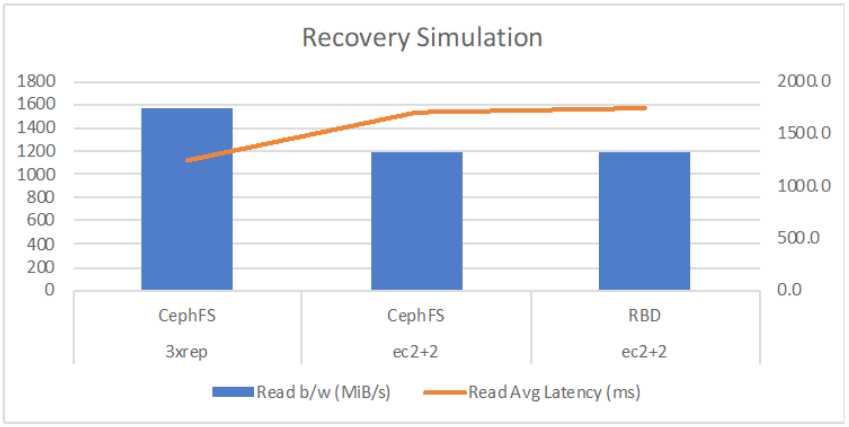
Data Protection	Protocol	Backup Simulation Write bandwidth (MiB/s)	Write IOPS	Write Avg Latency (ms)
ec2+2	RBD	1058	17	1949.8
3xrep	RBD	854	13	2289.1
3xrep	CephFS	728	11	2805.8
ec2+2	CephFS	901	14	2337.4



Recovery

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

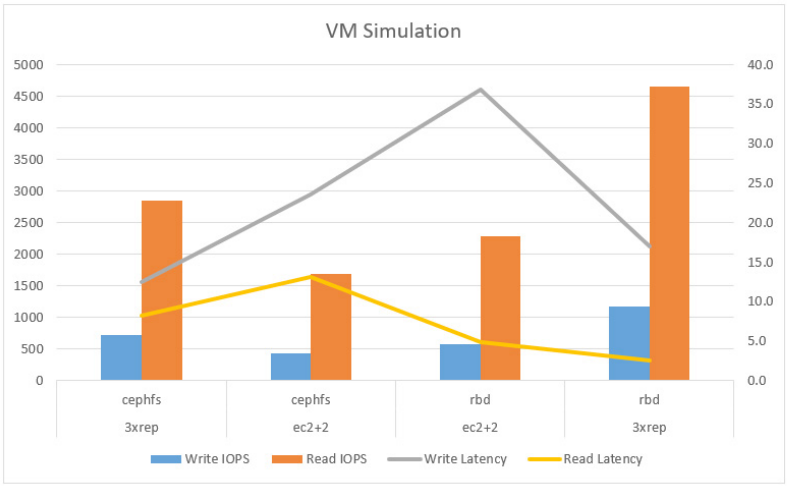
Recovery Simulation				
Data Protection	**Protocol**	**Read bandwidth (MiB/s)**	**Read IOPS**	**Read Avg Latency (ms)**
3xrep	CephFS	1570	25	1248.6
ec2+2	CephFS	1189	19	1695.9
ec2+2	RBD	1191	19	1751.8



KVM Virtual Guest

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.

VM Simulation						
Write b/w (MiB/s)	Write IOPS	Write Avg Latency (ms)	Test Description	Read Avg Latency (ms)	Read IOPS	Read b/w (MiB/s)
2	718	12.4	KVM on 3xrep CephFS	8.2	2842	11
1	421	23.5	KVM on EC CephFS	13.1	1689	6
2	572	36.9	KVM on EC RBD	4.9	2283	8
4	1167	17.0	KVM on 3xrep RBD	2.5	4655	18



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.

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 b/w (MiB/s)	Write IOPS	Write Avg Latency (ms)	Write Max Latency (ms)	Test Description	Read Avg Latency (ms)	Read IOPS	Read b/w (MiB/s)
4	640	44.6	1293	CephFS EC	17.4	157	1
14	1916	16.3	744	CephFS 3xrep	2.0	476	3
4	591	53.1	2847	RBD EC	15.9	146	1
14	1806	17.6	1167	RBD 3xrep	2.9	448	3

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 b/w (MiB/s)	Write IOPS	Write Avg Latency (ms)	Write Max Latency (ms)	Test Description	Read Avg Latency (ms)	Read IOPS	Read b/w (MiB/s)
2	380	24.6	3818	CephFS EC	14.8	1526	11
5	670	15.4	757	CephFS 3xrep	8.2	2654	20
4	562	40.0	3387	RBD EC	4.3	2232	17
14	1819	12.6	6030	RBD 3xrep	2.0	7277	56

Resources

- SUSE Enterprise Storage Technical Overview

  https://www.suse.com/docrep/documents/1mdg7eq2kz/suse_enterprise_storage_technical_overview_wp.pdf

- SUSE Enterprise Storage v5 - Administration Guide

  https://www.suse.com/documentation/suse-enterprise-storage-5/book_storage_admin/data/book_storage_admin.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

About Super Micro Computer, Inc.

Supermicro® (NASDAQ: SMCI), the leading innovator in high-performance, high-efficiency server technology is a premier provider of advanced server Building Block Solutions® for Data Center, Cloud Computing, Enterprise IT, Hadoop/Big Data, HPC and Embedded Systems worldwide. Supermicro is committed to protecting the environment through its "We Keep IT Green®" initiative and provides customers with the most energy-efficient, environmentally-friendly solutions available on the market.

Learn more on www.supermicro.com

Printed in USA ♻ Please Recycle 14_SSG-SUSE-Storage-Guide_180816_Rev3