SUSE Enterprise Storage on HPE Apollo 4200/4500 System Servers

Choosing HPE density-optimized servers as
SUSE Enterprise Storage building blocks

Technical white paper

Contents

Executive summary

Traditional file and block storage architectures are being challenged by the explosive growth of data, fueled by the expansion of Big Data, unstructured data, and the pervasiveness of mobile devices. Emerging storage architectures like object storage can help businesses deal with these trends, providing cost-effective storage solutions that keep up with capacity growth while providing service-level agreements (SLAs) to meet business and customer requirements.

Enterprise-class storage subsystems are designed to address storage requirements for business-critical transactional data latencies. However, they may not be an optimal solution for unstructured data and backup/archival storage. In these cases, enterprise-class reliability is still required, but massive scale-out capacity and lower investment drive solution requirements.

Object storage software solutions are designed to run on industry-standard server platforms, offering lower infrastructure costs and scalability beyond the capacity points of typical file server storage subsystems. HPE Apollo 4000 series hardware provides a comprehensive and cost-effective storage capacity building block for these solutions.

When considering an open source-based solution, most enterprise environments will require a strong support organization and a vision to match or exceed the capabilities and functionality they currently experience with their traditional storage infrastructure. Using SUSE Enterprise Storage fills both of these needs with a world-class support organization and a leadership position within the Ceph community. SUSE Enterprise Storage helps ensure customers are able to deploy and operate their storage clusters as backing stores for cloud and applications with object storage interfaces with confidence.

HPE hardware combined with SUSE Enterprise Storage delivers an open source object storage solution that:

- Has software that offers practical scaling from one petabyte to well beyond a hundred petabytes of data
- Lowers upfront solution investment and total cost of ownership (TCO) per gigabyte
- Provides a single software-defined storage (SDS) cluster for both object and low to mid-range performance block storage
- Uses open source, minimizing concerns about proprietary software vendor lock-in
- Provides a better TCO for operating and maintaining the hardware than ‘white box’ servers
- Can be configured to offer low-cost, low-performance block storage in addition to object storage

HPE hardware gives you the flexibility to choose the configuration building blocks that are right for your business needs. The HPE Apollo 4510 Gen9, Apollo 4530 Gen9, and Apollo 4200 Systems are most suited for the task and allow you to find the right balance between performance, cost-per-gigabyte, building block size, and failure domain size.

Target audience

CTOs and solution architects looking for a storage solution that can handle the rapid growth of unstructured data, cloud, and archival storage while controlling licensing and infrastructure costs. This paper assumes knowledge of enterprise data center administration challenges and familiarity with data center configuration and deployment best practices, primarily with regard to storage systems. It also assumes the reader appreciates both the challenges and benefits open source solutions can bring.

Overview

Business problem

Businesses are looking for better and more cost-effective ways to manage their exploding data storage requirements. In recent years, the amount of storage required for businesses to meet increased data retention requirements has increased dramatically. Cost-per-gigabyte and ease of retrieval are important factors for choosing a solution that can scale quickly and economically over many years of continually increasing capacities and data retention requirements.

Organizations that have been trying to keep up with data growth using traditional file and block storage solutions are finding that the complexity of managing and operating them has grown significantly—as have the costs of storage infrastructure. Storage hosting on a public cloud may not meet cost or data control requirements in the long term. The performance and control of on-premises equipment still offers real business advantages.

Traditional infrastructure is costly to scale massively and offers extra performance features that are not needed for cold or warm data. Object storage on industry standard infrastructure is optimized for this use case and is an ideal supplement to existing infrastructure. Offloading archive data to Ceph—an open source storage platform that stores data on a single distributed computer cluster—can reduce overall storage costs while freeing existing capacity for applications that require traditional infrastructure capabilities.

Challenges of scale

There are numerous difficulties around storing unstructured data at massive scale:

Cost

- Unstructured and archival data tends to be written only once or become stagnant over time. This stale data takes up valuable space on expensive block and file storage.
- Tape is an excellent choice for achieving the lowest cost per GB but suffers extremely high latencies. Unstructured and archival data can sit dormant for long stretches of time yet need to be available in seconds.

Scalability

- Unstructured deployments can accumulate billions of objects and petabytes of data. File system limits on the number and size of files and block storage limits on the size of presented blocks become significant deployment challenges.
- Additionally, block and file storage methods suffer from metadata bloat at a massive scale, resulting in a large system that cannot meet service level agreements.

Availability and manageability

- Enterprise storage is growing from smaller-scale, single-site deployments to geographically-distributed, scale-out configurations. With this growth, the difficulty of keeping all the data safe and available is also growing.
- Many existing storage solutions are a challenge to manage and control at massive scale. Management silos and user interface limitations make it harder to deploy new storage into business infrastructure.

Why SUSE Enterprise Storage?

- Leveraging industry-standard servers means the lowest possible cost for a disk-based system with a building block your organization already understands.
- SUSE Enterprise Storage provides all the benefits of Ceph with the addition of a world-class support organization.
- It is designed to scale indefinitely and scales from one petabyte to well beyond a hundred petabytes of data.
- A flat namespace and per-object metadata means little space is wasted on overhead and the interface scales efficiently to billions of objects.
- A single SUSE Enterprise Storage cluster can be configured to meet the requirements of many different storage needs all at once.
- It is designed to be deployed, accessed, and managed from any location.

SUSE Enterprise Storage use cases

OpenStack cloud storage

SUSE Enterprise Storage integrates well into an OpenStack® cluster. A typical setup uses block storage behind OpenStack Cinder and Ceph object storage in lieu of Swift. Ceph can perform the dual role of ephemeral virtual machine storage for OpenStack Nova and image storage for OpenStack Glance. For security, OpenStack Keystone can be configured to provide authentication to the Ceph cluster. In this setup, Ceph can still be used as block and/or object storage for non-OpenStack applications.

Content repository

For a company that can't or does not want to use a publicly-hosted content repository like Box, Dropbox, or Google Drive, SUSE Enterprise Storage is a low-cost private option. The Ceph object store can be configured to meet appropriate latency and bandwidth requirements for whatever the business need. The widespread S3 and Swift REST interfaces can both be used to access data, which means many existing tools can be used, and new tools do not require significant development work.

Content distribution origin server

Content distribution networks (CDNs) come in both private and public flavors. A business hosting their own, private CDN controls both the origin servers and edge servers. A business using a public CDN must use the content provider's edge servers but may choose to use a private origin server. SUSE Enterprise Storage object interfaces make an excellent origin in both cases. At scale, SUSE Enterprise Storage offers a lower TCO versus closed source object storage solutions or a content provider's origin servers.

Video archive

As video surveillance use grows in commercial, government, and private sectors, the need for low cost, multi-protocol storage is growing rapidly. HPE hardware with SUSE Enterprise Storage provides a platform that is an ideal target for these streams as the various interfaces; iSCSI, S3, and Swift service a wide array of applications. The added ability to provide a write-back cache tier enables the system to also service high-performance short-term streams where only a percentage of requests actually end up being served from the long-term archive.

Backup target

Most, if not all, modern backup applications provide multiple disk-based target mechanisms. These applications are able to leverage the distributed storage technology provided by SUSE Enterprise Storage as a disk backup device. The advantages of this architecture include high-performance backups, quick restores without loading tape medium, and integration into the multi-tier strategy utilized by most customers today. The economics of HPE servers running SUSE Enterprise Storage provide a superior TCO to utilizing traditional storage for these environments.

Solution introduction

How does object storage work?

Object storage allows the storage of arbitrary-sized "objects" using a flat, wide namespace where each object can be tagged with its own metadata. This simple architecture makes it much easier for software to support massive numbers of objects across the object store. The APIs provided by the object storage gateway add an additional layer above objects—called 'containers' (Swift) and 'buckets' (S3)—to hold groupings of objects.

To access the storage, a RESTful interface is used to provide better client independence and remove state tracking load on the server. HTTP is typically used as the transport mechanism to connect applications to the data, so it's very easy to connect any device over the network to the object store.

SUSE Enterprise Storage architecture—powered by Ceph

A SUSE Enterprise Storage cluster is a software defined storage (SDS) architecture built off of Ceph and layered on top of industry-standard servers. It provides a federated view of storage across the cluster where each individual server uses well-known block storage and file systems. This approach has the advantages of leveraging existing work and standard hardware where appropriate while still providing the scale and performance needed to the overall solution. See the SUSE Enterprise Storage page for more details on Ceph.

Cluster roles

There are three primary roles in the SUSE Enterprise Storage cluster covered by this sample reference configuration:

- **OSD Host**—Ceph server storing object data. Each OSD host runs several instances of the Ceph OSD Daemon process. Each process interacts with one Object Storage Disk (OSD), and for production clusters there is a 1:1 mapping of OSD Daemon to logical volume. The default file system used on each logical volume is XFS, although Btrfs is also supported.
- **Monitor (MON)**—Maintains maps of the cluster state, including the monitor map, the OSD map, the Placement Group Map, and the CRUSH map. Ceph maintains a history (called an "epoch") of each state change in the Ceph Monitors, Ceph OSD Daemons, and Placement Groups (PGs). Monitors are expected to maintain quorum to keep an updated cluster state record.
- **RADOS Gateway (RGW)**—Object storage interface to provide applications with a RESTful gateway to Ceph Storage Clusters. The RADOS Gateway supports two interfaces: S3 and Swift. These interfaces support a large subset of their respective APIs as implemented by Amazon and OpenStack Swift.

Keeping data safe

SUSE Enterprise Storage supports data replication as well as erasure coding. Erasure coding mathematically encodes data into a number of chunks that can be reconstructed from partial data into the original object. This is more space efficient than replication on larger objects, but it adds latency and is more computationally intensive. The overhead of erasure coding makes it space inefficient for smaller objects, and SUSE Enterprise Storage block requires a replicated cache tier to utilize it. As such, erasure coding is recommended for capacity efficiency, whereas replication is most appropriate for lower capacity block storage and small objects.

Putting data on hardware

One of the key differentiating factors between different object storage systems is the method used to determine where data is placed on hardware. SUSE Enterprise Storage calculates data locations using a deterministic algorithm called Controlled Replication Under Scalable Hashing (CRUSH). CRUSH uses a set of configurable rules and placement groups (PGs) in this calculation. Placement groups tell data where it is allowed to be stored and are architected in such a way that data will be resilient to hardware failure.

Solution

Solution architecture

HPE Apollo 4500 System

This block diagram has connections to infrastructure outside of the sample reference configuration. Each 4U Apollo 4530 chassis contains three compute nodes populated with 12 spinning disks and three SSD journals each. With high compute-to-storage ratios and SSD journals, these nodes make good low-performance and/or ephemeral block storage hosts. Contrast this to the Apollo 4510 chassis, which each contain only one compute node and are populated with up to 68 spinning disks (with co-located journals). This achieves maximum data density, and these nodes make good object storage hosts. Each Apollo 4530 can be considered a building block for block storage, and each Apollo 4510 can be considered a building block for object storage.

The external link on the blue-labeled 10GbE Data Network connects the cluster to client machines and load balancing. Notice that the Apollo 4530 units have a connection for each compute node. The orange-labeled 10GbE Cluster Network routes traffic only between cluster nodes. The 1GbE Management Network labeled in purple has an uplink to the larger management network. Each compute node in the Apollo 4530 chassis shares a single iLO[1] management port.

Sample block diagram

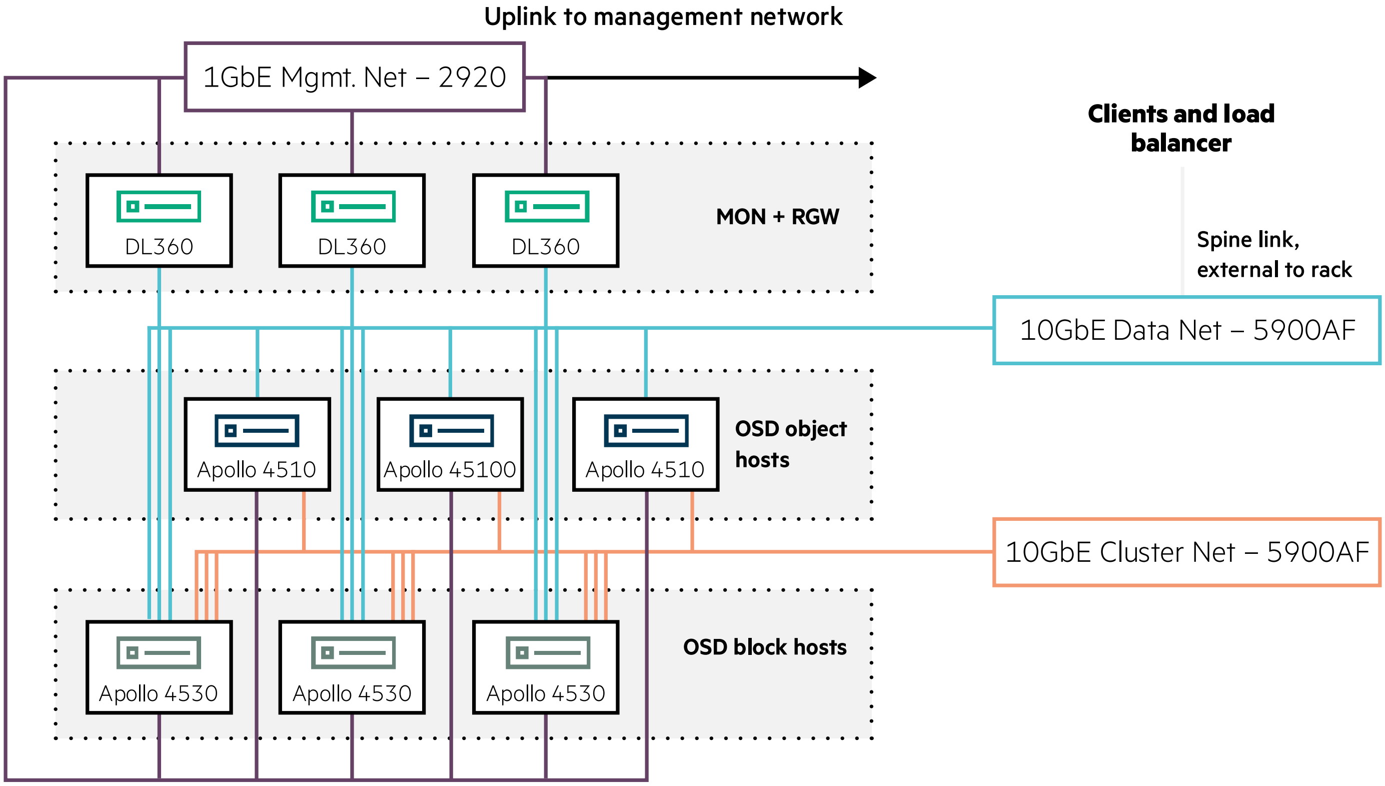

Figure 1. Apollo 4500 reference configuration block diagram

[1] Integrated Lights-Out (iLO) is HPE's advanced baseboard management controller (BMC). Of note, iLO features a robust REST API.

HPE Apollo 4200 System

Apollo 4500 servers are tuned for density optimized tasks, but may not fit all data center requirements. A 1200 mm rack may be too large, or perhaps your enterprise has standardized on 2U servers. Some businesses aim to use the same platform for multiple rack use cases. In these deployments, the Apollo 4200 may be a better fit. Standardizing on 2U servers also creates smaller failure domain and expansion blocks. Not all use cases are big enough to start with an infrastructure where 60+ drives go offline when a server goes down. The resulting solution is not as dense per rack unit; however, the building blocks are individually smaller, and the resulting reference architecture's OSD hosts fit in a total of 12U instead of 24U. In this case, the Apollo 4200 serves as the building block for both block and object storage using SUSE Enterprise Storage and are only differentiated by the configuration of the rear drive cage.

With the Apollo 4200 SUSE Enterprise Storage reference architecture, object storage hosts now have 28 spinning disks—24 in front with four in an LFF rear cage—in a 2U chassis. Contrast this to the 4510 with 68 spinning disks in a 4U chassis—34 disks per 2U. Block storage hosts use the same 24 spinning disks in front but utilize SSD journals in the rear cage for a 6:1 HDD to SSD ratio. This 2U chassis contains one compute node with 24 spinning disks—two compute nodes per 4U. The Apollo 4530 hosts three compute nodes with 12 spinning disks each in a 4U chassis. SUSE Enterprise Storage block storage will benefit from the Apollo 4530's higher compute to disk ratio.

The network configuration here is the same as the Apollo 4500 sample configuration, which you can read about above.

Sample block diagram

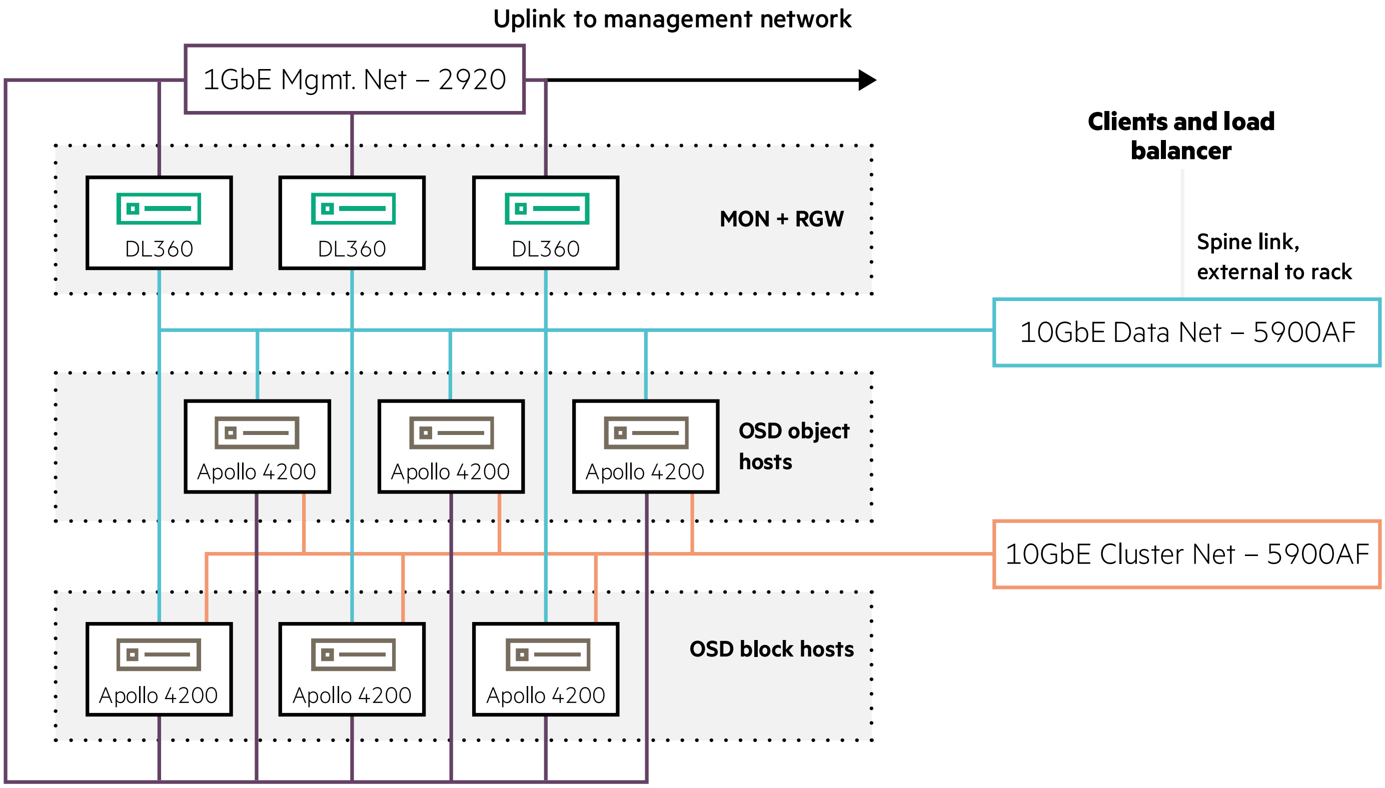

Figure 2. Apollo 4200 reference configuration block diagram

HPE value

Clusters built on a 'white box' server architecture work for business at small scales, but as they grow the complexity and cost make them less compelling than enterprise-focused hardware. With 'white box' solutions, IT has to standardize and integrate platforms and supported components themselves. Support escalation becomes more complicated. Without standardized toolsets to manage the hardware at scale, IT must chart their own way with platform management and automation. Power consumption and space inefficiencies of generic platform design also limit scale and increase cost over time.

The result is IT staff working harder and the business spending more to support the quantity and complexity of a 'white box' hardware infrastructure. The lowest upfront cost does not deliver the lowest total cost or easiest solution to maintain.

Using HPE hardware and software solutions provides advantages like:

- Platform management tools that scale across data centers
- Server components and form factors that are optimized for enterprise use cases
- Hardware platforms where component parts have been qualified together
- A proven, worldwide hardware support infrastructure

Disk encryption

In addition to the benefits above, all Apollo 4000 configurations include an HPE Smart Array card capable of Secure Encryption where enterprise-class encryption is needed. Encryption is FIPS-2 certified for security, has been tested as not affecting IOPS on spinning media for low performance impact, and is transparent to the operating system for ease-of-use. This means any drive supported on the server can be used, giving much more cost/performance flexibility than encryption on drive solutions. Key management is simple and can be managed locally or via an enterprise key management system.

SUSE Value

As a 20+ year provider of mission-critical open source solutions, SUSE is accustomed to ensuring customers have the best engineered and supported solutions possible. SUSE Enterprise Storage is no exception. SUSE has been on the front edge of storage technology for many years and is putting all of its expertise and experience behind making Ceph consumable by enterprise customers. Ensuring stability means a tight marriage between the most reliable enterprise Linux available and the industry leading Ceph distribution, SUSE Enterprise Storage.

With SUSE Enterprise Storage, customers get a solid build of Ceph with additional feature add-ons by SUSE, including: iSCSI support, encryption for data at rest, and optional installation mechanisms. Backed by a world-class support organization, customers can have confidence that SUSE Enterprise Storage is the best place to store data today and into the future.

Server platforms

This section gives some reasons and benefits around the industry standard servers chosen for the reference configuration. Decisions made for component sizing in the cluster (compute, memory, storage, networking topology) are described under the Configuration Guidance section.

ProLiant DL360 Gen9

- 1U rack space density
- Compute and memory scalability appropriate for connection and monitor roles
- Low-cost, low-footprint management platform

HPE Apollo 4500 System

Apollo 4500 System improvements over SL4540 Gen8:

- 4U instead of 4.3U Chassis
- 4 PCIe eight slots and one FlexibleLOM instead of single PCIe 8x mezzanine
- Higher performance Socket R instead of Gen8 Socket B
- Optional H or P series controller option for two boot drives gives more controller and cabling flexibility for drives
- M.2 support improves storage density
- BROC Gen9 provides hardware RAID with a battery-backed write cache and uses an open source driver
- Uses common-slot power supplies

Apollo 4500 System key points:

Apollo 4510 Gen9	Apollo 4530 Gen9
Maximum storage density for object storage	Better compute to disk ratio for block storage
Lowest cost per GB of storage	Smaller failure domain
Cabling and storage controller flexibility	More networking to maximize bandwidth
Tuned for low OPEX and TCO	Tuned for low OPEX and TCO
Space-efficient power and cooling	Space-efficient, shared power and cooling

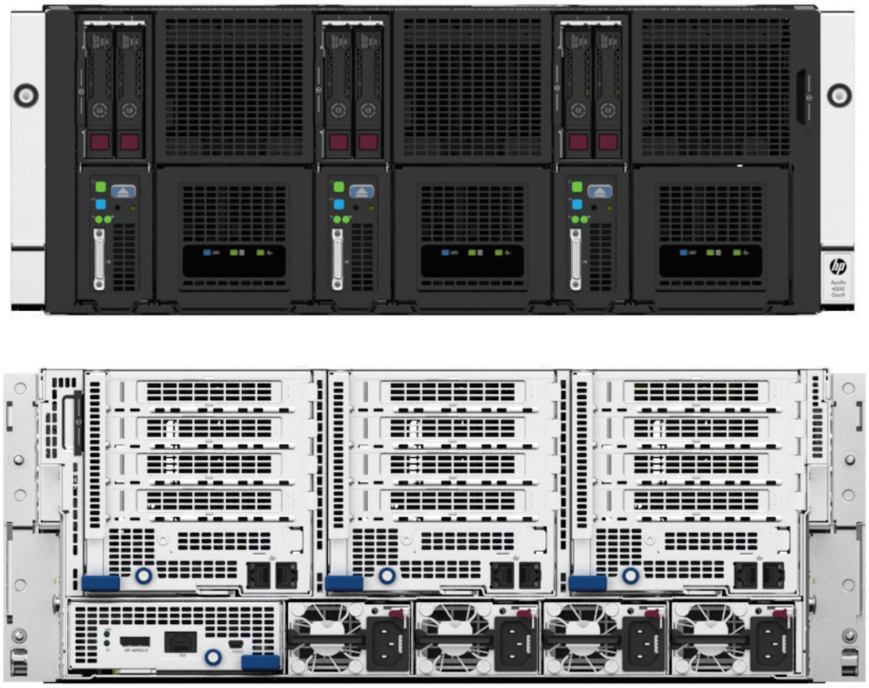

Figure 3. HPE Apollo 4530 Gen9 front (top) and rear (bottom)

HPE Apollo 4200 System

- Maintains 2U industry standard form factors while dramatically improving storage density
  - 24 + four large form factor or 48+2 small form factor max
  - Good for co-located data centers where space is at a premium
- 1075 mm rack compliant
- Front/rear loading drives
- Up to seven PCIe slots, including x16
  - Can increase Ceph object throughput with extra networking and/or additional controllers
  - Can easily be repurposed as base building block for other enterprise/data center use cases
- Rear cage with either SFF + two PCI slots for additional add-on card flexibility or LFF rear cage to improve capacity
- M.2 support further improves storage density
- Lower initial investment versus Apollo 4500 means reduced CAPEX

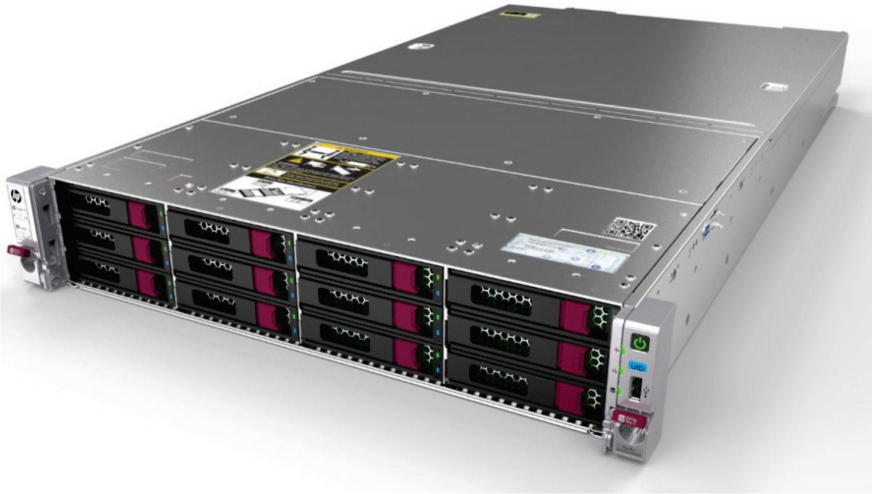

Figure 4. HPE Apollo 4200 System with large form factor drives with drawer closed

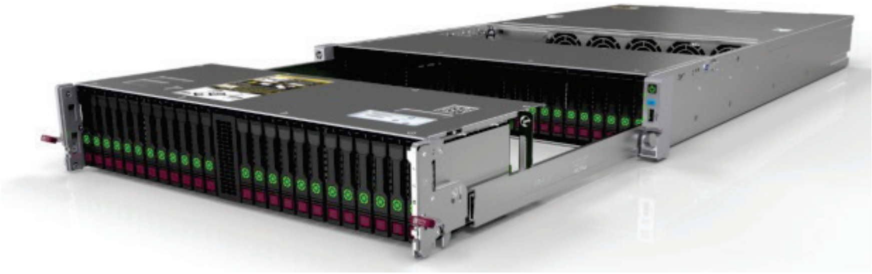

Figure 5. HPE Apollo 4200 System with small form factor drives with drawer open

Configuration guidance

This section covers how to create a SUSE Enterprise Storage cluster to fit your business needs. The basic strategy of building a cluster is this: with a desired capacity and workload in mind, understand where performance bottlenecks are for the use case, and what failure domains the cluster configuration introduces.

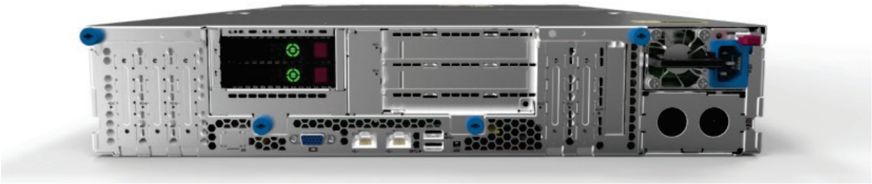

Figure 6. HPE Apollo 4200 System rear view with two small form factor rear drive cage plus 2 PCI card expander option

After choosing hardware, the SUSE Enterprise Storage Deployment & Administration Guide is an excellent place to start for instructions on installing software.

General configuration recommendations

- The slowest performer is the weakest link for performance in a pool. Typically, OSD hosts should be configured with the same quantity, type, and configuration of storage. There are reasons to violate this guidance (pools limited to specific drives/hosts, federation being more important than performance), but it's a good design principle.
- A minimum recommended size cluster would have at least six compute nodes. The additional nodes provide more space for unstructured scale, help distribute load per node for operations, and make each component less of a bottleneck. When considering rebuild scenarios, look at the capacity of a node in relation to available bandwidth. Higher density nodes work better in larger, faster clusters, while less dense nodes should be used in smaller clusters.
- If the minimum recommended cluster size sounds large, consider whether SUSE Enterprise Server is the right solution. Smaller amounts of storage that don't grow at unstructured data scales could stay on traditional block and file or leverage an object interface on a file-focused storage target.
- SUSE Enterprise Server clusters can scale to hundreds of petabytes, and you can easily add storage as needed. However, failure domain impacts must be considered as hardware is added. Design assuming elements will fail at scale.

SSD journal usage

If data requires significant write or PUT bandwidth, consider SSDs for data journaling.

Advantages

- Separation of the highly sequential journal data from object data—which is distributed across the data partition as RADOS objects land in their placement groups—means significantly less disk seeking. It also means that all bandwidth on the spinning media is going to data I/O, approximately doubling bandwidth of PUTs/writes.
- Using an SSD device for the journal keeps storage relatively dense because multiple journals can go to the same higher bandwidth device while not incurring rotating media seek penalties.

Disadvantages

- Each SSD in this configuration is more expensive than a spinning drive that could be put in the slot. Journal SSDs reduce the maximum amount of object storage on the node.
- Tying a separate device to multiple OSDs as a journal and using XFS—the default file system the **ceph-deploy** tool uses—means that loss of the journal device is a loss of all dependent OSDs. With a high enough replica and OSD count this isn't a significant additional risk to data durability, but it does mean architecting with that expectation in mind.
- OSDs can't be hot swapped with separate data and journal devices.
- Additional setup and planning is required to efficiently make use of SSDs.
- Small object IO tends to benefit much less than larger objects.

Configuration recommendations

- For bandwidth, four spinning disks to one SSD is a recommended performance ratio for block storage. It's possible to go with more spinning to solid state to improve capacity density, but this also increases the number of OSDs affected by an SSD failure.
- SSDs can become a bottleneck with high ratios of disks to SSD journals; balance SSD ratios vs. peak spinning media performance. Ratios larger than eight spinning disks to one SSD are typically inferior to just co-locating the journal with the data.
- Even where application write performance is not critical, it may make sense to add an SSD journal purely for rebuild/rebalance bandwidth improvements.
- Journals don't require a lot of capacity, but larger SSDs do provide extra wear leveling. Journaling space reserved by SUSE Enterprise Server should be 10–20 seconds of writes for the OSD the journal is paired with.
- A RAID1 of SSDs is not recommended outside of the monitor nodes. Wear leveling makes it likely SSDs will be upgraded at similar times. The doubling of SSDs per node also reduces storage density and increases price per gigabyte. With massive storage scale, it's better to expect drive failure and plan such that failure is easily recoverable and tolerable.
- Erasure coding is very flexible for choosing between storage efficiency and data durability. The sum of your data and coding chunks should typically be less than or equal to the OSD host count, so that no single host failure can cause the loss of multiple chunks.
- Keeping cluster nodes single function makes it simpler to plan CPU and memory requirements for both typical operation and failure handling.
- Extra RAM on an OSD host can boost GET performance on smaller I/Os through file system caching.

Choosing hardware

The SUSE Enterprise Storage Deployment and Administration Guide provides minimum hardware recommendations. In this section, we expand and focus this information around the reference configurations and customer use cases.

Choosing disks

Choose how many drives are needed to meet performance SLAs. That may be the number of drives to meet capacity requirements, but may require more spindles for performance or cluster homogeneity reasons.

Object storage requirements tend to be primarily driven by capacity, so plan how much raw storage will be needed to meet usable capacity and data durability. Replica count and data to coding chunk ratios for erasure coding are the biggest factors determining usable storage capacity. There will be additional usable capacity loss from journals co-located with OSD data, xfs/btrfs overhead, and logical volume reserved sectors. A good rule of thumb for three-way replication is 1:3.2 for usable to raw storage capacity ratio.

Some other things to remember around disk performance:

- Replica count or erasure encoding chunks mean multiple media writes for each object PUT.
- Peak write performance of spinning media without separate journals is around half due to writes to journal and data partitions going to the same device.
- With a single 10GbE port, the bandwidth bottleneck is at the port rather than controller/drive on any fully disk-populated HPE Apollo 4510 Gen9 server node.
- At smaller object sizes, the bottleneck tends to be on the object gateway's ops/sec capabilities before network or disk. In some cases, the bottleneck can be the client's ability to execute object operations.

Configuring disks

When array controller write cache is available, it is recommended to configure drives in RAID0 with controller write cache enabled to improve small object write performance.

For a fully disk-populated HPE Apollo 4510 Gen9 with 68 drives, significant CPU cycles must be reserved for 68 OSDs on a single compute node. Configuring RAID0 volumes across two drives at a time—resulting in 34 OSDs—could reduce CPU usage. Configuring multiple drives in a RAID array can reduce CPU cost for colder storage in exchange for reduced storage efficiency to provide reliability. It can also provide more CPU headroom for error handling or additional resources if cluster design dictates CPU resource usage outside of cluster specific tasks.

Choosing a network infrastructure

Consider desired bandwidth of storage calculated above, the overhead of replication traffic, and the network configuration of the object gateway's data network (number of ports/total bandwidth). Details of traffic segmentation, load balancer configuration, VLAN setup, or other networking configuration/best practice are very use-case specific and outside the scope of this document.

- Typical choices of configuration for data traffic will be LACP bonded 10GbE links. These links provide resiliency if spanned across switches and aggregated bandwidth.
- Network redundancy (active/passive configurations, redundant switching) is not recommended, as scale-out configurations gain significant reliability from compute and disk node redundancy and proper failure domain configuration. Consider the network configuration (where the switches and rack interconnects are) in the CRUSH map to define how replicas are distributed.
- A cluster network isolates replication traffic from the data network and provides a separate failure domain. Replication traffic is significant, as there are multiple writes for replication on the cluster network for every actual I/O. It is recommended to bond all 10GbE links with LACP and segment the public and backend traffic via VLANs
- It is recommended to reserve a separate 1GbE network for management as it supports a different class and purpose of traffic than cluster I/O.

Matching object gateways to traffic

Start by selecting the typical object size and I/O pattern then compare to the sample reference configuration results. The object gateway limits depend on the object traffic, so accurate scaling requires testing and characterization with load representative of the use case. Here are some considerations when determining how many object gateways to select for the cluster:

- Object gateway operation processing tends to limit small object transfer. File system caching for GETs tends to have the biggest performance impact at these small sizes.
- For larger object and cluster sizes, gateway network bandwidth is the typical limiting factor for performance.
- Load balancing does make sense at scale to improve latency, IOPS, and bandwidth. Consider at least three object gateways behind a load balancer architecture.
- While very cold storage or environments with limited clients may only ever need a single gateway, two is the recommended minimum to protect against a single point of failure.

With the monitor process having relatively lightweight resource requirements, the monitor can run on the same hardware used for an object gateway. Performance and failure domain requirements dictate that not every monitor host is an object gateway, and vice versa.

Monitor count

Use a minimum of three monitors for a production setup. While it is possible to run with just one monitor, it's not recommended for an enterprise deployment, as larger counts are important for quorum and redundancy. With multiple sites it makes sense to extend the monitor count higher to maintain a quorum with a site down.

Use physical boxes rather than virtual machines to have separate hardware for failure cases. It is recommended that the monitors utilize mirrored SSDs due to the high number of fsync calls on these nodes.

Bill of materials

This section contains SKUs for components of the RA servers used as SUSE Enterprise Storage building blocks. This helps demonstrate configuration guidance, and provides a practical starting point for sizing a real POC or deployment. Because of the focus on industry standard servers in this RA, we do not present a comprehensive BOM for an entire solution.

Components selected for operational requirements, inter-rack and/or inter-site networking, and service and support can vary significantly per deployment and are complex topics in their own right. Work with your HPE representative to complete the picture and create a SUSE Enterprise Storage cluster that fits all requirements.

3 x HPE Apollo 4530 Gen9

Qty	Part Number	Description
9	786595-B23	HPE ProLiant XL450 Gen9 Configure-to-order Server Node for Apollo 4530 Chassis
9	783934-L21	Intel® Xeon® E5-2670v3 (2.3 GHz/12-core/30 MB/120 W) FIO Processor Kit
72	726719-B21	HPE 16 GB (1 x 16 GB) Dual Rank x4 DDR4-2133 CAS-15-15-15 Registered Memory Kit
27	797303-B21	HPE 480 GB 6G SATA Value Endurance LFF 3.5-in LP Enterprise Value 3yr Warranty Solid State Drive
108	805334-B21	HPE 8 TB 6G SATA 7.2 k rpm LFF (3.5-inch) Low Profile Midline 512e 1yr Warranty Hard Drive
18	655708-B21	HPE 500 GB 6G SATA 7.2 k rpm SFF (2.5-inch) SC Midline 1yr Warranty Hard Drive
9	665243-B21	HPE Ethernet 10 Gb 2-port 560FLR-SFP+ Adapter
9	808965-B21	HPE Apollo 4500 Mini SAS P440 Cable Kit
9	761878-B21	HPE H244br FIO Smart HBA
9	726821-B21	HPE Smart Array P440/4G Controller
3	727258-B21	HPE 96w Megacell Batt Cntrl Bd
3	799581-B23	HPE Apollo 4530 Gen9 CTO Chassis
12	720479-B21	HPE 800 W Flex Slot Platinum Plus Hot Plug Power Supply Kit
3	681254-B21	HPE 4.3U Server Rail Kit
9	512485-B21	HPE iLO Advanced including 1yr 24x7 Technical Support and Updates Single Server License

3 x HPE Apollo 4510 Gen9

Qty	Part Number	Description
3	786593-B21	HPE ProLiant XL450 Gen9 Configure-to-order Server Node for Apollo 4510 Chassis
3	783936-L21	Intel Xeon E5-2690 v3 (2.6 GHz/12-core/30 MB/135 W) FIO Processor
3	783936-B21	Intel Xeon E5-2690v3 (2.6 GHz/12-core/30 MB/135 W) Processor
48	726722-B21	HPE 32 GB (1 x 32 GB) Dual Rank x4 DDR4-2133 CAS-15-15-15 Registered Memory Kit
204	805334-B21	HPE 8 TB 6G SATA 7.2 k rpm LFF (3.5-inch) Low Profile Midline 512e 1yr Warranty Hard Drive
6	655708-B21	HPE 500 GB 6G SATA 7.2 k rpm SFF (2.5-inch) SC Midline 1yr Warranty Hard Drive
6	777262-B21	HPE 120 GB 6G SATA Read Intensive M.2 2280 3yr Warranty Solid State Drive
6	665243-B21	HPE Ethernet 10 Gb 2-port 560FLR-SFP+ Adapter
3	808967-B21	HPE Apollo 4500 Mini SAS Dual P440 Cable Kit
3	761878-B21	HPE H244br FIO Smart HBA
3	726897-B21	HPE Smart Array P840/4G Controller
3	727258-B21	HPE 96 w Megacell Batt Cntrl Bd
3	799581-B21	HPE Apollo 4510 Gen9 CTO Chassis
3	799377-B21	HPE Apollo 4510 8HDD Rear Cage Kit
6	720620-B21	HPE 1400 W Flex Slot Platinum Plus Hot Plug Power Supply Kit
3	681254-B21	HPE 4.3U Server Rail Kit
3	512485-B21	HPE iLO Advanced including 1yr 24x7 Technical Support and Updates Single Server License

6 x Apollo 4200 System

In total, this BOM lists components for three block storage servers and three object storage servers. The configuration is as consistent as possible across the two server types. The key difference between the two is the block storage server has SSDs in the rear slots for better write bandwidth. M.2 devices are used for boot storage to maximize storage density from SUSE Enterprise Server OSDs.

Qty	Part Number	Description
6	808027-B21	HPE Apollo 4200 Gen9 24LFF CTO Server
6	803314-L21	HPE Apollo 4200 Gen9 Intel Xeon E5-2680 v3 (2.5 GHz/12-core/30 MB/120 W) FIO Processor Kit
6	803314-B21	HPE Apollo 4200 Gen9 E5-2680 v3 Kit
72	728629-B21	HPE 32 GB (1 x 32 GB) Dual Rank x4 DDR4-2133 CAS-15-15-15 Registered Memory Kit
6	777894-B21	HPE Dual 120 GB Value Endurance Solid State M.2 Enablement Kit
6	806563-B21	HPE Apollo 4200 Gen9 LFF Rear HDD Cage Kit
156	797269-B21	HPE 6 TB 6G SATA 7.2 k rpm LFF (3.5-inch) Low Profile Midline 1yr Warranty Hard Drive
12	665243-B21	HPE Ethernet 10 Gb 2-port 560FLR-SFP+ Adapter
12	720479-B21	HPE 800 W Flex Slot Platinum Hot Plug Power Supply Kit
6	822731-B21	HPE 2U Shelf-Mount Adjustable Rail Kit
6	813546-B21	HPE SAS Controller Mode for Rear Storage
12	797303-B21	HPE 480 GB 6G SATA Value Endurance LFF 3.5-in LPC Enterprise Value 3yr Warranty Solid State Drive
6	512485-B21	HPE iLO Advanced including 1yr 24x7 Technical Support and Updates Single Server License

3 x HPE ProLiant DL360 Gen9

Qty	Part Number	Description
3	755259-B21	HPE DL360p Gen9 4-LFF CTO Server
3	764097-L21	HPE DL360 Gen9 Intel Xeon E5-2630L v3 FIO Processor Kit
6	726719-B21	HPE 16 GB (1 x 16 GB) Dual Rank x4 DDR4-2133 CAS-15-15-15 Kit
3	749976-B21	HPE H240ar 12 Gb 2-ports Int FIO Smart Host Bus Adapter
3	665243-B21	HPE Ethernet 10 Gb 2-port 560FLR-SFP+ Adapter
6	652745-B21	HPE 500 GB 6G SAS 7.2 k rpm SFF (2.5-inch) SC Midline
6	720478-B21	HPE 500 W Flex Slot Platinum Hot Plug Power Supply Kit
3	789388-B21	HPE 1U LFF Gen9 Easy Install Rail Kit
3	512485-B21	HPE iLO Advanced including 1yr 24x7 Technical Support and Updates Single Server License

SUSE Enterprise Storage 3 Deployment for HPE DL Series

Hardware configuration

This is a representative configuration used for testing. The hardware more than meets the recommendations in the architectural overview document.

OSD Node (Qty 4)

- DL380G9
  - 2 x E2660 v3
  - 128 GB RAM
  - 2200 GB Value SSD in rear bay
    - Configure this first
    - Set to RAID 1 leaving all defaults
  - 250 GB Write Intensive SSD in 3.5" bay
    - Individual RAID-0
    - 90% write-back cache leaving all other settings at default
    - Configured as logical devices 2 & 3
  - 10 6 TB 7200 RPM SAS
    - Individual RAID-0
    - 90% write-back cache leaving all other settings at default
  - 2 Intel 530SFP+

Monitor Node (Qty 3) & Admin Node (Qty 1)

- DL360 Gen9
  - 2 x E5-2690 v3
  - 64 GB
  - 2 x SSD
  - 1 Intel 530SFP+ dual port 10GbE

2 x HPE FlexFabric 5700-40XG-2QSFP+ switches

Step-by-step instructions

Planning and prerequisites

Preplan the IP range to be utilized. Then create a single storage subnet where all nodes, gateways, and clients will connect. In many cases, this may entail using a range larger than a standard /24. While storage traffic can be routed, it is generally discouraged to help ensure lower latency.

Set up DNS A records for the OSD and Mon nodes.

Decide on subnets and VLANs to be utilized and configure the switch ports accordingly. Guidance is provided in the SUSE Enterprise Storage Architecture Overview with Recommendations.

Subscription Management Tool—This service provides a local mirror of the SUSE repositories, allowing for rapid software deployment and updating. More information can be found in the Subscription Management Tool (SMT) for SUSE Linux Enterprise 11 SP3.

Configure all nodes of similar types identically. With the RAID controllers, this is especially important as it greatly simplifies drive configuration.

Network

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 to take care of are creating the IRF topology (stacking), LACP groups, and enabling jumbo frames. Each aggregation group needs a unique number, planned ahead of time. It is also desirable to disable the spanning tree on the ports utilized for storage.

Configure IRF as described in the HPE FlexFabric 5700 Switch Series IRF Configuration Guide.

To configure an LACP group for switch 1 port 1 and switch 2 port 1 and enable jumbo frame support, perform the following commands:

```
System-view
Interface Bridge-aggregation 1
Link-aggregation mode dynamic
quit
interface Ten-GigabitEthernet 1/0/1
port link-aggregation group 1
jumboframe enable 9100
stp disable
interface Ten-GigabitEthernet 2/0/1
port link-aggregation group 1
jumboframe enable 9100
stp disable
quit
save
```

Repeat these steps for each aggregation group required.

Create at least one VLAN for cluster communication. In this example, we are using VLAN 3001 for the cluster traffic.

```
system-view
vlan 3001
name ceph-cluster
```

```
        description vlan for ceph cluster back end communication
        quit
        save
```

Assign VLANs to ports. The configuration below assumes a port-based VLAN of one (default).

```
        system-view
        interface bridge-aggregation 1
        port link-type hybrid
        port hybrid vlan 1 untagged
        port hybrid vlan 3001 tagged
        quit
        save
```

Operating system deployment

The second major step is to install the operating system. In the reference configuration, there are two SSD drives, ten 6 TB near-line SAS drives for data, and a RAID-1 of SSD drives for the OS. All nodes of the same role should be configured the same.

When deploying the operating system, be sure to utilize only the first device. This is the RAID-1 of 200 GB SSD drives.

- Perform a basic network configuration during installation. This will include host name, domain name, and IP address.
- Minimal install with SUSE Enterprise Storage repository selected.
- Deselect x-windows and Gnome.

After the installation is complete, it is recommended that you rewrite the /etc/udev/rules.d/70-persistent-network-rules file to have ports set up identically on each node. After the file has been modified, each node will need to be rebooted and the final network configurations restored.

Verify connectivity and then run zypper up to ensure there no further updates.

The table below shows the IP assignments used for this exercise. These should be adjusted as needed for each individual deployment.

Node	Front end IP (VLAN 1)	Back end IP (VLAN 3001)
sesadmin	192.168.124.90	192.168.100.90
monnode1	192.168.124.91	192.168.100.91
monnode2	192.168.124.92	192.168.100.92
monnode3	192.168.124.93	192.168.100.93
osdnode1	192.168.124.101	192.168.100.101
osdnode2	192.168.124.102	192.168.100.102
osdnode3	192.168.124.103	192.168.100.103
osdnode4	192.168.124.104	192.168.100.104

SUSE Enterprise Storage deployment

The third major step is to deploy Ceph as described in the Ceph Object Gateway SUSE documentation.

On sesadmin, perform the following:

Configure ntp:

```
yast <return> ->Network Services ->NTP Configuration->[Start NTP Daemon]
Now and on Boot->Add->Server->(your preferred NTP Server)->OK->OK->Quit
```

Configure sudoers

```
visudo
```

Add the following to the end of the file

```
ceph ALL = (root) NOPASSWD:ALL
```

On each node, create the cephadm user and set the password:

```
useradd -m cephadm && passwd cephadm
```

Create and distribute the ssh-key for the cephadm user. From sesadmin:

```
su - cephadm

ssh-keygen

ssh-copy-id cephadm@osdnode1

Repeat for each node
```

Copy /etc/sudoers to each node:

```
sudo scp /etc/sudoers root@osdnode1:/etc/sudoers

Repeat for each node
```

Install ntp on each node:

```
sudo zypper in ntp yast2-ntp-client

Repeat for each node
```

Copy /etc/ntp.conf to each node:

```
sudo scp /etc/ntp.conf root@osdnode1:/etc/ntp.conf

Repeat for each node
```

On each node:

```
sudo systemctl enable ntpd

sudo systemctl start ntpd
```

Install ceph on all nodes:

```
sudo zypper in ceph
```

Install ceph-deploy on the admin node

```
sudo zypper in ceph-deploy
```

Disable IPv6:

Edit /etc/sysctl.conf and add these lines to the bottom of the file:

net.ipv6.conf.all.disable_ipv6 = 1

net.ipv6.conf.default.disable_ipv6 = 1

net.ipv6.conf.lo.disable_ipv6 = 1

Copy /etc/sysctl.conf to all nodes and reboot them:

```
sudo scp -p /etc/sysctl.conf root@osdnode1:/etc/sysctl.conf
Repeat for each node
```

Run ceph-deploy install for all nodes:

```
ceph-deploy install sesadmin osdnode1 osdnode2 osdnode3 osdnode4 monnode1 monnode2 monnode3
```

Set up the monitor nodes:

```
ceph-deploy new monnode1 monnode2 monnode3
```

Modify ceph.conf to reflect the networks. In the [global] section, add/edit the following:

```
public network = 192.168.124.0/24
cluster network = 192.168.100.0/24
```

Make sure the firewall is off on all nodes:

```
sudo /sbin/SuSEfirewall2 status
Repeat on each node
```

Create the initial monitor service:

```
ceph-deploy mon create-initial
```

Next prepare the OSD Nodes using the ceph-deploy osd create command. The first device (e.g., sdd) is the data drive and the last device (sdb) is the journal device.

ceph-deploy osd prepare osdnode{1..4}:sd{d..h}:sdb

ceph-deploy osd prepare osdnode{1-4}:sd{i..m}:sdc

ceph-deploy osd activate osdnode{1..4}:sd{d..h}1:sdb

ceph-deploy osd activate osdnode{1..4}:sd{i..m}1:sdc

Deploy the admin node(s):

```
ceph-deploy admin sesadmin
```

Install Romana on the admin nodes

As cephadm:

```
ceph-deploy admin monnode1 monnode2 monnode2
```

As root:

```
zypper in romana

calamari-ctl initialize

su - cephadm

ceph-deploy calamari --master sesadmin connect osdnode1 osdnode2 osdnode3 osdnode4 monnode1
monnode2 monnode3
```

The Romana interface may now be accessed via a browser at http://sesadmin.

Basic SUSE Enterprise Storage cluster validation

After basic configuration of the SUSE Enterprise Storage cluster is complete, perform a basic set of checks to ensure the cluster is operating as expected.

From the admin node:

```
ceph status

ceph osd pool create test 4096

rados bench -p test 300 write --no-cleanup

rados bench -p test 300 seq
```

After validation is complete, remove the test pool.

```
ceph osd pool delete test test -yes-i-really-really-mean-it
```

For further information on SUSE Enterprise Storage, please visit SUSE Enterprise Storage, powered by Ceph.

Summary

With rapid growth of unstructured data and backup/archival storage, traditional storage solutions are lacking in their ability to scale or efficiently serve this data. For unstructured data, performance capabilities of SAN and NAS are often less important than cost per gigabyte of storage at scale. Management of the quantity of storage and sites is complicated, and guaranteeing enterprise reliability to the clients becomes difficult or impossible.

SUSE Enterprise Storage on HPE hardware uses object storage and industry-standard servers to provide the cost, reliability, flexibility, and centralized management businesses need for petabyte unstructured storage scale and beyond. Industry-standard server hardware from HPE is a reliable, easy-to-manage, and supported hardware infrastructure for the cluster. SUSE Enterprise Storage provides the same set of qualities on the software side. Together, they form a solution with a lower TCO than traditional storage that can be designed and scaled for current and future unstructured data needs.

Importantly, the solution brings the control and cost benefits of open source to those enterprises that can leverage it. Enterprise storage features and functionality with a supported open source cost provides great TCO. All this with no inherent vendor lock-in from the cluster software.

This paper shows HPE Apollo 4200 and Apollo 4500 servers as the foundation of a SUSE Enterprise Storage solution for enterprise scale-out storage needs. With these pieces, your business can create a solution that meets scale and reliability requirements at massive scale, realize the TCO improvements of software-defined storage on industry-standard servers, and leverage the strengths of open source in your operations.

Glossary

- **Cold, warm, and hot storage**—Temperature in data management refers to frequency and performance of data access in storage. Cold storage is rarely accessed and can be stored on the slowest tier of storage. As the storage 'heat' increases, the bandwidth over time, as well as instantaneous (latency, IOPS) performance requirements increase.
- **CRUSH**—Controlled Replication Under Scalable Hashing. CRUSH uses 'rules' and placement groups to compute the location of objects deterministically in a SUSE Enterprise Server cluster.
- **Failure domain**—Area of the solution impacted when a key device or service experiences failure.
- **Federated storage**—Collection of autonomous storage resources with centralized management that provides rules about how data is stored, managed, and moved through the cluster. Multiple storage systems are combined and managed as a single storage cluster.
- **Object storage**—Storage model designed for massive scale implemented using a wide, flat namespace. Focuses on data objects instead of file systems or disk blocks, and metadata is applied on a per-object basis to give the object context. Typically accessed by a REST API. A subset of SDS.
- **Placement Group (PG)**—A mapping of objects onto OSDs; pools contain many PGs, and many PGs can map to one OSD.
- **Pool**—Logical, human-understandable partitions for storing objects. Pools set ownership/access to objects, the number of object replicas, the number of placement groups, and the CRUSH rule set to use.
- **RADOS**—A Reliable, Autonomic Distributed Object Store. This is the core set of SUSE Enterprise Server software that stores the user's data.
- **REST**—Representational State Transfer is stateless, cacheable, layered client-server architecture with a uniform interface. In SUSE Enterprise Server, REST APIs are architected on top of HTTP. If an API obeys REST principles, it is said to be 'RESTful.'
- **SDS**—Software-defined storage is a model for managing storage independently of hardware. Also typically includes user policies and may include advanced features like replication, deduplication, snapshots, and backup.

For more information

With increased density, efficiency, serviceability, and flexibility, the HPE Apollo 4000 Server family is the perfect solution for scale-out storage needs. To learn more about density optimized servers visit the HPE Apollo Systems webpage.

SUSE Enterprise Storage is built on Ceph and has excellent documentation available at its website; this white paper has sourced it extensively. The documentation master page starts at SUSE Enterprise Storage.

Learn more at

hpe.com/servers/apollo

Sign up for updates

4AA6-6494ENW, July 2016