SUSE Enterprise Storage 5.5 with HPE ProLiant

By Bryan Gartner, SUSE
bryan.gartner@suse.com

Table of Contents

1. Executive Summary

This reference configuration is intended to help organizations plan and install a Ceph-based, software-defined storage infrastructure.

For most enterprise-level businesses, the demand for data storage is growing much faster than the rate at which the price for storage is shrinking. As a result, you could be forced to increase your budget dramatically to keep up with data demands. This intelligent, software-defined storage solution—powered by SUSE Enterprise Storage technology and HPE ProLiant system hardware—enables you to transform your enterprise storage infrastructure to reduce costs, while providing unlimited scalability to keep up with your future demands. With this completely tested approach, you will have the confidence to deploy a working solution in an agile manner and be able to maintain and scale it over time, without capacity-based increases in software subscriptions.

2. Target Audience

This document is intended for IT decision makers, architects, system administrators and technicians who are implementing the HPE ProLiant platform and need a flexible, software-defined storage solution—such as SUSE Enterprise Storage—that can provide multiple protocol access. To ensure optimum results, you should have a solid understanding of your storage use cases, along with sizing/characterization concepts and limitations within your environment.

3. Solution Overview

The coupling of an industry-leading infrastructure platform such as HPE ProLiant and a software-defined storage solution such as SUSE Enterprise Storage provides an incredibly powerful and flexible combination. Nodes and storage capacity can be quickly added, replaced or substituted over time as demands dictate. Use cases for such a flexible solution range from dynamically allocated storage pools for physical, virtualized or containerized environments, to a custom testing and development solution, to more specific use cases, such as integrations with:

- HPE Helion OpenStack
- SUSE OpenStack Cloud
- SUSE CaaS Platform
- SUSE Cloud Application Platform

4. Solution Components

4.1 Facility

While beyond the scope of this document, the heating, ventilation and air conditioning (HVAC) requirements of hosting such an infrastructure solution should be carefully considered and planned. To aid in determining the power requirements for system deployment, use the HPE Power Advisor (available online or as a downloadable application). Using this tool, you can plan the needs for your solution and order the correct Power Distribution Unit (PDU) to account for the local power conditions and connections in the final installation location.

4.2 Network

Networking components and their associated services typically require the most advanced planning. Connectivity, capacity and bandwidth requirements for a software-defined storage infrastructure have a fair amount of complexity, especially within the context of an existing IT infrastructure.

With the range of available network interface card options for HPE ProLiant, the baseline network bandwidth can be easily provided for each resource node and collectively within the software-defined storage deployment. For improved resiliency and performance, it is also possible to bond multiple network interfaces together. While beyond the scope of this document, the only remaining consideration is for capacity and bandwidth for the interconnecting top-of-rack switches.

Note: Given the available options for HPE ProLiant with 10, 25 or 100 GbE speeds, you are encouraged to utilize those with higher speed ones to help address the access of many clients and to deal with the back-end replication and recovery aspects, especially as the usage and scale increases.

4.3 Computing

***HPE ProLiant Servers*.** These servers are the most flexible, reliable and performance-optimized ProLiant rack servers ever. HPE continues to provide industry-leading compute innovations. The new HPE ProLiant rack portfolio—with flexible choices and versatile design, along with improved energy efficiencies—ultimately lowers your TCO. Integrated with a simplified (yet comprehensive) management suite and industry-leading support, the ProLiant rack portfolio delivers a more reliable, fast and secure infrastructure solution; helps to increase IT staff productivity; and accelerates service delivery. In addition, the rack portfolio is performance-optimized for multi-application workloads to significantly increase the speed of IT operations and enable IT to respond faster to business needs of any size.

HPE ProLiant DL360 Compute. In a form factor that provides performance-drive density, this machine delivers security, agility and flexibility without compromise. It supports the Intel® Xeon® Scalable processor with up to a 60% performance gain and 27% increase in cores, along with 2933 MT/s HPE DDR4 SmartMemory, supporting up to 3.0 TB with an increase in performance of up to 82%. With the added performance that persistent memory, HPE NVDIMMs and 10 NVMe bring, the HPE ProLiant DL360 means business. You can deploy, update, monitor and maintain it with ease by automating essential server lifecycle management tasks with HPE OneView and HPE Integrated Lights Out 5 (iLO 5). In addition, you can deploy this 2P secure platform for diverse workloads in space-constrained environments.

For this implementation, the DL360 and DL380 models of HPE ProLiant were used for the various cluster roles, as defined in the next section and detailed in the bill of materials referenced in the Resources and additional links section.

Tip: Any SUSE YES certified HPE platform can be used for the physical nodes of this deployment, as long as the certification refers to the major version of the underlying SUSE operating system required by the SUSE Enterprise Storage release.

4.4 Storage

Each of the resource nodes is also expected to have local, direct-attached storage, which is used for the node's operating system. For this deployment, a pair of disk drives is configured as a RAID1 volume.

For the OSD nodes, configure the remaining drives as RAID0 LUNs (or HBA mode on the controller). Configure the number and type of these accessible drives to provide the software-defined storage capacity and functions, such as BlueStore's WAL and DB, cache and data drives. Follow the recommended number, type, capacity and ratios from the SUSE Enterprise Storage documentation.

4.5 Software

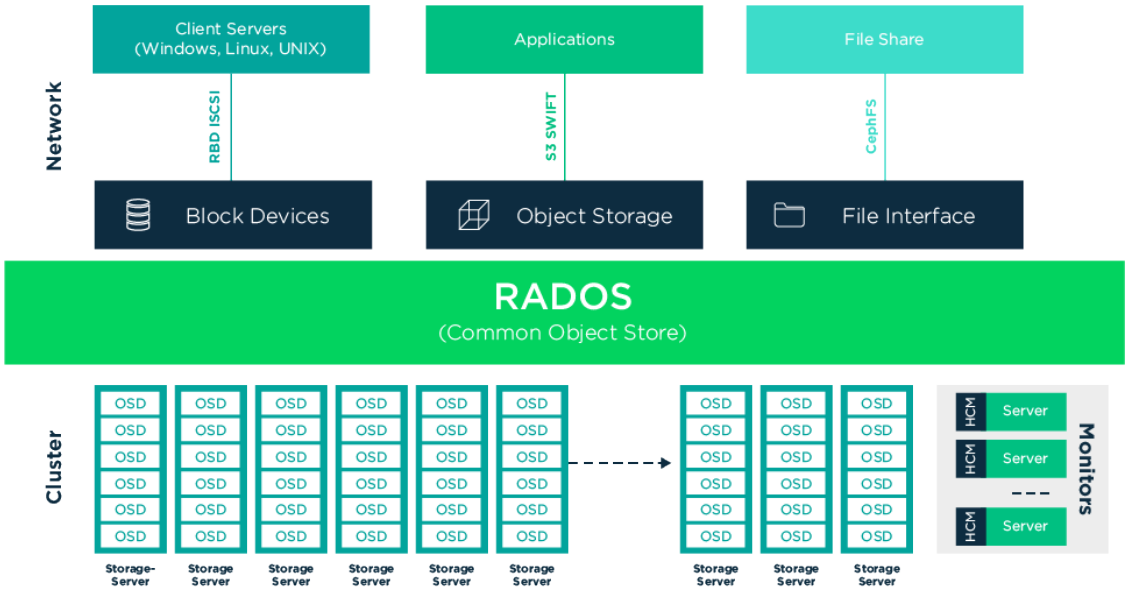
SUSE Enterprise Storage is an intelligent software-defined storage solution, powered by Ceph technology, which enables you to transform your enterprise storage infrastructure. While a software-defined approach might seem new, the logic within enterprise storage devices has always been written in software. It has only been in the last few years that hardware has progressed enough that enterprise storage software and dedicated hardware can now be separated. This provides IT organizations with a simple-to-manage, agile infrastructure approach—yielding increased speed of delivery, durability and reliability. This helps to accelerate innovation, reduce costs and alleviate proprietary hardware lock-in by transforming your enterprise storage infrastructure with a truly open and unified, intelligent software-defined storage solution.

Our single, truly unified, software-defined storage cluster provides applications with object, block and file system storage. It offers ubiquitous and universal access to your legacy and modern applications, as well as automated durability for your data, with high availability and disaster recovery options.

Key features include:

- Unlimited scalability with a distributed storage cluster designed to scale to thousands of nodes and multi-hundred petabyte environments (and beyond) to meet your growing data requirements.
- A highly redundant storage infrastructure design maximizes application availability with no single points of failure.
- Self-healing capabilities minimize storage administration involvement and optimize data placement, enabling rapid reconstruction of redundancy and maximizing system resiliency and availability.
- Utilize commodity, off-the-shelf hardware that is at minimum 30 percent less expensive than average capacity optimized solutions to drive significant CAPEX savings.
- Automated re-balancing and optimized data placement with an easy-to-manage intelligent solution that continuously monitors data utilization and infrastructure—without any manual intervention and without growing IT staff.
- Included with the solution are the following supported protocols:
    - Native
        - RBD (Block)
        - RADOS (Object)
        - CephFS (with multiple active MDS Servers)
        - S3 & SwiftRBD (Block)
    - Traditional
        - iSCSI
        - NFS
        - CIFS/SMB

As noted above, 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 might be considered a limiting factor, the ISCSI and S3 gateways can scale horizontally using load balancing techniques.



Ceph Architecture

Cluster networking. The network environment where you intend to run Ceph should ideally be a bonded set of at least two network interfaces that are logically split into a public part and a trusted internal part, using VLANs. The bonding mode is recommended to be 802.3ad, if possible, to provide maximum bandwidth and resiliency. The public VLAN provides the service to the client nodes, while the internal part provides for the authenticated Ceph network communication. The main reason for this is that although Ceph provides authentication and protection against attacks, once the secret keys are in place, they can become vulnerable if the messages used to configure them are transferred openly.

Administration Node. The Administration Node is a central point of the Ceph cluster, because it manages the rest of the cluster nodes by querying and instructing their Salt minion services. It usually includes other services as well—for example, the Ceph Dashboard (formerly known as openATTIC) web interface with the Grafana dashboard backed by the Prometheus monitoring toolkit.

Ceph Monitor. Ceph Monitor (often abbreviated as MON) nodes maintain information about the cluster health state, a map of all nodes and data distribution rules. If failures or conflicts occur, the Ceph Monitor nodes in the cluster decide by majority which information is correct. To form a qualified majority, it is recommended to have an odd number of Ceph Monitor nodes, starting with at least three.

Ceph Manager. The Ceph manager (MGR) collects the state information from the whole cluster. The Ceph manager daemon runs alongside the monitor daemons. It provides additional monitoring, and interfaces the external monitoring and management systems. The Ceph manager requires no additional configuration, beyond ensuring that it is running.

Ceph OSD. A Ceph OSD is a daemon that handles Object Storage Devices, which are physical or logical storage units (hard disks or partitions). Object Storage Devices can be physical disks/partitions or logical volumes. The daemon additionally takes care of data replication and rebalancing, in case of added or removed nodes. Ceph OSD daemons communicate with monitor daemons and provide them with the state of the other OSD daemons.

Optional Roles

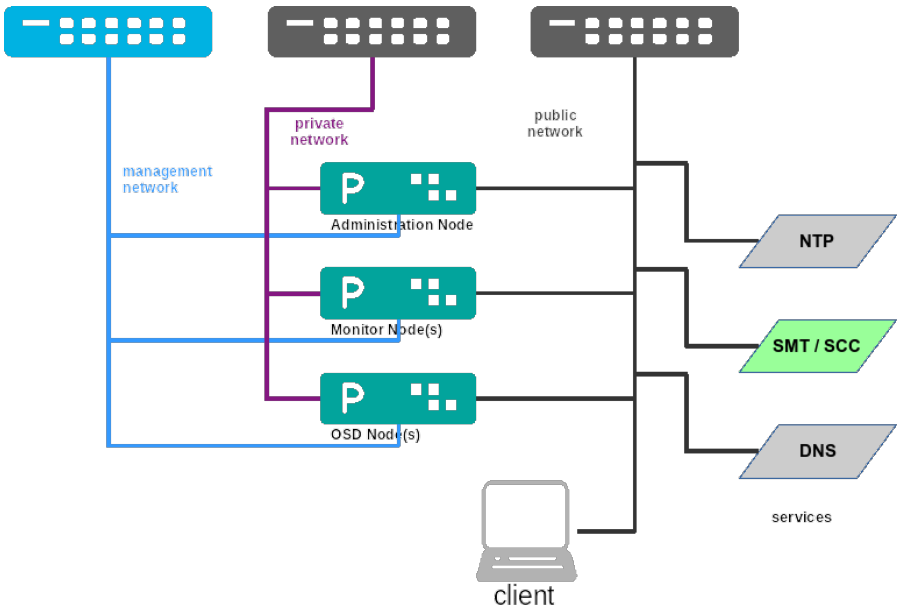
- ***Metadata Server.*** The metadata servers (MDS) store metadata for the CephFS. By using an MDS, you can execute basic file system commands, such as ls, without overloading the cluster.
- ***Object Gateway.*** The Ceph Object Gateway provided by Object Gateway is an HTTP REST gateway for the RADOS object store. It is compatible with OpenStack Swift and Amazon S3 and has its own user management.
- ***NFS Ganesha.*** NFS Ganesha provides an NFS access to either the Object Gateway or the CephFS. It runs in the user instead of the kernel space and directly interacts with the Object Gateway or CephFS.
- ***iSCSI Gateway.*** iSCSI is a storage network protocol that enables clients to send SCSI commands to SCSI storage devices (targets) on remote servers.

Additional Network Infrastructure Components/Services

- Domain Name Service (DNS): An external network-accessible service to map IP Addresses to hostnames for all cluster resource nodes.
- Network Time Protocol (NTP): An external network-accessible service to obtain and synchronize system times to aid in timestamp consistency. It is recommended to point all resource nodes to a physical system/device that provides this service.
- Software Update Service: Access to a network-based repository for software update packages. This can be accessed directly from each node via registration to the SUSE Customer Center (SCC) or from local servers running a SUSE Subscription Management Tool (SMT) instance. As each node is deployed, it can be pointed to the respective update service; then update notification and applications will be managed by the configuration management web interface.

5. Solution Details

This document focuses on a new, basic SUSE Enterprise Storage deployment on the HPE ProLiant platform, which can be scaled over time. More physical nodes can also be added to augment the cluster's functionality and capacity or to replace some of the initial resource nodes. To provide a production-ready cluster and to take advantage of the HPE ProLiant platform, the following figure shows the target logical cluster deployment:



Deployment Logical View

5.1 Deployment Flow

This section is meant as a companion guide to the official network, system and software product deployment documentation, citing specific settings as needed for this reference implementation. Default settings are assumed to be in use, unless otherwise cited to accomplish the respective best practices and design decisions herein.

Given the detailed information contained in the SUSE Enterprise Storage Deployment Guide, only the following additional, incremental configurations and modifications are described below.

Pre-Installation Checklist

Start by collecting the necessary install media and validating the inventory of systems to be deployed.

- Obtain the following software media and documentation artifacts:
  - From the SUSE site, download the install media and product media as noted below:
    - The SUSE Enterprise Storage x86_64 install media (DVD1)
    - The corresponding SUSE Linux Enterprise Server 12-SP3 x86_64 install media (DVD1)

Note: Utilize either trial or purchased subscriptions for all the resource nodes to ensure access to support and software updates. The bill of materials in the Resources and additional links section outlines the type and quantity of subscriptions needed.

- Obtain and preview the SUSE Enterprise Storage documentation, focusing on these documents:
  - Release Notes
  - Deployment Guide
  - Administration Guide
- Validate that the necessary CPU, memory and interconnect quantity and type are present for each node and intended role. Refer to the minimum CPU/Memory/Disk/Networking requirements as noted in the SUSE Enterprise Storage Deployment Guide.
- Ensure that a pair of local, direct attached disk drives is present on each node (SSDs are preferred); these will become the target for the operating system installation.
- Network: Prepare an IP addressing scheme and create both a storage cluster public and private network, along with the desired subnets and VLAN designations.
- Boot Settings: Manage the boot node and select UEFI mode, with the primary device being hard disk.
- BIOS/uEFI settings are reset to defaults for a known baseline, consistent state, or perhaps with desired localized values.
  - Use consistent and up-to-date versions for BIOS/uEFI/device firmware to reduce potential troubleshooting issues later.

Resource Node Installation

Install the SUSE Linux Enterprise Server operating system on each node type, starting with the Administration Server, then the Monitor Nodes and finally the OSD Nodes.

- Include only the minimal pattern and components, according to the procedure from deployment. This can be accomplished in any number of ways, such as with the virtual media option through HPE iLO or from a PXE network-boot environment.
- Use the suggested, default-partitioning scheme on each node and validate that the target LUN for the operating system installation corresponds to the RAID1 pair of local disk drives.
- After the operating system installation is complete across all the nodes, perform the following checks:

- Ensure that each node has access to the necessary software repositories, for later operations and updates. It is suggested that you apply all software updates, via zypper up.
- NTP is configured and operational, synchronizing with a source outside the cluster via ntpq -pn.
- DNS is configured and operational, referencing a source outside the cluster.
- If necessary, adjust the udev rules to ensure that network interfaces are identified (as needed) in the same logical order across the systems, to make later steps easier. Ensure that the respective network interfaces are bonded together with the associated VLANs configured. While configuring these interfaces, it is also recommended to disable IPv6 functionality and the firewall on each node.

Note: For environments that require firewalls to be in place, refer to the "Cluster Deployment" section in the SUSE Enterprise Storage Deployment Guide.

Cluster Deployment

Follow the steps noted in "Cluster Deployment" section of the SUSE Enterprise Storage Deployment Guide. You are strongly encouraged to utilize the "DeepSea" approach, which saves the administrator time and helps to confidently perform complex operations on a Ceph cluster in a staged fashion:

- Complete Stage 0 (preparation) and Stage 1 (discovery).
- Before executing Stage 2:
  - You might want to take advantage of SSD devices for BlueStore WAL/DB (if available) to help improve the overall performance of your cluster. This can be accomplished from the command line on the Administration Node via:

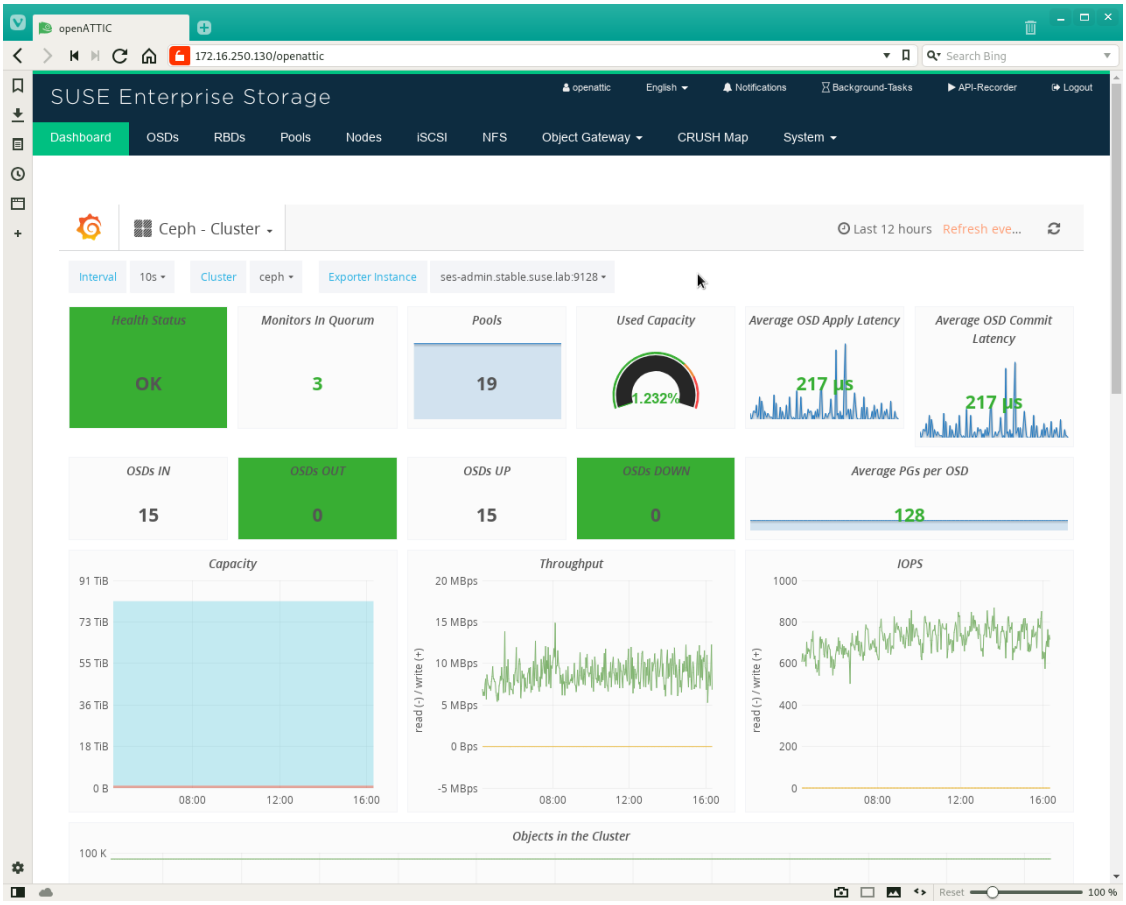
```
# salt-run proposal.populate \
    name=proliant ratio=6 target='cephosd*' format=bluestore \
    wal-size=2g db-size=50g db=745 wal=745 ssd=spinner=True data=279
```

Note: The size values might need adjustment to effectively use the drive capacity and quantity in your configuration.

- Create a policy.cfg file that references the /srv/pillar/ceph/proposals/profile-proliant contents generated above.
- At this point, you should also review and validate the /srv/pillar/ceph/proposals/profile-proliant/stack/default/ceph/minions/*.yml content for the OSD nodes to ensure that the correct disk designations are in place.
  - Then execute Stage 2 (configuration), Stage 3 (deployment) and Stage 4 (services) to complete the deployment.
  - Upon successful completion of the previous stages, you can check the cluster status via:

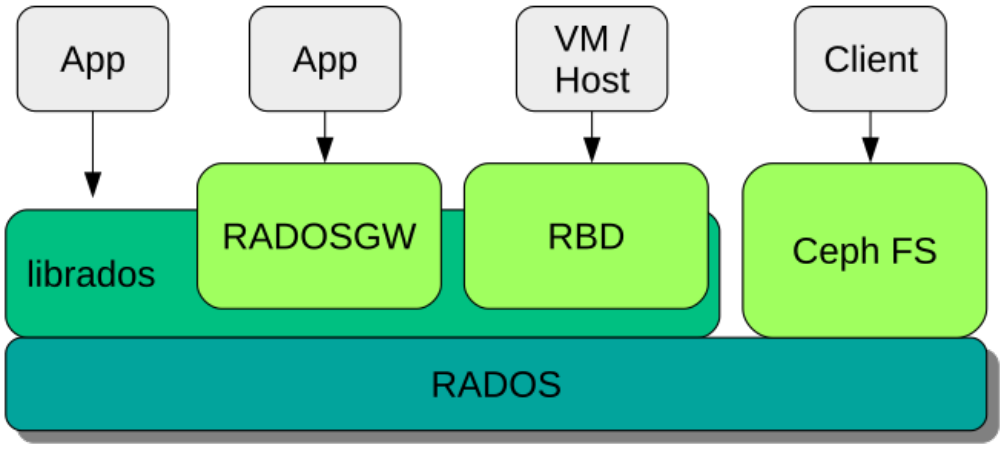
```
ceph health
ceph status
```

Access the deployed Ceph Dashboard (openATTIC) web interface to view the management and monitoring dashboard.



Ceph Dashboard

To complete the solution and implement a common use case, this deployment can provide guest virtual machines, running on a Linux-based KVM hypervisor host with access to block devices for their root filesystem. This is accomplished through the libvirtd interaction of Linux/KVM and librbd to the storage cluster, as shown in the following figure:



Interfaces to the Ceph Object Store

- At this point, you are ready to begin creating the block device and the virtual machines that use this storage service. Refer to the "Integration with Virtualization Tools" section of the Administration Guide.

- With this setup configuration completed, you can install virtual machines, using the RBD functionality of SUSE Enterprise Storage to provide inherent resiliency to their operating system volumes. Even shutting down or experiencing a hardware failure on one of the Ceph nodes will not affect the running virtual machine.

6. Additional Considerations

To understand the administration aspects of the cluster, review these sections of the SUSE Enterprise Storage Administration Guide:

- Changing or scaling the cluster node count by adding and removing resource nodes
- Operating the cluster and managing the storage resources and accessing the data
- Monitoring and managing the services
- Troubleshooting hints and tips

Additionally, this solution can be used for multiple access methods (object, block, file) and accessed by clients over many different protocols (iSCSI, S3, NFS, SMB). Each of these can be incrementally added over time, using the SUSE Enterprise Storage deployment framework and adding the respective gateway roles and nodes.

7. Conclusion

After reviewing and working through the steps described in this document, you should have a working software-defined storage platform that is scalable through the addition of more resource nodes, as needed. SUSE Enterprise Storage provides a complete suite of the necessary software and processes, which leverage the feature set of HPE ProLiant to create a production-ready and agile platform.

8. Resources and Additional Links

HPE ProLiant

- HPE ProLiant Rack Servers

Bill of Materials – HPE ProLiant DL360 Systems

Role	Quantity	Product Number	Product Name	Description
Admin	1	DL360 Gen9	HPE ProLiant DL360 Gen9	HPE ProLiant DL360 Gen9 4LFF Configure-to-order Server HPE Dynamic Smart Array B140i, 4-Bay LFF Drive Cage
	1	755382-L21	Processor	Intel Xeon E5-2620v3 (2.4GHz/6-core/15MB/85W)
	4	726718-B21	Memory	8GB (1x8GB) Single Rank x4 DDR4-2133 CAS-15-15-15 Registered Memory Kit
	1	665243-B21	Flexible LOM	HPE Ethernet 10Gb 2-port 560FLR-SFP+ FIO Adapter

	2	720478-B21	Power Supplies	HPE 500W Flex Slot Platinum Hot Plug Power Supply Kit
	1	749974-B21	Storage Controller	HPE Smart Array P440ar/2GB FBWC 12Gb 2-ports Int FIO SAS Controller
	1	766211-B21	Storage Cable	HPE DL360 Gen9 LFF P440ar/H240ar SAS Cbl
	2	652757-B21	Hard Drives	HPE 2TB 6G SAS 7.2K rpm LFF (3.5-inch) SC Midline 1yr Warranty Hard Drive
	1	663202-B21	Rail Kit	HPE 1U Large Form Factor Ball Bearing Rail Kit
	1	663203-B21	Cable Management Arm	HPE 1U Cable Management Arm for Ball Bearing Rail Kit
Monitor	1	755259-B21	ProLiant DL360 Gen9	HPE ProLiant DL360 Gen9 4LFF Configure-to-order Server HPE Dynamic Smart Array B140i, 4-Bay LFF Drive Cage
	1	755382-L21	Processor 1	Intel® Xeon® E5-2690v3 (2.6GHz/12-core/30MB/135W) FIO Processor Kit
	1	755382-B21	Processor 2	Intel® Xeon® E5-2690v3 (2.6GHz/12-core/30MB/135W) FIO Processor Kit
	4	726718-B21	Memory for 1st processor	8GB (1x8GB) Single Rank x4 DDR4-2133 CAS-15-15-15 Registered Memory Kit
	4	726718-B21	Memory for 2nd processor	8GB (1x8GB) Single Rank x4 DDR4-2133 CAS-15-15-15 Registered Memory Kit
	1	665243-B21	FlexibleLOM	HPE Ethernet 10Gb 2-port 560FLR-SFP+ FIO Adapter
	2	720478-B21	Power Supplies	HPE 800W Flex Slot Platinum Hot Plug Power Supply Kit

	2	765424-B21	Hard Drives	HPE 600GB 12G SAS 15K rpm LFF (3.5-inch) SC Enterprise 3yr Warranty Hard Drive
	1	749974-B21	Storage Controller	HPE Smart Array P440ar/2GB FBWC 12Gb 2-ports Int FIO SAS Controller
	1	766211-B21	Storage Cable	HPE DL360 Gen9 LFF P440ar/H240ar SAS Cbl
	1	663202-B21	Rail Kit	HPE 1U Large Form Factor Ball Bearing Rail Kit
	1	663203-B21	Cable Management Arm	HPE 1U Cable Management Arm for Ball Bearing Rail Kit
OSD	1	719061-B21	ProLiant DL380 Gen9	HP ProLiant DL380 Gen9 12LFF Configure-to-order Server HP Dynamic Smart Array B140i 12-Bay LFF Drive Cage
	1	719044-L21	Processor 1	Intel® Xeon® E5-2690v3 (2.6GHz/12-core/30MB/135W) FIO Processor Kit
	1	719044-B21	Processor 2	Intel® Xeon® E5-2690v3 (2.6GHz/12-core/30MB/135W) FIO Processor Kit
	8	726718-B21	Memory for 1st processor	8GB (1x8GB) Single Rank x4 DDR4-2133 CAS-15-15-15 Registered Memory Kit
	8	726718-B21	Memory for 2nd processor	8GB (1x8GB) Single Rank x4 DDR4-2133 CAS-15-15-15 Registered Memory Kit
	1	665243-B21	FlexibleLOM	HPE Ethernet 10Gb 2-port 560FLR-SFP+ FIO Adapter
	2	720479-B21	Power Supply	HPE 800W Flex Slot Platinum Hot Plug Power Supply Kit
	1	724864-B21	Additional Bay	HP DL380 Gen9 2SFF Bay Kit
	2	749974-B21	Storage Controller	HP 600GB 12G SAS 15K rpm LFF (3.5-inch) SC Enterprise 3yr Warranty Hard Drive

	10	761477-B21	Rail Kit	HP 6TB 6G SAS 7.2k rpm LFF (3.5-inch) SC Midline 1yr Warranty Hard Drive
	1	761874-B21	Cable Management Arm	HP Smart Array P840/4G FIO Controller
	1	785991-B21	Storage Cable	HP DL380 Gen9 12LFF SAS Cable Kit
	1	783007-B21	Storage Cable	HP DL380 Gen9 P840/440 SAS Cable Kit
	1	665249-B21	Network	HP Ethernet 10Gb 2-port 560SFP+ Adapter
	1	720864-B21	Rail Kit	HP 2U Large Form Factor Ball Bearing Rail Kit
	1	720865-B21	Cable Management Arm	HP 2U Cable Management Arm for Ball Bearing Rail Kit

Note: Each of the above node roles is detailed as singular configurations so that it is easier to figure out how to order the appropriate quantity for the desired cluster.

SUSE Enterprise Storage

- Website
- Product Documentation

Bill of Materials – SUSE Enterprise Storage Software FixMe

Role	Quantity	Product Number	Description	Notes
Software	1	P9P49AAE	SUSE Enterprise Storage Base Configuration, x86-64, 4 OSD Nodes with 1-2 Sockets, L3-Priority Subscription, 3 Year	Includes 4 OSD Nodes plus 6 infrastructure nodes (e.g., 1-Admin, 3-MON, 2-gateway)
	1	P9P50AAE	SUSE Enterprise Storage Expansion Node, x86-64, 1 OSD Node with 1-2 Sockets, L3-Priority Subscription, 3 Year	For scaling, includes 1 additional OSD Node plus 1 infrastructure node

Note: With the Base Configuration subscription, two more resource nodes can be added to the documented eight-node cluster, to potentially provide other protocol gateways.