Reference Architecture

Storage

SUSE Enterprise Storage Reference Architecture for Cisco UCS

Table of Contents

SUSE Enterprise Storage™ solutions offer a single, unified software-defined storage cluster that provides unified object, block and file storage. Designed with unlimited scalability from terabytes to petabytes and no single point of failure, SUSE Enterprise Storage maximizes system resiliency and application availability in the event of unexpected hardware failures.

Included in this solution are storage-rich nodes such as the Cisco UCS S3260 Storage Server[1], a modular dual-node x86 server designed for investment protection and architectural flexibility to provide high performance or high capacity for your data-intensive workloads. Adding a few other system nodes of your choice from the Cisco UCS portfolio yields a storage cluster that is ideal for delivering multi-protocol file services to many applications or use cases like integration with OpenStack.

Tested Configuration

This reference implementation has been tested and validated through the Cisco Solution Partner IVT[2] program and is listed in the Cisco Technology Solutions Catalog[3].

Target Audience

This reference implementation is focused on administrators who deploy such a solution within their datacenter, making the file services accessible to various consumers. By following this document and those referenced herein, the administrator should have a complete view of the overall architecture, basic deployment and administrative tasks, with a specific set of recommendations for deployment on this hardware and networking platform.

Business Problem and Business Value

This SUSE Enterprise Storage solution delivers a highly scalable and resilient storage environment designed to scale from terabytes to petabytes. It can reduce IT costs with an intelligent, software-defined storage management solution that uses industry-standard servers and disk drives. It can also take advantage of the inherent features present in the platform. SUSE Enterprise Storage is able to seamlessly adapt to changing business and data demands. With its capability to automatically optimize, upgrade or add storage when needed, the solution is optimized for bulk and large data storage requirements.

1 **www.cisco.com/c/en/us/products/servers-unified-computing/ucs-s3260-storage-server/index.html**
2 **http://solutionpartnerdashboard.cisco.com/web/memberservices/testing**
3 **https://marketplace.cisco.com/catalog/companies/suse-llc/products/suse-enterprise-storage**

Business Problem

For most enterprise-level businesses, the demand for data storage is growing much faster than the rate at which the price for storage is shrinking. As a result, you could be forced to increase your budget dramatically to keep up with data demands.

Business Value

This intelligent software-defined storage solution—powered by Ceph[4] technology and incorporating feature-rich, manageable system hardware—enables you to transform your enterprise storage infrastructure to reduce costs while providing unlimited scalability to keep up with your future demands. With this completely tested solution, you will have the confidence to deploy a working solution and be able to maintain and scale it over time, without capacity-based increases in software subscriptions.

Requirements

While this reference implementation focuses on the initial setup and deployment of the storage solution, it also offers several other operational benefits in the overall lifecycle of the solution.

A SUSE Enterprise Storage solution is:

- *Simple to set up and deploy, within the documented guidelines for system hardware, networking and environmental prerequisites*
- *Adaptable to the physical and logical constraints needed by the business, both initially and as needed over time for performance, security or scalability concerns*
- *Resilient to changes in physical infrastructure components that are the result of failure or required maintenance*
- *Capable of providing optimized object and block services to client access nodes, either directly or through gateway services*

Architectural Overview

This reference implementation is intended to complement the SUSE Enterprise Storage Architectural Overview[5], which presents the concepts behind Software-Defined Storage and Ceph. You are encouraged to read and understand that overview first, since it provides suggested sizing information for the nodes providing each role.

Solution Architecture

SUSE Enterprise Storage provides unified block and object access based on Ceph, open source distributed storage software designed for scalability, reliability and performance. As opposed to conventional systems, which have allocation tables to store and fetch data, Ceph uses a pseudo-random data distribution function, which reduces the number of lookups required. In addition to the required network interfaces, switches and desired topology, the minimum Ceph storage cluster includes one Administration Server, a minimum of four object storage devices (OSD Nodes), three Monitor Nodes and one or more Ceph Object Gateways.

- *The Administration Server is used to deploy and configure SUSE Enterprise Storage on the other nodes that will be the OSDs, Monitor Nodes and Object Gateways. Calamari, the web-based interface for Ceph, can also be hosted on the Administration Server. In this implementation, a Cisco UCS C460 M4 Rack Server[6] was utilized as the Administration Server.*

---

4 **http://ceph.com**
5 **www.suse.com/docrep/documents/1mdg7eq2kz/suse_enterprise_storage_technical_overview_wp.pdf**
6 **www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c460-m4-rack-server/index.html**

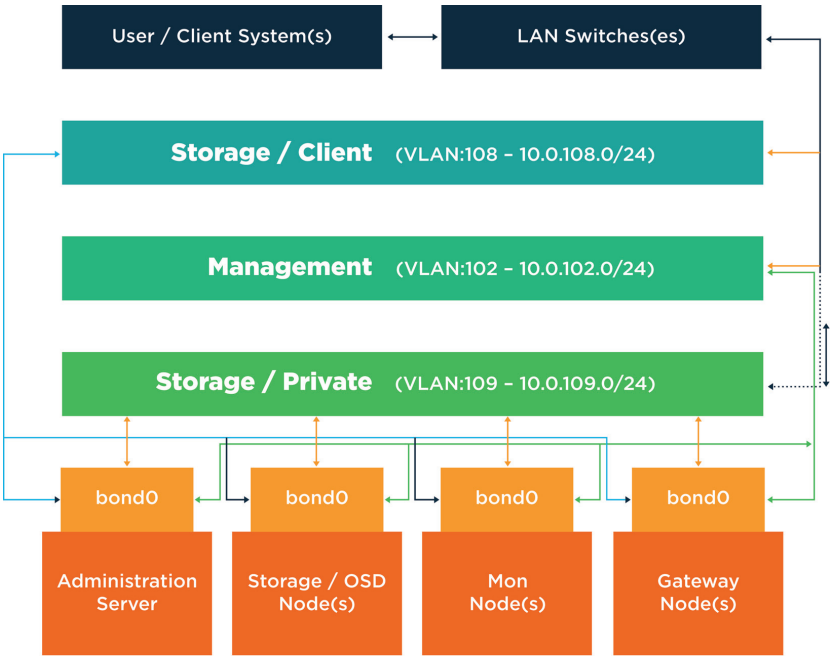
- *Data is stored on hosts containing one or more intelligent object storage devices (OSDs). OSDs automate data management tasks such as data distribution, data replication, failure detection and recovery using the CRUSH[7] algorithm. The CRUSH algorithm determines how to store and retrieve data by computing data storage locations. CRUSH empowers Ceph clients to communicate with OSDs directly, rather than through a centralized server or broker. For each of the four OSD Nodes, a Cisco UCS S3260 M4 Storage Server[8] was utilized.*
- *The monitors (MONs) maintain maps of the cluster state, including the monitor map, the OSD map, the Placement Group (PG) 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. For high availability, at least three monitors should be run in a Ceph cluster, with Cisco UCS C220 M4 Rack Servers[9] utilized for this role.*
- *A Ceph Object Gateway is an object storage interface that supports S3-compatible and Swift-compatible object storage functionality, using either the Amazon S3 or the OpenStack Swift RESTful APIs. Any SUSE YES Certified™ model of Cisco UCS Server that meets the requirement for the particular gateway function can be used for these roles.*

Networking Architecture

As with any software-defined solution, networking is a critical component. For a Ceph-based storage solution, there is a public or client-facing network and a private, backend network for the communication and replication by the OSD Nodes. In this implementation, all systems had at least two 10GbE NICs bonded together and utilized VLANs to segment the client and backend network traffic. The NICs were connected to Cisco Nexus 5000 Series Switches[10], configured with the corresponding VLANs.

The following diagram shows the logical layout of the network:

LOGICAL NETWORK DIAGRAM



7 **http://docs.ceph.com/docs/master/rados/operations/crush-map/**
8 **www.cisco.com/c/en/us/products/servers-unified-computing/ucs-s3260-storage-server/index.html**
9 **www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c220-m4-rack-server/index.html**
10 **www.cisco.com/c/en/us/products/switches/nexus-5000-series-switches/index.html**

Component Model

Note that the Server and each of the Nodes in the logical network diagram contain the respective Cisco UCS physical system platform and provide the software-defined storage solution with a couple layers of software from SUSE to deliver the various roles and services.

Component Overview

The following is a brief description of the components that make up the solution:

- **SUSE® Linux Enterprise Server 12 SP1**
  A world-class, secure, open source server operating system—built to power physical, virtual and cloud-based mission-critical workloads. SUSE Linux Enterprise Server is an infrastructure foundation that enables you to maximize service uptime, provide maximum security for all of your application needs and create a cost-effective infrastructure with the support of a wide range of hardware platforms.
- **SUSE Enterprise Storage 3**
  Provided as an add-on to SUSE Linux Enterprise Server, SUSE Enterprise Storage 3 combines the capabilities from the Ceph storage project with the enterprise engineering and support of SUSE. SUSE Enterprise Storage provides IT organizations with the ability to deploy a distributed storage architecture that can support a number of use cases, using industry-standard hardware platform.

Deployment

This reference implementation is a complement to the SUSE Enterprise Storage Deployment and Administration Guide[11]. As such, only notable design decisions, material choices and specific configuration changes that might differ from the published defaults in the deployment guide are highlighted in the remainder of this document.

Network Deployment Configuration (Suggested)

The following considerations for the network physical components should be attended to. Taking the time to complete these at the outset will help to prevent networking issues later and will facilitate any troubleshooting processes.

- *When racked, place identical models of each Top-of-Rack Cisco Nexus 5000 switch in two or more distinct racks to take advantage of different power distribution; this will allow the resiliency of the solution architecture to protect against potential outages.*
- *Ensure that all network switches are updated with consistent firmware versions.*
- *Configure 802.3ad for system port bonding and vLAG between the switches; also enable jumbo frames. It might be desirable to disable the spanning tree on the ports utilized for storage as well.*
- *Pre-plan 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 subnet mask to account for later growth. While storage traffic can be routed, it is generally discouraged to help ensure lower latency.*
- *Configure the desired subnets and VLANs to be utilized and configure the switch ports accordingly. See the Logical Network Diagram on the previous page for those used in this reference implementation.*
- *Properly and neatly cable and configure each node's port on the switches.*

---

11 **www.suse.com/documentation/ses-3/singlehtml/book_storage_admin/book_storage_admin.html**

Network Services Configuration

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

- *Ensure that you have access to a valid, reliable NTP service. This is a critical requirement in order for all nodes in the SUSE Enterprise Storage cluster to be maintained in time sync.*
- *Set up and reserve DNS A records for the storage nodes.*
- *Set up or use an existing SUSE Subscription Management Tool (SMT)[12] system. This service provides a local mirror of the SUSE product package repositories, allowing for rapid software deployment and later updating of packages.*
  - For larger environments, SUSE Manager[13] enables administrators to manage many Linux systems and keep them up to date, so SUSE Manager could be used in place of the SMT solution.

System Deployment Configuration

The following considerations for the system platforms should be attended to:

- *When racked, place identical models of each system platform or functional role in two or more distinct racks; this enables you to take advantage of different power distribution and networking paths and allows the resiliency of the solution architecture to protect against potential outages.*
- *If possible, set up Cisco UCS Manager service profile templates for each of the solution roles, Administration Server, OSD Nodes, Monitor Nodes and Gateway. This enables you to repeatedly and reliably set up the systems to a known state, including at least the following attributes:*
  - Confirm that BIOS/uEFI are updated with consistent versions across the same model, along with other firmware-based devices and components. For reference, check that the physical servers correspond to the versions (or later versions) listed on the SUSE YES certification for the Cisco platforms and respective SUSE Linux Enterprise Server 12 SP1 operating system. Specifically:
    - Cisco UCS-C460-M4 (SUSE YES Bulletin 144675[14])
    - Cisco UCS-C220-M4 (SUSE YES Bulletin 144340[15])
    - Cisco UCS-C3X60M4 (SUSE YES Bulletin 144909[16])
  - Ensure that Boot Mode is set to UEFI for all physical nodes that comprise the SUSE Enterprise Storage cluster.
  - On each system type, first configure a pair of SSDs into a RAID 1 LUN as the operating system deployment target volume.
- *To make later deployment tasks easier, consistently arrange the large number of remaining drives (by type and physical location) on each of the OSD Nodes. For each of the OSD Nodes, configure all non-operating system drives as RAID 0 / JBOD mode devices, with 3 SSDs targeted as a cache tier and 9 SSDs set aside as journal devices for the large number of HDDs.*

---

12 **www.suse.com/documentation/sles-12/book_smt/data/book_smt.html**
13 **www.suse.com/products/suse-manager/**
14 **www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=144675**
15 **www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=144340**
16 **www.suse.com/nbswebapp/yesBulletin.jsp?bulletinNumber=144909**

SW Deployment Configuration
Perform the following steps, in order, for this reference implementation:

- *Start by "Preparing Each Ceph Node" of the cluster by installing the SUSE Linux Enterprise Server operating system. Include only the minimal pattern and components, according to the procedure from the SUSE Enterprise Storage Deployment and Administration Guide*[17]*. This can be accomplished in any number of ways: using physical ISO media, with the virtual media option through Cisco UCS, or from a PXE network-boot environment. Use the pair of SSDs configured as a RAID 1 LUN, and the suggested default partitioning scheme, on each node as the target for the operating system. The SUSE Enterprise Storage extension can be installed either concurrently with the OS or post-installation as an add-on.*
  - The `ceph-deploy` command-line process will be used throughout the deployment, so please refer only to those procedures.
  - After installation of the operating system, ensure that each node has
    - *Access to the necessary software repositories, for later operations and updates. It is suggested that you apply all software updates, via* `zypper up`.
    - *NTP configured and operational, synchronizing with a source outside the cluster.*
    - *If necessary, adjust the udev rules to ensure that network interfaces are identified (as needed) in the same logical order across the systems, to make later steps easier. Ensure that the respective network interfaces are bonded together with the associated VLANs configured. While configuring these interfaces, it is also convenient to disable IPv6 functionality and the firewall on each node.*
      - For environments that require firewalls to be in place, refer to the "Firewall Settings for Ceph" section for detailed configuration settings.
    - *A cephadm user is configured with password-less SSH access from the Administration Server.*
- *The "Running ceph-deploy" section, and other portions noted below, will guide you through the steps to:*
  - Ensure that `ceph-deploy` is present on each node and distribute the necessary keyring and configuration files.
  - Set up each of the 3 Monitor Nodes and ensure that they are in a healthy, functional state via `ceph mon status`.
    - *An additional recommendation is to set up the Monitor Nodes as NTP peers.*
  - Set up each of the 4 OSD Nodes. Because SSD devices are present to act as journals, ensure that you invoke the ceph-deploy command with the appropriate syntax as you iterate across the devices to include this functionality:
    - `ceph-deploy osd prepare` *HOST:DISK[:JOURNAL]*
    - *Then check their status as you progress via* `ceph health`.
    - *The cache tier SSD devices can be set up by following the "Setting up and Example Tiered Storage" section. This entails the setup of a specific pool by adapting the* CRUSH *maps and the use of the* `ceph osd` *command line for pool and tier operations. In this reference implementation, the cache mode was set to write-back.*
  - Set up the Calamari Server, (using the "Install Calamari" section) and connect it to the cluster to help manage and monitor the nodes and services.
  - Set up an iSCSI Gateway (beginning with the "Ceph iSCSI Gateway" section) and also to set up a client to access the features of this SUSE Enterprise Storage cluster.
  - Use this simple set of steps to validate the overall cluster health:
    - *Set up or use existing client machines to access the cluster's block, object and iSCSI services and to view the Calamari dashboard.*

17 **www.suse.com/documentation/ses-3/singlehtml/book_storage_admin/book_storage_admin.html**

- *Within the cluster. perform some basic assessments of functionality*
  - `ceph health`
  - `ceph statusceph`
  - `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 via:*
  - `ceph osd pool delete test -yes-i-really-really-mean-it`

Deployment Considerations

As mentioned in the Functional Requirements section, there are considerations beyond the initial installation to take into account. SUSE Enterprise Storage is inherently resilient and can help protect from component failures and allow replacement of components over time or to address obsolescence or needed upgrades, as described in the SUSE Enterprise Storage Deployment and Administration Guide[18]. This includes:

- *Upgrading from Previous Releases*
- *Monitoring*
- *Recovery*
- *Maintenance*
- *Performance Diagnosis*

Deployment Considerations

Some additional considerations (with corresponding information in the SUSE Enterprise Storage Deployment and Administration Guide[19]) are important to highlight as well:

- *The "Best Practice" section contains a wealth of knowledge on diagnosis and tuning aspects.*
- *With the default replication setting of 3, remember that the client-facing network will have about half or less of the traffic of the backend network. This is especially true when component failures occur or rebalancing happens on the OSD Nodes. For this reason, it is important not to under-provision this critical cluster and service resource.*
- *It is important to maintain the minimum number of Monitor Nodes at three. As the cluster increases in size, it is best to increment it in pairs, keeping the total number of Monitor Nodes as an odd number. However, only really large or very distributed clusters would likely need beyond the three Monitor Nodes cited in this reference implementation. For performance reasons, it is recommended to use distinct nodes for the monitor roles, so that the OSD Nodes can be scaled as capacity requirements dictate.*

18 **www.suse.com/documentation/ses-3/singlehtml/book_storage_admin/book_storage_admin.html**
19 **www.suse.com/documentation/ses-3/singlehtml/book_storage_admin/book_storage_admin.html**

- *As described in this implementation, a minimum of four OSD Nodes is recommended, with the default replication setting of 3. This ensures that the cluster will continue working, without data loss, even with an outage of one of the OSD Nodes. There is no a practical or theoretical known limit of OSD Nodes within a cluster, subject to ensuring that each meets the minimum requirements. In fact, performance of the overall cluster increases as properly configured OSD Nodes are added.*
- *As noted in the deployment section, a cache tier was added to this reference implementation. This can actually be added any time during the cluster lifecycle, as performance needs dictate. Some workloads might not see improved performance with this cache tier, as noted in the "When to Use Cache Tiering" section.*

While the inclusion of iSCSI Gateways was built into this reference implementation, other types of gateway services are also available to deploy (such as RADOS), to provide Amazon S3 or Swift-compatible functionality and compliment the various needs of object and block storage use cases. And, depending on the actual storage usage, technologies like Erasure Coded Pools are available as well.

Appendix: Bill of Materials

Component / System / Software

Note: *Specific system platforms are not listed for the iSCSI Gateway nor the Calamari Server since any SUSE YES Certified model of Cisco UCS server will work. As a starting point, the same model used for the Monitor Nodes should be sufficient.*

Role	Quantity	Component	Notes
Top-of-Rack Network Switch	2	Cisco Nexus 5000	Add necessary network interconnect cables
Administration Server	1	Cisco UCS-C460-M4	4 x E7-480 v3 CPU 512 GB RAM 2 x 800GB Enterprise SSDs 6 x 1TB 10000 RPM HDDs 4 x 10GB NICs 2 x Power Supply
Monitor Node(s)	3	Cisco UCS-C220-M4	2 x E5-2650 v3 32GB RAM 2 x 800GB Enterprise SSDs 2 x 10GB NICs 2 x Power Supply
OSD Node(s)	4	Cisco UCS-C3260 M4	2 x E5-2697 v2 320 GB RAM Battery-backed, cache storage controller 2 x 800GB Enterprise SSDs 40 x 4TB 7200 RPM NL-SAS HDDs 12 x 400GB SAS SSDs 4 x 10GB NICs 4 x Power Supply
Software	1	SUSE Enterprise Storage Subscription	Includes 6 Infrastructure Nodes (for Administration Server, Monitor Nodes, Gateways), 4 OSD Nodes and limited use SUSE Linux Enterprise Server for 1-2 Socket Systems

Resources

Glossary

Ceph[20]: Open source software for a distributed object store with no single point of failure, which is the upstream software project that SUSE Enterprise Storage is based upon.

RADOS: A Reliable, Autonomic Distributed Object Store. This is the core set of SUSE Enterprise Server software that stores the user's data.

CRUSH[21]: Controlled Replication Under Scalable Hashing uses rules and placement groups to compute the location of objects deterministically in a SUSE Enterprise Server cluster.

Products

- *Cisco Nexus 5000 Series Switches*[22]
- *Cisco Unified Computing Products*[23]
- *SUSE Enterprise Storage*[24]

Component Model

Note that the Server and each of the Nodes in the logical network diagram contain the respective Cisco UCS physical system platform and provide the software-defined storage solution with a couple layers of software from SUSE to deliver the various roles and services.

---

20 **http://ceph.com**
21 **http://docs.ceph.com/docs/master/rados/operations/crush-map/**
22 **www.cisco.com/c/en/us/products/switches/nexus-5000-series-switches/index.html**
23 **www.cisco.com/c/en/us/products/servers-unified-computing/product-listing.html**
24 **www.suse.com/products/suse-enterprise-storage/**

Contact your local SUSE Solutions Provider, or call SUSE at:

1 800 796 3700 U.S./Canada
1 801 861 4500 Worldwide

SUSE
Maxfeldstrasse 5
90409 Nuremberg
Germany

www.suse.com