Hewlett Packard Enterprise

Technical white paper

SUSE Enterprise Storage on HPE Apollo 4200/4500 System Servers

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

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 open source storage architectures such as Ceph 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. Additionally, modern businesses must provide data access from anywhere at any time, through a variety of legacy and modern access protocols.

Ceph software-defined storage is 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 storage servers provide a comprehensive and cost-effective storage capacity building block for Ceph-based solutions.

Ceph has its code roots in the open source community. 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 to build enterprise-ready Ceph solutions 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 Ceph software-defined storage on industry-standard x86 server systems, to serve their block, file, and object needs.

Hewlett Packard Enterprise hardware combined with SUSE Enterprise Storage delivers an open source unified block, solution that:

- Has software that offers practical scaling from a few hundred terabytes to petabytes and well beyond traditional storage systems
- Lowers up-front solution investment and total cost of ownership (TCO) per gigabyte
- Provides a single software-defined storage (SDS) cluster for object, file, and low to midrange performance block storage
- Uses open source software, 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 and file 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 4000 Gen10 server 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

This paper is written for administrators and solution architects who deploy software-defined storage solutions within their data centers. 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. Ceph software-defined storage on industry-standard infrastructure is optimized for this use case and is an ideal supplement to existing infrastructure by creating a network-based active archive repository. Offloading archive data to Ceph—an open source storage platform that stores data on a single distributed 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 tend 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 and 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 SLAs.

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
- Designed to scale indefinitely and scales from hundreds of terabytes to petabytes and well beyond traditional storage systems
- 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
- 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 use cases, the need for low-cost, multiprotocol 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. Additionally, most backup software now also supports the S3 protocol. This protocol allows for high performance backup to be achieved in the data center while also enabling backup to remote sites where latency may be a factor, thus making it a frequent choice for data center backups. The economics of HPE servers running SUSE Enterprise Storage provide a superior TCO to utilizing traditional storage for these environments.

HPC Archive

Ceph provides a compelling solution as an HPC archive solution. HPC clusters are typically utilized for many projects requiring significant amounts of storage. After the project has run its course, the original data set and results may be unused for a long period of time. By pairing SUSE Enterprise Storage with HPE's DMF product, Lustre environments are able to offload stale data to a lower cost archive location where it can be quickly recalled when the need arises.

Solution introduction

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 gateway. These gateways can scale horizontally using load balancing techniques.

SUSE Enterprise Storage architecture—powered by Ceph

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.

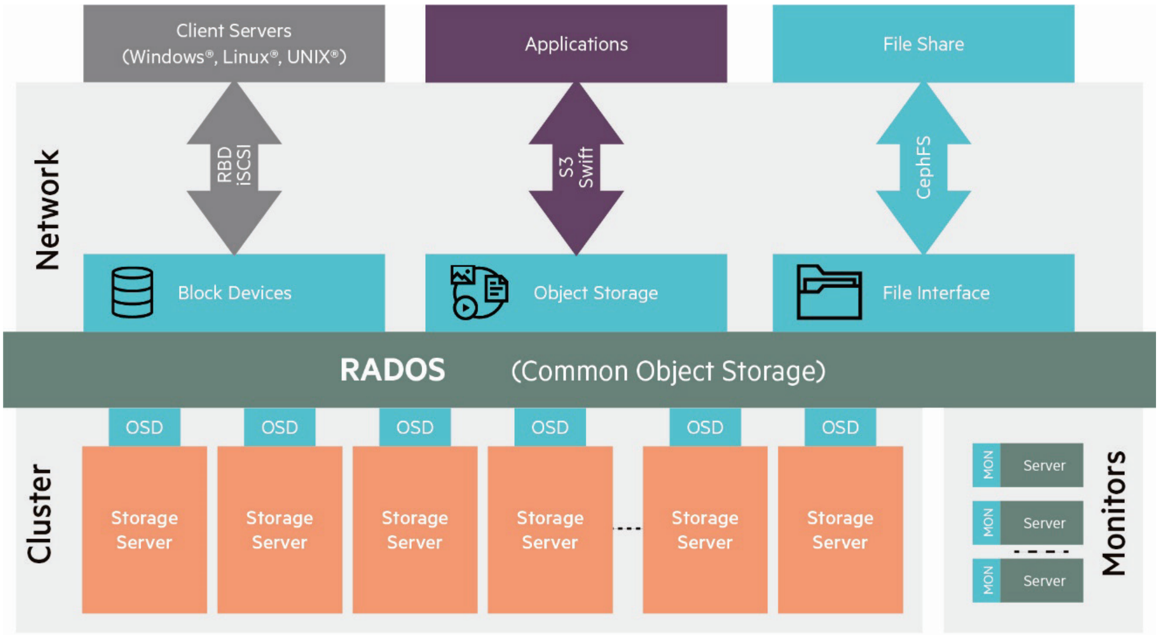

Figure 1. Ceph architecture diagram

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 storage system for SUSE Enterprise Storage is BlueStore. This technology provides native Ceph control of the storage back-end reducing latencies experienced with previous versions of Ceph.

Monitor (MON)—Combining both the monitor and ceph-manager functions, this node type 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 Group (PGs).

Monitors are expected to maintain quorum to keep an updated cluster state record. The ceph-manager daemon is also responsible for providing a centralized management daemon for the cluster and maintains information about the cluster status.

Administrator—This is the self-master and hosts the SUSE Enterprise Storage administrative interface, openATTIC. This web-based storage management console makes day-to-day administration of the cluster a relatively simple task by providing dashboards and configuration interfaces for the most common tasks.

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.

A minimum SES v5.5 cluster should contain:

- One administrator (typically a ProLiant DL360 server)
- Three or more MON nodes (typically ProLiant DL360 servers)
- Three or more OSD nodes for testing, four or more for production
- One or more RGW (typically ProLiant DL360 severs)
- Optional: iSCSI gateway (one or more ProLiant DL360 server)

Density-optimized Apollo 4000 servers are ideal for use as the bulk storage OSD nodes. Ceph supports mixing Apollo 4000 server types and generations, enabling seamless growth with current technologies.

Keeping data safe

SUSE Enterprise Storage brings Ceph's flexibility to bear by supporting 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. Leveraging the ISA plugin to create the erasure coded pools, leverages an accelerator built into Intel® processors, thus reducing the computational load created by erasure coding. Recent versions of Ceph have added the ability to utilize erasure coding as a backing pool for both block and file storage.

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. Ceph calculates data locations using a deterministic algorithm called Controlled Replication Under Scalable Hashing (CRUSH). CRUSH uses a set of configurable rules and Placement Group (PGs) in this calculation. Placement Group tells data where it is allowed to be stored and are architected in such a way that data will be resilient to hardware failure.

Solution

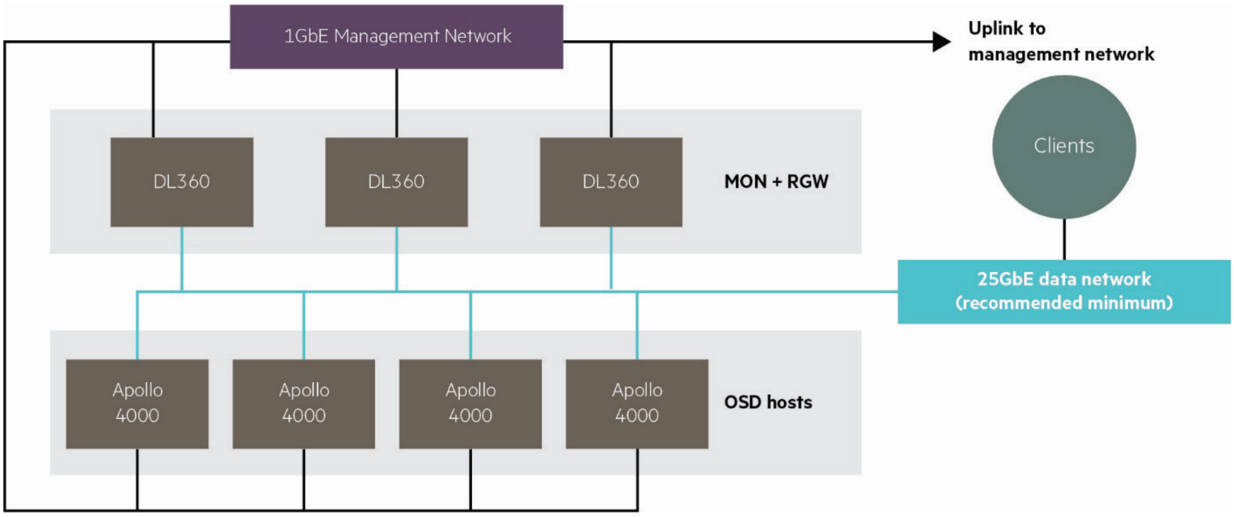

Figure 2. Apollo 4000 sample block diagram

SUSE Enterprise Storage v5.5

SUSE Enterprise Storage 5.5 based on the Ceph Luminous release includes many new and updated features that accelerate innovation including multiprotocol block, object and filesystem data access capabilities, multisite object replication and asynchronous block mirroring, a new open source management framework.

SES v5.5 release highlights include:

- Further reduce capital and operational costs for your storage infrastructure with a truly unified block, object and files solution with production ready Ceph filesystem (CephFS) providing native filesystem access.
- Advanced graphical user interface for simplified management and improved cost efficiency, using the openATTIC open source storage management system.
- Snapshots management for RBD.
- Support for SMB/CIFS in low-throughput production environments.
- Enhanced cluster orchestration using Salt for simplified storage cluster management.

Note
SUSE Enterprise Storage 5.5 uses SLES 12 SP3 as the base operating system

Hewlett Packard Enterprise value for a Ceph storage environment

Software-defined storage running on Linux servers can be deployed on a variety of hardware platforms. However, clusters built on a white-box server infrastructure work for business at small scale, but as they grow, the complexity and cost make them less compelling than enterprise hardware-based solutions. With white-box server infrastructure, IT has to standardize and integrate platforms as well as supported components themselves, and 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. Often the result is the IT staff working harder and the businesses spending more to support a white-box hardware infrastructure than the one-time CAPEX savings realized in buying the white-box servers.

Using an HPE hardware and software solution provides advantages that reduce OPEX spending not available in an infrastructure built on white-box servers. Key OPEX savings from using an integrated HPE solution are:

- 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. hpe.com/servers/secureencryption

Multigenerational Ceph support

Ceph cluster support mixing multiple generations of server storage nodes. Apollo 4000 Gen10 can be used to expand existing storage clusters based on Apollo Gen9 systems.

SUSE value

As a provider of mission-critical open source solutions for over 20 years, 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 optimized 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 provides 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.

HPE Apollo 4510 Gen10 System

The Apollo 4510 Gen10 is a third-generation density-optimized platform recommended for use as Ceph OSD nodes. Key attributes of the HPE Apollo 4510 Gen10 server include:

- Chassis
  - The Apollo 4500 Gen10 chassis is 4 RU
  - Uses Gen10 HPE Flexible Slot Power Supplies, which provides support for 800W 48 VDC and 277 VAC environments, in addition to standard AC environments for 800W and 1400W Platinum and 800W Titanium hot-plug power supply kits
- Processor
  - Intel® Xeon® Processor Scalable Family from 4-26 cores
  - HPE DDR4 SmartMemory up to 2666 MT/s. Up to 1024 GB of LRDIMM memory with 2 processors
- Memory
  - HPE DDR4 SmartMemory up to 2666 MT/s. Up to 1024 GB of LRDIMM memory with 2 processors
- OS drive controller/drives
  - HPE Smart Array P408i-a SR Gen10 Controller (for the two server node SFF drives and M.2 drives.)
  - Optional Smart Array E208i-p or P408i-p SR Gen10 Controllers for 60 bulk LFF drives
  - 12G SAS supported
- Storage
  - 60 LFF data drives with 12G SAS support, side accessible
  - 2 SFF SATA drives, front accessible
  - UFF Dual M.2

- Networking
  - Two 1GbE NIC ports, optional FlexibleLOM
  - Two 1GbE dedicated management ports
- PCIe slots
  - x2 FIO I/O Module with 4 PCIe 3.0 slots, 3 x16 slots, 1 x8 slot for FlexLOM
  - x1 FIO I/O Module with 4 PCIe 3.0 slots, 2 x8 slots, 1 x16 slot, 1 x8 slot for FlexLOM
- On System Management
  - HPE iLO 5 Management Engine
  - Optional HPE iLO Advanced and Premium Security Edition products
- Cluster Management (optional)
  - HPE Insight Cluster Management Utility (CMU)

Figure 3. HPE Apollo 4510 Gen10 storage server system front view

HPE Apollo 4200 Gen10 System

- Chassis
  - The Apollo 4200 Gen10 is a 2 RU server that fits in a standard 1075 mm rack
  - Uses Gen10 HPE Flexible Slot Power Supplies, which provides support for 800W 48 VDC and 277 VAC environments, in addition to standard AC environments for 800W and 1400W Platinum and 800W Titanium hot-plug power supply kits
- Processor
  - Intel Xeon Processor Scalable Family processors
  - HPE DDR4 SmartMemory up to 2666 MT/s. Up to 1024 GB of LRDIMM memory with 2 processors
  - 2x 2666 MT/s NVDIMMs per processors
- OS drive controller/drives
  - HPE Smart Array S100i, optional HPE Smart Array cards
  - Optional M.2 flash devices can be used for the OS drives
- Storage
  - Supports up to 24 LFF data drives with 4 SFF read drives for OS and metadata
  - Maximum storage capacity is 288 TB (24 x 12 TB)
- PCIe slots
  - Supports up to 5 x PCIe 3.0 x8 slots

- On System Management
  - HPE iLO 5 Management Engine
  - Optional HPE iLO Advanced products
- Cluster Management (optional)
  - HPE Insight Cluster Management Utility (CMU)

Figure 4. HPE Apollo 4200 Gen10 System with large form factor drives; 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. After choosing hardware, SUSE Enterprise Storage Administration and Deployment 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 OSD storage nodes. While the minimum is four 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 latency is a concern for the data in the cluster, consider utilizing SSD or NVMe for offloading the Write Ahead Log (WAL) and RocksDB or for a separate storage pool for metadata.

Advantages

- Both SSD and NVMe provide far superior latencies when compared to 7200 rpm disk mechanisms. By leveraging these to offload operations such as the WAL and RocksDB and/or the metadata and index pools, Ceph will spend less time waiting for completion of I/O requests for this data.

Disadvantages

Using flash type media is not without drawbacks.

- Cost is an obvious concern in leveraging these technologies. When considering SSD, a ratio of 12G SAS spinners to 12G SAS SSD should not exceed 8:1, while the lower throughput of SATA devices makes 5:1 near the maximum ratio. NVMe devices provide higher ratios with 12:1 being quite common.
- OSDs can't be hot swapped with separate data and journal devices.

Configuration recommendations

- Device type is an important consideration when selecting flash for WAL and RocksDB. As these devices are being used for write heavy workloads, it is important to select devices optimized for write-intensive environments, thus lowering the variety of devices which may be an optimal fit.
- It is recommended that OS devices be deployed on a RAID 1 of flash-based storage where possible. In the Apollo 4000 Gen10, the recommendation would be for 240 GB+ of M.2.
- Erasure coding is very flexible for choosing between storage efficiency and data durability. The sum of your data and coding chunks should typically be at least one less than 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.

Choosing hardware

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

Choosing CPUs

Proper consideration must be given to CPU selection when building a Ceph cluster. The recommendation from SUSE is to provide one 2 GHz core per spinning OSD. If the cluster is going to have a significant number of NVMe devices, it is recommended that those be provided one 2 GHz core per device.

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.

Some other things to remember around disk performance:

- Replica count or erasure encoding chunks mean multiple media writes for each object write/PUT.
- Mixing disk types results in the performance of a particular bus being reduced to that of the slowest device on a bus. E.g., pairing a 6 Gb/s SATA spinners with 12 Gb SAS SSDs, results in the SSDs operating at the lower performance of 6 Gb/s.
- 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 RAID 0 with controller write cache enabled to improve small object write performance.

Choosing a network infrastructure

Consider the desired bandwidth of storage calculated in the preceding paragraph, 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 25GbE or 100GbE 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 links with LACP and segment the public and back-end 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:

- Load balancing does make sense at scale to improve availability, latency, IOPS, and bandwidth. Consider at least two 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 iSCSI gateway, object gateway or metadata server. 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.

Sample bill of materials

This section contains SKUs for components of the 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 paper, 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.

1x HPE Apollo 4510 Gen10 as object storage servers

Table 1. 1x HPE Apollo 4510 Gen10 as object storage servers

Quantity	Part number	Description
1	864668-B21	HPE Apollo 4510 Gen10 CTO Chassis
1	864625-B21	HPE XL450 Gen10 CT 1x Svr Node
1	882020-B21	HPE Apollo 4500 Gen10 CPU0 x2/CPU1 x2 FIO I/O module
1	872549-L21	HPE XL450 Gen10 4114 FIO Kit
1	872549-B21	HPE XL450 Gen10 4114 FIO Kit
6	815100-B21	HPE 32 GB 2Rx4 PC4-2666V-R Kit
1	825111-B21	HPE InfiniBand EDR/Ethernet 100Gb 2-port 840QSFP28 Adapter
2	JL271A	HPE x240 100G QSFP28 to QSFP28 1m Direct Attach Copper Cable
1	804331-B21	HPE Smart Array P408i-a SR Gen10 Ctrlr
1	874779-B21	HPE Apollo 4510 E208i-a/P408i-a Cable Kit
1	830824-B21	HPE Smart Array P408i-p SR Gen10 (8 Internal Lanes/2GB Cache)
1	874777-B21	HPE Apollo 4510 E208i-p/P408i-p Cable Kit
1	P01366-B21	HPE 96W Smart Storage Battery (up to 20 Devices/145mm Cable) Kit
2	655710-B21	HPE 1TB 6G SATA 7.2k 2.5in SC MDL HDD
1	877827-B21	HPE 3.2TB PCIe x8 MU HH DS Card
60	881787-B21	HPE 12TB 6G SATA 7.2K LFF 512e LP HDD
2	830272-B21	HPE 1600W FS Plat Ht Plg LH Pwr Sply Kit
1	878571-B21	HPE Apollo 4500 Chassis 4U Rail Kit

1x Apollo 4200 Gen10 System as block storage servers

Table 2. 1x Apollo 4200 Gen10 System as block storage servers

Quantity	Part number	Description
1	808027-B21	HPE Apollo 4200 Gen10 24LFF CTO Svr
1	806563-B21	HPE Apollo 4200 Gen10 LFF Rear HDD Cage Kit
1	830724-L21	HPE Apollo 4200 Gen10 Intel Xeon E5-2630v4 FIO Processor Kit
1	830724-B21	HPE Apollo 4200 Gen10 Intel Xeon E5-2630v4 Processor Kit
6	805349-B21	HPE 16 GB 1Rx4 PC4-2400T-R Kit
1	665243-B21	HPE Ethernet 10 Gb 2P 560FLR-SFP+ Adptr
1	813546-B21	HPE SAS Controller Mode for Rear Storage
2	797275-B21	HPE 1 TB 6G SATA 7.2k rpm LFF Low Profile Midline 1yr Warranty Hard Drive
1	846788-B21	HPE 1.6 TB 6G SATA Mixed Use-2 LFF 3.5-in LPC 3yr Wty Solid State Drive
4	867261-B21	HPE 8 TB 6G SATA 7.2K LFF LP 512e FIO HDD (Bundle)
1	806565-B21	HPE Apollo 4200 Gen10 IM Card Kit
1	806562-B21	HPE Apollo 4200 Gen10 Redundant Fan Kit
2	720479-B21	HPE 800W FS Plat Ht Plg Pwr Supply Kit
1	822731-B21	HPE Apollo 4200 Gen10 Hardware Rail Kit

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.

1x ProLiant DL360 Gen10 as infrastructure nodes

Table 3. 1x HPE ProLiant DL360 Gen10 as infrastructure nodes (Monitors, Managers, Gateways)

Quantity	Part number	Description
1	867959-B21	HPE ProLiant DL360 Gen10 4LFF Configure-to-order Server
1	860657-L21	HPE DL360 Gen10 Xeon-S 4110 FIO Kit
1	860657-B21	HPE DL360 Gen10 Xeon-S 4110 FIO
2	835955-B21	HPE 16 GB (1x16GB) Dual Rank x8 DDR4-2666
1	727055-B21	HPE Ethernet 10 Gb 2-port 562SFP+ Adapter
1	JD096C	HPE X240 10G SGP+ 1.2m DAC Cable
1	804326-B21	HPE Smart Array E208i-a SR Gen10 Controller
2	765453-B21	HPE 1 TB 6G SATA 7.2 K rpm Gen10 (3.5-inch) SC Midline 1yr Warranty Hard Drive
2	865408-B21	HPE 500 W FS Plat Ht Plg LH Power Supply Kit
1	789388-B21	HPE 1U Gen10 Easy Install Rail Kit

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

- **Ceph**—Open source software for a distributed object store with no single point of failure.
- **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.
- **Controlled Replication Under Scalable Hashing (CRUSH)**—CRUSH uses “rules” and Placement Group 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 Group, and the CRUSH rule set to use.
- **A Reliable, Autonomic Distributed Object Store (RADOS)**—This is the core set of SUSE Enterprise Server software that stores the user's data.
- **Representational State Transfer (REST)**—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."
- **Software-defined storage (SDS)**—A model for managing storage independently of hardware. Also typically includes user policies and may include advanced features like replication, deduplication, snapshots, and backup.
- **Swift**—object storage in the open source OpenStack project, used to build clusters of redundant, scalable, distributed object stores.

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 storage dense servers, visit: hpe.com/us/en/servers/hpc-apollo-4000.html.

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 here: SUSE Enterprise Storage Documentation.

Learn more at
h22168.www2.hpe.com/us/en/partners/suse/

Make the right purchase decision. Click here to chat with our presales specialists.

Share now

Get updates

4AA6-3911ENW, December 2018, Rev. 2