Dell EMC PowerEdge Server reference architecture for SUSE Enterprise Storage 5

Written by:

Edward Hayes-Hall, SUSE LLC

Kishore Gagrani, global product director, PowerEdge Emerging Technologies, Dell EMC

Table of Contents

Introduction

Dell EMC's PowerEdge Servers Offer Innovative Designs to Transform IT

The PowerEdge rack-based portfolio offers flexible designs to optimize your applications. The one-socket server portfolio provides balanced performance and storage capacity for future growth. The two-socket server portfolio brings a mix of features to maximize performance, scale for meet future demands and adapt to virtually any workload with an optimum balance of compute and memory. Dell EMC's four-socket server portfolio fills the top end with the highest performance and extensive scalability for your applications, from in-database workloads and HPC to data analytics, AI and GPU database acceleration.

SUSE Enterprise Storage (SES) offers 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.

This reference implementation has been tested and validated through the Dell Partner Lab, in Austin, TX, in collaboration with Dell EMC and SUSE teams.

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 that is 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. Dell EMC's PowerEdge servers offer multiple storage configurations, giving end users the ability to optimize for performance and capacity. This flexibility makes Dell EMC servers an optimum platform on which to implement 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 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.

Solution Architecture

SUSE Enterprise Storage provides unified, file, block and object storage based on Ceph, an open source software-defined distributed storage solution designed for scalability, reliability and performance. Unlike 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 device (OSD) nodes and three monitor (MON) nodes.

The solution includes storage-rich nodes of Dell EMC's PowerEdge R740 Rack Server, an x86 server designed for investment protection and architectural flexibility to provide high performance or high capacity for your data-intensive workloads. The infrastructure nodes are designed using the Dell EMC PowerEdge R640 Rack Servers. Together, SUSE Enterprise Storage and the Dell EMC PowerEdge Rack Server yields a storage cluster that is ideal for delivering unified storage servicing block, file and object protocols to many applications or use cases—such as integration with SUSE Containers as a Service Platform, Disk to Disk Backup and Archive solutions.

Ceph supports both native and traditional client access. The native clients are aware of the storage topology and communicate directly with the storage daemons, resulting in horizontally scaling performance. Non-native protocols, such as ISCSI, S3 and NFS, require the use of gateways. While these gateways might be considered a limiting factor, the ISCSI and S3 gateways can scale horizontally using load balancing techniques.

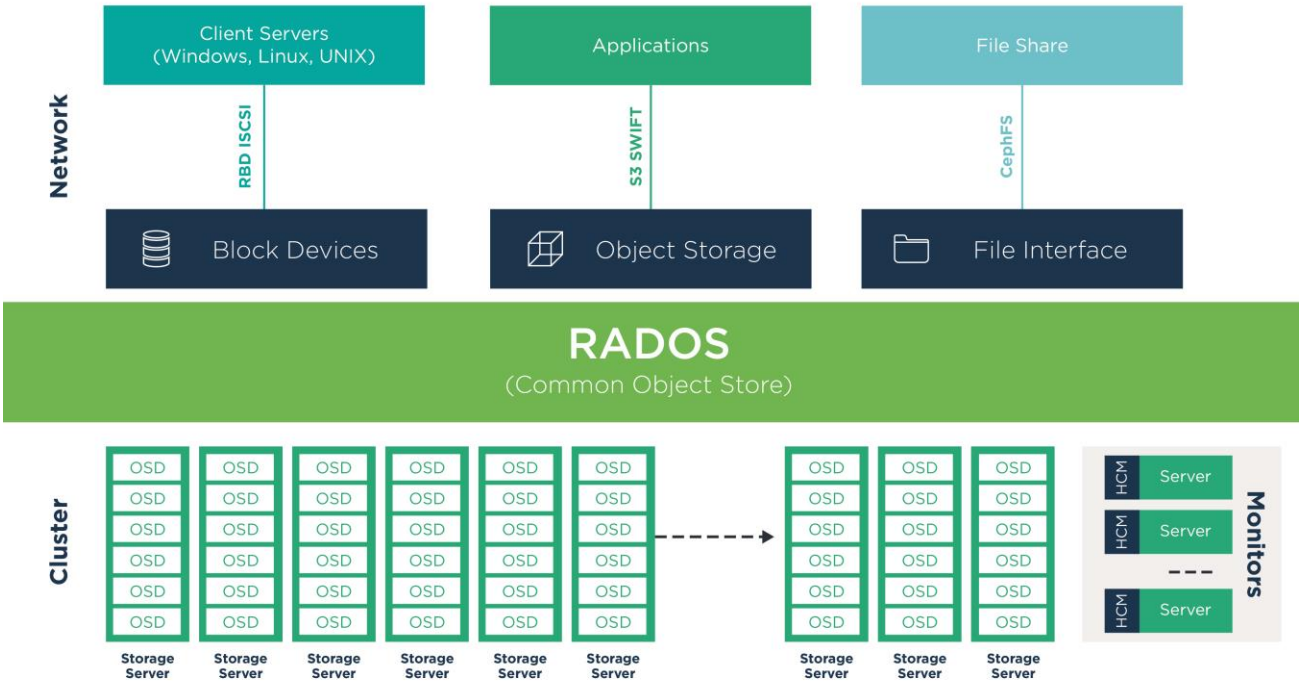

Figure 1: Ceph architecture diagram

Solution Admin Host

Due to the need for various administrative-like services, the team decided to create a Solution Admin Host (SAH) that consolidates services that would typically be found in an enterprise environment. Given a finite number of physical systems, this consolidation helps to preserve other system nodes for more resource-intensive uses by deploying virtual machine guests for various administrative functions.

Design Decision: A simple hypervisor host, using KVM, provides the platform for the SAH and enables grouping of administrative functions (such as the SUSE Subscription Management Tool (SMT), DNS and other services) as virtual machines. SMT enables you to provision updates for all of your devices running a product based on SUSE Linux Enterprise. By downloading these updates once and distributing them throughout the enterprise, you can set more restrictive firewall policies. This also reduces bandwidth usage because there is no need to download the same updates for each device.

Note: *Avoid running the NTP service in a virtualized instance.*

Process: *Using one of the available Dell PowerEdge R640 Rack Systems, perform a bare-metal installation of the SUSE Linux Enterprise Server 12 SP3 operating system with either physical media or virtual media through iDRAC.*

Note: *The default partitioning scheme can be used, but remember to store any virtual machine images into the larger home directory partition.*

- *A minimal system can be installed, with at least the following patterns included: base, minimal, kvm_server, kvm_tools.*
- *Register the system in the SUSE Customer Center (SCC) during or after the installation to ensure that all the latest software updates are present.*
- *After the installation completes, use YaST® to configure the desired networking, including:*
    - *An external network interface for access from beyond the storage environment, using one or both of the 1GbE NICs (e.g., em3, em4). This is because the SMT server needs to access the Internet to download updates, whereas the 25GbE should be isolated to storage traffic.*
    - *A bond (mode 802.3ad to match the switch configuration) across all 25GbE NICs being used (e.g., p3p1, p2p1) to provide the necessary services to the SUSE Enterprise Storage cluster nodes.*
    - *A bridge for virtualization (in addition to the previously bonded network interfaces), configured with an IP address in the public network (for 25GbE) and for internet access (for 1GbE).*
- *Ensure that an Administrative VNC server is set up and running to remotely access this system from other systems, which provides a graphical user interface.*
- *Ensure that the system is configured to have a common Network Time Protocol (NTP) source, since this will become a reference for the storage resource nodes as well. For the purposes of this exercise, all nodes were configured to sync with a router connected to the cluster.*

Node Roles

The Administration Server that is used to deploy and configure SUSE Enterprise Storage via Salt on all the cluster nodes should be installed on bare metal. Additionally, the web-based interface for Ceph, openATTIC, should also be hosted on the Administration Server.

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 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 physical Dell EMC PowerEdge R740 Rack Server was utilized.

The monitors (MONs) maintain maps of the cluster state, including the monitor map, the OSD map, the Placement Group (PG) 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 three Dell EMC PowerEdge R640 Rack Servers utilized for this role.

The various types of gateway nodes and MDS nodes utilized for various protocol translations complete the cluster roles. While none were specifically used during this effort, gateways for S3/Swift, ISCSI and NFS can be utilized to add protocol interfaces. MDS nodes (the metadata nodes required for CephFS) are also a key component for many SUSE Enterprise Storage deployments. All of these roles are easily provided by Dell-EMC PowerEdge R640 class systems.

Networking Architecture

As with any software-defined solution, networking is a critical component. For a Ceph-based storage solution, there can be a public or client-facing network and a private, backend network for communication and replication by the OSD Nodes. In this implementation, all four OSD nodes utilized two 25GbE NICs bonded together to provide both the public and private network functionality. Clients used for testing were connected to the public network.

Although the monitor nodes have a much lighter network traffic requirement, in order to provide lower latency and avoid potential issues with network timing, all physical systems leveraged two 25GbE NICs bonded to connect to the public network.

Network/IP Address Scheme

Specific to this implementation, shown in Table 1, the following naming and addressing scheme was utilized:

Function	Hostname	Public Network	iDRAC Network
Solution Admin Host (SAH)	sah.suse.xyz	10.204.92.86	10.204.92.135
SUSE Enterprise Storage Admin	admin.suse.xyz	10.204.92.155	10.204.92.163
SMT	smt.suse.xyz	10.204.92.125	N/A
Monitor	mon1.suse.xyz	10.204.92.85	10.204.92.164
Monitor	mon2.suse.xyz	10.204.92.86	10.204.92.165
Monitor	mon3.suse.xyz	10.204.92.87	10.204.92.166
OSD Node	osd1.suse.xyz	10.204.92.221	10.204.92.101
OSD Node	osd2.suse.xyz	10.204.92.222	10.204.92.102
OSD Node	osd3.suse.xyz	10.204.92.223	10.204.92.103
OSD Node	osd4.suse.xyz	10.204.92.224	10.204.92.104

Table 1: Network/IP address scheme

Hardware Configuration

PowerEdge rack servers offer storage flexibility, capacity and performance optimization. The reference section of this document includes a list of technical architecture to help you understand some salient features of the PowerEdge storage sub-systems' capabilities. While not all of these features were tested during this reference architecture, they should be considered while deploying your own production environment (e.g., using NVMe drives for the Cache Tier).

A complete BOM of the PowerEdge configuration that was used in this paper's testing is included in Appendix C.

Dell EMC Networking S5248F 25GbE switch
Dell EMC S4112-ON10GbE switch
Dell EMC PowerEdge R640 Rack Server
Dell EMC PowerEdge R740 Rack Server

Component Model

Component Overview (SUSE)

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

SUSE® Linux Enterprise Server 12 SP3: 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 5: Provided as an add-on to SUSE Linux Enterprise Server, SUSE Enterprise Storage 5 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 an industry-standard hardware platform.

Component Overview (Dell EMC)

Note: Specific system platforms are not listed for the iSCSI Gateway nor the openATTIC Server, since any SUSE YES Certified model of DELL EMC PowerEdge 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 Switch	2	Dell S5248S Dell S4112-ON	Add necessary network interconnect cables
Administration Server	1	Dell EMC PowerEdge R640	128GB RAM 2x 240GB SATA using RAID-1 2x Intel(R) Xeon(R) Gold 5115 CPU @ 2.40GHz Broadcom 57414 Dual Port 25Gb NIC

Monitor Node(s)	3	Dell EMC PowerEdge R640	128GB RAM 2x 240GB SATA SSD using RAID-1 2x 1TB SATA HDD using RAID-1 2x Intel(R) Xeon(R) Gold 5115 CPU @ 2.40GHz Broadcom 57414 Dual Port 25Gb NIC
OSD Nodes	4	Dell EMC PowerEdge R740	128GB RAM 2 240GB SATA using RAID-1 2x Intel(R) Xeon(R) Gold 5115 CPU @ 2.40GHz 2x Broadcom 57414 Dual Port 25Gb NIC 10x 8TB SATA
Software	1	SUSE Enterprise Storage Subscription	Subscription includes: 6 Infrastructure Nodes (for Administration Server, Monitor Nodes and/or Gateways), 4 OSD Nodes and limited use SUSE Linux Enterprise Server for 1-2 Socket Systems

Table 2: Component overview

Deployment

This reference implementation is a complement to the SUSE Enterprise Storage Deployment and Administration Guides. 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 (Recommended)

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 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.
- 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 might 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 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. Use the “ip a” command to confirm that network configuration and bonding is correct.
- Properly and neatly cable and configure each node’s port on the switches.

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. Use the following on all nodes to sync time:

```
sudo service ntp stop
sudo ntpd -gq
sudo service ntp start
```

- In a production environment, the NTP configuration should reference two or more servers to ensure availability to time synchronization service. It is also necessary to ensure that the NTP server does not reside on virtualized infrastructure.
- Set up or use an existing SUSE Subscription Management Tool (SMT) system. This service provides a local mirror of the SUSE product package repositories, allowing for rapid software deployment and later updating of packages. For the reference environment, this was provided on the SAH host described earlier in this document.

Hardware Deployment Configuration

The following information considerations for the system platforms should be attended to.

When racked, distribute services as evenly as possible across the racks. For example, place one mon node in each rack and two OSD nodes in the first rack. Then distribute the remaining two OSD nodes with one in each remaining rack. 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.

Confirm that the BIOS is updated with consistent versions across the same model, along with other firmware-based devices and components. Ensure that all nodes in the cluster have the BIOS set to performance optimization settings. For reference, check that the physical servers correspond to the versions (or later versions) listed on the SUSE YES certification for the Dell platforms and respective SUSE Linux Enterprise Server 12 SP3 operating system. Specifically: Distribute the drives (by type and physical location) equally to the OSD nodes, creating a symmetric configuration. For each of the OSD Nodes, configure all non-operating system drives as RAID 0 / JBOD mode devices. For each of the OSD Nodes, it is recommended to configure 2 SSDs for Write-Ahead Log (WAL) and RocksDB purposes for the data residing on the non-OS system HDDs. In this reference implementation, the cache mode for the PowerEdge RAID Controller (PERC) was set to write-back, as this provides some benefits for small I/O operations on the cluster.

All nodes should have the operating system (OS) deployed on a mirrored pair of SSDs whose total size is greater than the memory in the host. This allows for a full memory dump to be saved, should the need occur. The use of SSDs also lowers the power and thermal footprint of the system. These could be installed in the mid-plane or backplane drive cages. Additionally, the R740XD supports the use of M.2 media for boot, which is a valid option as well.

Software Deployment Configuration

Perform the following steps, in the order shown, for this reference implementation.

Start by preparing each node of the cluster by installing the SUSE Linux Enterprise Server operating system. Include only the minimal components, according to the procedure from the SUSE Enterprise Storage Deployment Guide. This can be accomplished in any number of ways: using physical ISO media, with the virtual media option through Dell iDRAC, or from a PXE network-boot environment. The SUSE Enterprise Storage extension can be installed either concurrently with the OS or post-installation as an add-on. SUSE Enterprise Linux and SUSE Enterprise Storage are installed with the root user id.

After installation of the operating system, ensure that each node has the following:

- Access to the necessary software repositories on the SMT server for later operations and updates. zypper lr can be used to list the repositories configured on each server.
- Each node is registered to access necessary updates and apply those software updates, via zypper up.
- NTP is configured and operational, synchronizing with a source outside the cluster.
- ssh is operating correctly.

Salt, along with DeepSea, is a stack of components that help to deploy and manage server infrastructure. It is very scalable and fast and is relatively easy to get running.

The “Deploying using DeepSea” section of the SUSE Enterprise Storage Deployment Guide will guide you through the steps to deploy SUSE Enterprise storage.

These three key Salt imperatives must need to be followed and are described in detail in section 4. Deploying with DeepSea and Salt:

- The Salt master is the host that controls the entire cluster deployment. Ceph itself should NOT be running on the master as all resources should be dedicated to Salt. In our scenario, we used the Admin host as the Salt master.
- Salt minions are nodes controlled by the Salt master. OSD and monitor nodes are all Salt minions in this installation. Salt minions need to correctly resolve the Salt master’s hostname over the network. This can be achieved through configuring unique host names per interface in DNS and/or local /etc/hosts files.
- DeepSea consists of a series of Salt files to automate the deployment and management of a Ceph cluster. It consolidates the administrator’s decision-making in a single location around cluster assignment, role assignment and profile assignment. DeepSea collects each set of tasks into a goal or stage.

Then follow the process steps from the Cluster Deployment section of the Deployment Guide.

To complete the setup, execute stage. Zero and stage. One of the five-step deployment process on the Administration Server. A default proposal is generated by stage. One and is used for this configuration. For information on offloading RocksDB and WAL, see Appendix A.

The openATTIC Server (SUSE Enterprise Storage Administrative Interface) is deployed as part of stage four. Connect to the GUI by pointing a browser to the Administration Server. Log in with the default User ID and Password (for details, see SUSE Enterprise Storage Administration – Managing Cluster with GUI Tools – openATTIC Deployment and Configuration) to help manage and monitor the nodes and services.

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 openATTIC dashboard.
- Within the cluster, perform some basic assessments of functionality.

ceph health

```
admin:~ # ceph health
HEALTH_OK
admin:~ #
```

ceph status

```
admin:~ #
admin:~ # ceph status
  cluster:
    id:     9116f500-f2f2-34ee-9e87-b75a8e1b4ea1
    health: HEALTH_OK

  services:
    mon: 3 daemons, quorum mon3,mon1,mon2
    mgr: mon3(active), standbys: mon2, mon1
    osd: 24 osds: 24 up, 24 in

  data:
    pools:   0 pools, 0 pgs
    objects: 0 objects, 0 bytes
    usage:   24897 MB used, 22329 GB / 22353 GB avail
    pgs:

admin:~ #
```

ceph osd df

```
admin:~ # ceph osd df
ID CLASS WEIGHT  REWEIGHT SIZE   USE    AVAIL  %USE VAR  PGS
 5   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 140
 7   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 118
12   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 145
15   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 133
17   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 140
21   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 124
 1   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 123
 2   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 143
 6   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 145
 9   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 126
11   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 116
14   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 125
 0   hdd 0.90959  1.00000   931G  1048M   930G 0.11 1.01 123
 3   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 112
 4   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 120
16   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 124
22   hdd 0.90959  1.00000   931G  1040M   930G 0.11 1.00 125
23   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 136
 8   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 123
10   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 121
13   hdd 0.90959  1.00000   931G  1038M   930G 0.11 1.00   0
18   hdd 0.90959  1.00000   931G  1048M   930G 0.11 1.01 132
19   hdd 0.90959  1.00000   931G  1044M   930G 0.11 1.00 118
20   hdd 0.90959  1.00000   931G  1048M   930G 0.11 1.01 132
                    TOTAL 22353G 25022M 22329G 0.11
MIN/MAX VAR: 1.00/1.01  STDDEV: 0
admin:~ #
```

The following commands allow a short performance test to create and exercise a storage pool:

```
ceph osd pool create test 1024

rados bench -p test 300 write -no-cleanup

rados bench -p test 300 seq
```

Once the test has completed, the created pool should be deleted, using the following commands:

```
ceph tell mon* injectargs -mon_allow_pool_delete=true

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

ceph tell mon* injectargs -mon_allow_pool_delete=false
```

Summary

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

- Upgrading from previous releases
- Monitoring
- Recovery
- Maintenance
- Performance diagnosis

Deployment Considerations

These additional considerations (with corresponding information in the SUSE Enterprise Storage Administration Guide) are important to highlight as well.

Section 18 “FAQ, Tips and Troubleshooting” of the Administration Guide contains a wealth of information 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 when 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 very large or very distributed clusters would likely need beyond the three Monitor nodes cited in this reference implementation. For

performance reasons, it is recommended to use distinct nodes for the monitor roles. This ensures that OSD performance is not impacted by other services running on the same node.

As described in this implementation, a minimum of four OSD Nodes is required, 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 practical or theoretical known limit of OSD Nodes within a cluster, subject to ensuring that each meets the minimum requirements. In fact, the performance of the overall cluster increases as properly configured OSD Nodes are added.

Resources

SUSE Enterprise Storage 5 Administration Guide

SUSE Enterprise Storage 5 Deployment Guide

DELL PowerEdge R640 Spec Sheet

DELL PowerEdge R740 Spec Sheet

Dell S5048F-ON 25GbE switch Data Sheet

Dell S4820T 10GbE switch Data Sheet

Appendix A: Recommended OSD Drive and Journal Proposal Changes

The DeepSea deployment tool uses a profile to define the disk usage and layout. It is generally recommended that the RocksDB and Write Ahead Log be relocated to SSD or NVME devices to improve performance.

An example proposal for using a 480GB SSD in a 5:1 ratio of spinners to SSD is generated by executing:

```
salt-run proposal.populate name=740xd ratio=5 wal=440-490 db=440-490 target='osd*' db-size=60g
wal-size=2g data=7000-9000
```

Where:

- `ratio` is the ratio of SSDs to HDDs in each OSD node
- `wal` is the size range of the device to use for the Write Ahead Log partition
- `wal-size` is the size of individual partitions for the Write Ahead Logs
- `db` is the size range of the device to use for the RocksDB partition
- `db-size` is the size of individual partitions for RocksDB
- `name` is the Ceph Proposal name
- `target` is the list of the target OSDs hostnames
- `data` is the size of the device to use for the bulk of OSD storage

The contents of the proposal can be found in /srv/pillar/ceph/proposals/profile-740xd/stack/default/ceph/minions. Each file contains the disk devicesx with db and wal partitions clearly identified on separate devices.

Appendix B: Policy.cfg

Located in `/srv/pillar/ceph/proposals` on the Administration Server

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

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

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

## Roles
# ADMIN
role-master/cluster/admin*.sls
#
role-admin/cluster/admin*.sls
# MGR
role-mon/cluster/mon*.sls
# MON
role-mgr/cluster/mon*.sls
# openATTIC
role-openattic/cluster/admin*.sls
```

Appendix C: Component Bill of Materials

Solution Admin Host – PowerEdge R640

Description	Quantity	SKU
PowerEdge R640 Server	1	210-AKWU
PowerEdge R640 Motherboard	1	329-BDKC
Trusted Platform Module 2.0	1	461-AAEM
2.5 Chassis with up to 10 Hard Drives , 2x2.5" SATA/SAS Drives and 2PCIe slots, Dual Controller	1	321-BDKF
PowerEdge R640 Shipping	1	340-BKNE
PowerEdge R640 x4 and x10 Drive Shipping Material	1	340-BLUC
Intel Xeon Gold 5115 2.4G, 10C/20T, 10.4GT/s , 14M Cache, Turbo, HT (85W) DDR4-2400	1	338-BLUU
Intel Xeon Gold 5115 2.4G, 10C/20T, 10.4GT/s , 14M Cache, Turbo, HT (85W) DDR4-2400	1	374-BBPR
DIMM Blanks for System with 2 Processors	1	370-ABWE
Standard 1U Heatsink	2	412-AAIQ
2666MT/s RDIMMs	1	370-ADNU
Performance Optimized	1	370-AAIP
32GB RDIMM 2666MT/s Dual Rank	4	370-ADNF
No RAID	1	780-BCDI
HBA330 12Gbps SAS HBA Controller (NON-RAID), Minicard	1	405-AAJU
PERC H330 RAID Controller, Adapter, Low Profile	1	405-AANP
1TB 7.2K RPM SATA 6Gbps 512n 2.5in Hot-plug Hard Drive	2	400-ASHF
240GB SSD SATA Read Intensive 6Gbps 512 2.5in Flex Bay Drive, 1 DWPD,438 TBW	2	400-AWHG
No Operating System	1	619-ABVR
No Media Required	1	421-5736
iDRAC9,Enterprise	1	385-BBKT
OME Server Configuration Management	1	528-BBWT
iDRAC Group Manager, Enabled	1	379-BCQV

iDRAC,Legacy Password	1	379-BCSG
Riser Config 1, 1x16 LP	1	330-BBGX
Dell Networking, Transceiver, SFP28 25GbE SR, 85c, MMF Duplex, LC	2	407-BBZT
Broadcom 57414 Dual Port 25Gb, SFP28, rNDC	1	540-BBUM
No Internal Optical Drive	1	429-AAIQ
8 Standard Fans for R640	1	384-BBQJ
Dual, Hot-plug, Redundant Power Supply (1+1), 750W	1	450-ADWS
NEMA 5-15P to C13 Wall Plug, 125 Volt, 15 AMP, 10 Feet (3m), Power Cord, North America	2	450-AALV
No Bezel	1	350-BBBW
Dell EMC Luggage Tag for x10	1	350-BBJT
Quick Sync 2 (At-the-box mgmt)	1	350-BBKC
Power Saving Dell Active Power Controller	1	750-AABF
UEFI BIOS Boot Mode with GPT Partition	1	800-BBDM
Energy Star	1	387-BBMK
ReadyRails Sliding Rails With Cable Management Arm	1	770-BBBL
No Systems Documentation, No OpenManage DVD Kit	1	631-AACK
US Order	1	332-1286
Basic Hardware Services: Business Hours (5x10) Next Business Day On-Site Hardware Warranty Repair, 3 Years	1	813-9254
Dell Hardware Limited Warranty Plus On-Site Service	1	813-9255
On-Site Installation Declined	1	900-9997

OSD Nodes – PowerEdge R740XD

Description	Quantity	SKU
PowerEdge R740XD Server	4	210-AKZR
PowerEdge R740/R740XD Motherboard	4	329-BDKH
Trusted Platform Module 2.0	4	461-AAEM

Chassis with up to 12x3.5" HDDs on BP, No Mid-Bay and 2x3.5" HDDs Flexbay, 1 or 2CPU Config	4	321-BDSJ
PowerEdge R740XD Shipping	4	340-BLBE
PowerEdge R740 Shipping Material	4	343-BBFU
Intel Xeon Gold 5115 2.4G, 10C/20T, 10.4GT/s , 14M Cache, Turbo, HT (85W) DDR4-2400	4	338-BLUU
Intel Xeon Gold 5115 2.4G, 10C/20T, 10.4GT/s , 14M Cache, Turbo, HT (85W) DDR4-2400	4	374-BBPR
Standard 1U Heatsink	8	412-AAIQ
2666MT/s RDIMMs	4	370-ADNU
Performance Optimized	4	370-AAIP
32GB RDIMM 2666MT/s Dual Rank	16	370-ADNF
No RAID	4	780-BCDI
PERC H730P RAID Controller, 2GB NV Cache, Mini card	4	405-AAQU
8TB 7.2K RPM SATA 6Gbps 512e 3.5in Hot-plug Hard Drive	40	400-ASIF
240GB SSD SATA Read Intensive 6Gbps 512 2.5in Flex Bay Drive,3.5in HYB CARR, 1 DWPD,438 TBW	8	400-AWHH
No Operating System	4	619-ABVR
No Media Required	4	421-5736
iDRAC9,Enterprise	4	385-BBKT
OME Server Configuration Management	4	528-BBWT
iDRAC Group Manager, Enabled	4	379-BCQV
iDRAC, Factory Generated Password	4	379-BCSF
Riser Config 2, 3 x8, 1 x16 slots	4	330-BBHG
Dell Networking, Transceiver, SFP28 25GbE SR, 85c, MMF Duplex, LC	8	407-BBZT
Broadcom 57414 Dual Port 25Gb, SFP28, rNDC	4	540-BBUM
6 Performance Fans forR740/740XD	4	384-BBPZ
Dual, Hot-plug, Redundant Power Supply (1+1), 750W	4	450-ADWS

NEMA 5-15P to C13 Wall Plug, 125 Volt, 15 AMP, 10 Feet (3m), Power Cord, North America	8	450-AALV
PowerEdge 2U Standard Bezel	4	325-BCHU
PE R740XD Luggage Tag	4	389-BTTO
Quick Sync 2 (At-the-box mgmt)	4	350-BBJU
Performance BIOS Settings	4	384-BBBL
UEFI BIOS Boot Mode with GPT Partition	4	800-BBDM
ReadyRails Sliding Rails With Cable Management Arm	4	770-BBBR
No Systems Documentation, No OpenManage DVD Kit	4	631-AACK
US Order	4	332-1286
Basic Hardware Services: Business Hours (5x10) Next Business Day On-Site Hardware Warranty Repair, 3 Years	4	813-6067
Dell Hardware Limited Warranty Plus On-Site Service	4	813-6068
On-Site Installation Declined	4	900-9997
Declined Remote Consulting Service	4	973-2426

Ceph Infrastructure Nodes (Mon, Admin, Gateways) – PowerEdge R640

Description	Quantity	Sku
PowerEdge R640 Server	3	210-AKWU
PowerEdge R640 Motherboard	3	329-BDKC
Trusted Platform Module 2.0	3	461-AAEM
2.5 Chassis with up to 10 Hard Drives , 2x2.5" SATA/SAS Drives and 2PCIe slots, Dual Controller	3	321-BDKF
PowerEdge R640 Shipping	3	340-BKNE
PowerEdge R640 x4 and x10 Drive Shipping Material	3	340-BLUC
Intel Xeon Gold 5115 2.4G, 10C/20T, 10.4GT/s , 14M Cache, Turbo, HT (85W) DDR4-2400	3	338-BLUU
Intel Xeon Gold 5115 2.4G, 10C/20T, 10.4GT/s , 14M Cache, Turbo, HT (85W) DDR4-2400	3	374-BBPR
DIMM Blanks for System with 2 Processors	3	370-ABWE

Standard 1U Heatsink	6	412-AAIQ
2666MT/s RDIMMs	3	370-ADNU
Performance Optimized	3	370-AAIP
32GB RDIMM 2666MT/s Dual Rank	12	370-ADNF
No RAID	3	780-BCDI
HBA330 12Gbps SAS HBA Controller (NON-RAID), Minicard	3	405-AAJU
PERC H330 RAID Controller, Adapter, Low Profile	3	405-AANP
240GB SSD SATA Read Intensive 6Gbps 512 2.5in Flex Bay Drive, 1 DWPD,438 TBW	6	400-AWHG
No Operating System	3	619-ABVR
No Media Required	3	421-5736
iDRAC9,Enterprise	3	385-BBKT
OME Server Configuration Management	3	528-BBWT
iDRAC Group Manager, Enabled	3	379-BCQV
iDRAC,Legacy Password	3	379-BCSG
Riser Config 1, 1x16 LP	3	330-BBGX
Dell Networking, Transceiver, SFP28 25GbE SR, 85c, MMF Duplex, LC	6	407-BBZT
Broadcom 57414 Dual Port 25Gb, SFP28, rNDC	3	540-BBUM
No Internal Optical Drive	3	429-AAIQ
8 Standard Fans for R640	3	384-BBQJ
Dual, Hot-plug, Redundant Power Supply (1+1), 750W	3	450-ADWS
NEMA 5-15P to C13 Wall Plug, 125 Volt, 15 AMP, 10 Feet (3m), Power Cord, North America	6	450-AALV
No Bezel	3	350-BBBW
Dell EMC Luggage Tag for x10	3	350-BBJT
Quick Sync 2 (At-the-box mgmt)	3	350-BBKC
Power Saving Dell Active Power Controller	3	750-AABF
UEFI BIOS Boot Mode with GPT Partition	3	800-BBDM
Energy Star	3	387-BBMK

ReadyRails Sliding Rails With Cable Management Arm	3	770-BBBL
No Systems Documentation, No OpenManage DVD Kit	3	631-AACK
US Order	3	332-1286
Basic Hardware Services: Business Hours (5x10) Next Business Day On-Site Hardware Warranty Repair, 3 Years	3	813-9254
Dell Hardware Limited Warranty Plus On-Site Service	3	813-9255
On-Site Installation Declined	3	900-9997

References

- Tech Note provides an overview of PowerEdge hard drive technologies, including form factor, rotational speed, sector format and bus interface. These factors can be used in conjunction with trade-offs between performance, capacity and budget to help users select the appropriate HDD.
- Point of View, jointly written by Dell EMC and Brocade, explains that upgrading IT infrastructures with new PowerEdge servers and flash storage can deliver significant improvements to workload performance. However, to fully realize the benefits of these upgrades and to avoid performance bottlenecks, the network connecting servers to storage should also be upgraded.
- The Dell EMC PowerEdge R640 Unique NVMe Implementation, July 2017. This Tech Note explains how the new PowerEdge R640 takes advantage of an increase in PCI lanes to optimize the performance of up to 8 NVMe drives. This implementation lowers latency, decreases power consumption and reduces the cost for workloads needing up to 8 NVMe drives.