Reference Architecture: Red Hat Hyperconverged Infrastructure for Virtualization

Last update: 3 May 2019

Version 1.0

Describes the reference architecture for Red Hat Hyperconverged Infrastructure

Describes Lenovo ThinkSystem servers and network switches

Provides detailed networking configuration information

Includes validated and tested deployment information including bill of materials

Feng Xia

Miroslav Halas

Mike Perks

Table of Contents

1 Introduction

The intended audience for this document is technical IT architects, system administrators, and managers who are interested in executing workloads on the Red Hat Hyperconverged Infrastructure for Virtualization platform. This platform provides a hyper-converged infrastructure based on Red Hat Enterprise Linux. Some experience with Lenovo server and switch configurations is highly desirable. Experience with Red Hat implementation of virtualization and Red Hat Enterprise Linux (RHEL) version 7.5 is helpful, but it is not required.

Hyper-converged means incorporating multiple components like compute and storage into a single entity through software. A hyper-converged infrastructure seamlessly pools compute and storage to deliver high performance for the virtual workloads and provides flexibility to combine the local storage using a distributed file system to eliminate shared storage such as SAN or NAS. These factors make the solution cost effective without compromising the performance.

This document provides an overview of the business problem that is addressed by Red Hat Hyperconverged Infrastructure for Virtualization and the business value that is provided by such a hyperconverged platform. A description of customer requirements is followed by an architectural overview of the solution and a description of the logical components. The operational model describes the architecture for deploying Red Hat Hyperconverged Infrastructure for Virtualization into an Enterprise environment and then discusses some Lenovo tooling that helps bootstrap an installation of Red Hat Hyperconverged Infrastructure for Virtualization into a customer environment. The last chapter features a detailed Bill of Materials for the Lenovo ThinkSystem hardware that are used in the solution.

See also the Red Hat OpenStack Reference Architecture: https://lenovopress.com/lp0762.

2 Business problem and business value

2.1 Business problem

While software continues to be deployed on either physical servers or in virtual machine, container based deployment is gaining ground. However, virtual machines is a well-established technology that still has advantages in areas such as security, backwards compatibility, technology maturity, support, and operational experience.

Within the virtualization sphere, the Linux kernel-based virtual machine (as known as KVM) is a popular hypervisor because it is open source and production ready. However, there must be additional infrastructure that provides an adequate storage backend, networking, and management.

Building such a virtualized solution from scratch is not an easy task. There are many moving parts, each with lots of options, and many opportunities for things to go wrong. Customers also want to deploy multiple environments such as development, pre-production, and production. It is important to keep these environments synchronized; otherwise, there could be unexpected problems when moving from development to production. However, the needs for an agile, small footprint, experimental development environment are completely different to stable, secure, and scalable production environment.

Lastly, there is whole area of the hardware needed for a virtualized environment, understanding all of the various parts that are needed from drives and cables to power consumption, maintaining firmware, driver and Operating System code versions, and management of the whole hardware infrastructure.

2.2 Business value

This reference architecture describes using Red Hat Hyperconverged Infrastructure for Virtualization with Lenovo ThinkSystem severs and switches. Many customers have already embraced virtualization as a way to consolidate hardware, reduce costs, and increase the availability of their workloads.

Hyper-convergence as implemented by Red Hat Hyperconverged Infrastructure incorporates distributed, highly available storage into servers, can improve I/O performance, and eliminates the need to depend on complex, expensive, and difficult to manage SAN storage. Red Hat Hyperconverged Infrastructure provides an Open Source solution for those customers who would rather not have proprietary one.

Red Hat Hyperconverged Infrastructure for Virtualization integrates well with Red Hat Ansible Automation, which provides deployment and management automation. This reduces manual interventions creates repeatable deployment processes, and helps continuous development/continuous integration.

3 Requirements

The functional and non-functional requirements for this reference architecture are described below.

3.1 Functional requirements

The primary use case is to deploy a software application in a virtual machine (VM).Figure 1 shows the high-level use cases for creating and using virtual machines.
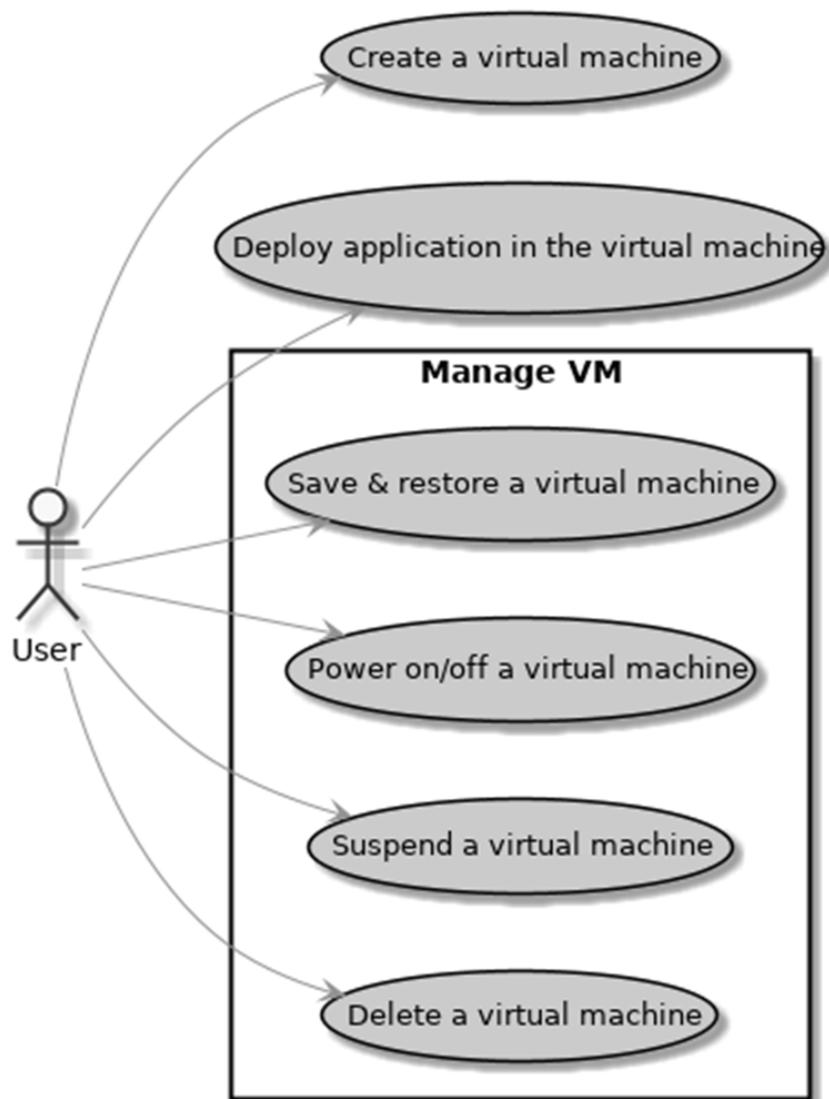

Figure 1: High-level Use Cases

These use cases can be summarized as follows:

1. As a cloud user, I want to deploy application in VMs, so that the deployment is portable to multiple hypervisors such as VMware ESXi and KVM.
2. As a cloud user, I want to create VM from disk image, so that I can reuse an existing VM image.

3. As a cloud user, I want to create VM from a configuration template that defines VM attributes, such as CPU, memory, disk size, base image, and so on, so that I will get an identical VM if using the same values.

4. As a cloud user, I want to pin a VM to a specific physical node, so that I can use a high availability model on my application to mandate VMs on different physical nodes.

5. As a cloud user, I want to pin a VM to CPU threads, IO threads, emulator threads, or NUMA nodes, so that I can meet the performance requirements of my workload.

6. As a cloud user, I want to take snapshot of a VM to preserve its execution states so that I can restore the states and resume the execution

3.2 . Non-functional requirements

In addition to the use cases above, a private cloud needs fault tolerance, high availability, maintenance, ease of installation, and high performance.

3.2.1 Fault tolerance and high availability

Fault tolerance ensures the integrity of the infrastructure when some hardware resource become unavailable. The most important considerations are the safety of data such as management data, workload execution states, and user data.

1. As a cloud admin, I want the data saved in the resilient storage in the event of single drive failure or single server failure.

2. As a cloud admin, I want to create a backup of VM snapshots

3. As a cloud admin, I want the infrastructure to continue to meet all the functional requirements when a single node is down or if any sub-components inside the node are down.

4. As a cloud admin, I want to have redundancy in network connections.

3.2.2 Maintenance

Infrastructure requires regular maintenance. The key consideration is to ensure maintenance has a minimal or zero impact on the infrastructure.

1. As a cloud admin, I want to know which VMs are running on which physical node, so that I can evaluate the impact of shutdown of that node in the event of server maintenance or repair.

2. As a cloud admin, I want to know which physical nodes and VMs are running on a particular network, so that I can evaluate the impact of shutdown of that network in the event of network maintenance or repair.

3. As a cloud admin, I want to manually or even automatically migrate VMs to another physical node in the event of server shutdown for maintenance or repair. Pinned VMs present difficulties and may need special intervention.

3.2.3 Ease of installation

It should be easy to build such the infrastructure.

1. As a cloud admin, I want to automatically build the infrastructure on bare-metal servers using a configuration file that describes the end state of the infrastructure.
2. As a cloud admin, I want to rebuild the infrastructure to the same state each time so that the construction process is deterministic and repeatable.
3. As a cloud admin, I want to automate the entire installation, preferably using one simple command or button click.

3.2.4 High performance

High performance is always important.

1. As a cloud admin, I want to use more than 1000 VMs each with two CPU cores, 8 GB memory, and 100 GB disk space to enable support for workloads of this size (for example virtual desktops).
2. As a cloud admin, I want any VM to access data storage with more than 200 MB/second write throughput and more than 500 MB/second read throughput.

4 Architectural overview

Red Hat Hyperconverged Infrastructure for virtualization (RHHI-V) meets the requirements described in the previous chapter. It consists of an integrated Red Hat Virtualization cluster and Red Hat Gluster cluster on a set of Lenovo servers and switches. Figure 2 below shows an overview using three servers.
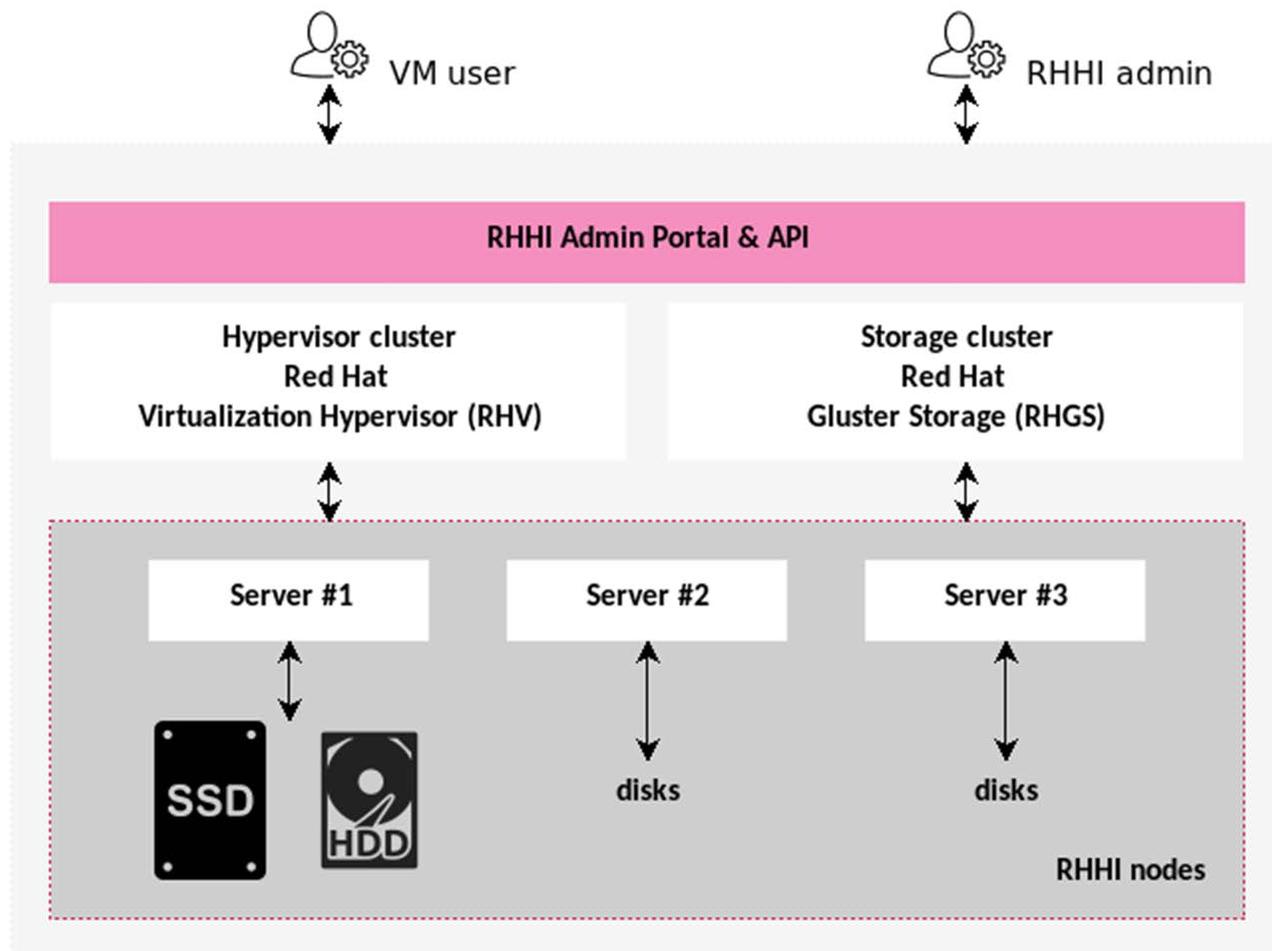

Figure 2: Architecture overview

The primary storage for Red Hat Hyperconverged Infrastructure is Gluster. It uses internal drives in servers to form distributed storage. By default data blocks are replicated to all servers in the cluster.

The server's compute resources are virtualized through a hypervisor, and workloads are deployed in virtual machines (VMs) that can then be located on any of the hosts in the cluster. The cluster's management is also in redundant VMs to provide uninterrupted management capability. There is also a management CLI application on all cluster hosts that provides administration capabilities.

The minimum configuration is three servers. Once the software is deployed, the customer can add more servers, which extends the cluster and provides additional compute, memory, and storage for workloads.

5 Component model

Red Hat Hyperconverged Infrastructure for Virtualization is comprised of two primary components:

- Red Hat Virtualization (RHV) to provide virtualized, highly available environment for virtual machines
- Red Hat Gluster Storage (RHGS) to provide distributed, highly available storage

5.1 Red Hat Virtualization

Red Hat Virtualization is an enterprise-grade virtualization platform built on Red Hat Enterprise Linux and has three key components:

- **Red Hat Virtualization Manager**: A service that provides a graphical user interface and a RESTful API to manage the resources in the environment.
- **Hosts**: Red Hat Enterprise Linux hosts (RHEL-based hypervisors) and Red Hat Virtualization Hosts (image-based hypervisors) are the two supported types of host. Hosts use Kernel-based Virtual Machine (KVM) technology and provide resources used to run VMs.
- **Storage**: A storage service is used to store the data associated with VMs.

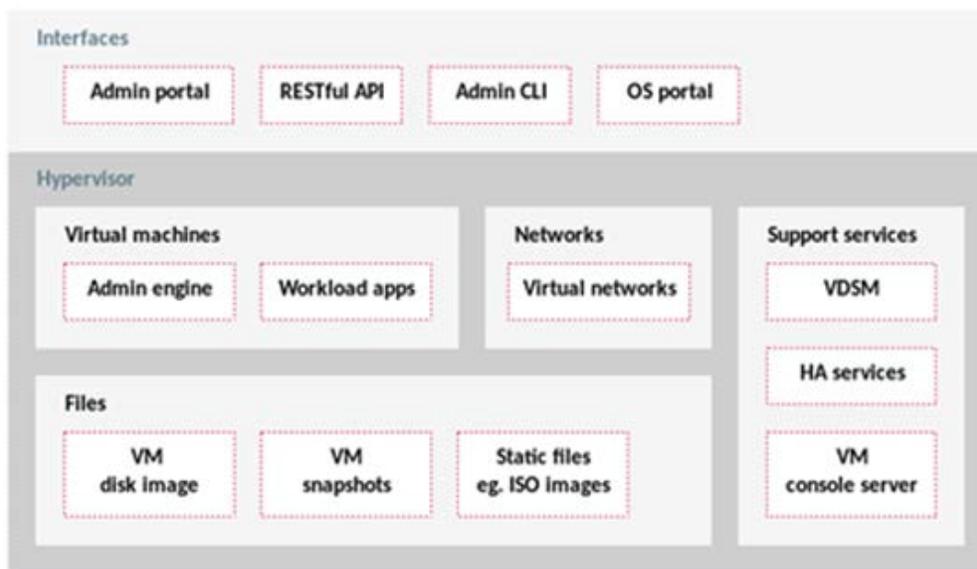Figure 3 shows a high-level model of the Red Hat Virtualization.

Figure 3: High-level Model

5.1.1 Interfaces

Interfaces provide entry points into the infrastructure for an administrator or user.

- **Admin interfaces**: There are three interfaces for the infrastructure administrator — a web portal, a RESTful API, and an Admin CLI, These interface cover both the hypervisor and the Gluster storage.
- **OS portal**: RHEL 7 has a built-in portal, the "cockpit", which makes it easy to administrate a RHEL-based host.

5.1.2 Hypervisor

The hypervisor, as known as KVM, provides the virtualization services and represents hardware resources as virtual resources.

Virtual machines are virtualized compute resources. There are two types of virtual machines:

- Workloads: Any software deployed on the infrastructure is a workload.
- Management: VMs are for administrating the infrastructure

Physical networks are virtualized so a server's physical NICs and cable connections are shareable by virtual machines. Virtual networks are not associated with any particular physical server. Thus regardless where the virtual machine is running, these networks can be available to it. In order for this to work, however, all servers in a cluster must have identical physical networks.

There are many support services for the infrastructure. Three key services are:

- **VDSM**: Daemon running on each host to support the admin application.
- **HA services**: responsible to keep the admin application running on a healthy host so that administrative capability of the infrastructure has high availability.
- **VM console server**: gives remote console access to the VMs. This allows access to the VM for administrative tasks, even before VM networking is configured.

There are three important types of files in Red Hat Hyperconverged Infrastructure:

- **VM disk image**: A virtual machine consists of one or many disk images that function like the physical disk in a server.
- **VM snapshot**: A snapshot allows a user to freeze the state of a virtual machine in time, and recover and resume from a snapshot.
- **Static files**: Other important files such as ISO images.

Red Hat Virtualization uses storage domains to data images. A storage domain can be either a block device such as "Red Hat Gluster Storage" (see page 9) or a POSIX compliant file system such as NFS.

Data storage domains store the virtual drives and OVF files for the VM templates, VMs, and VM snapshots. There must be at least one data domain for a Red Hat Hyperconverged Infrastructure for Virtualization deployment.

ISO domains store the ISO files used to install and boot operating systems and applications for the VMs. This removes the requirement for physical media. Only one ISO domain can be used for a Red Hat Hyperconverged Infrastructure for Virtualization deployment and the same one can be shared across multiple deployments. ISO domains can only be NFS-based.

Table 1 shows the mapping between Gluster volumes and domain types.

Table 1. Gluster volumes and domain types

Gluster Volume	Domain Name	Domain Type
engine	hosted_storage	Data
vmstore	vmstore	Data
Data	Iso	ISO

5.2 Red Hat Gluster Storage

Red Hat Gluster Storage is a software-defined, scale-out storage that provides flexible and affordable unstructured data storage for the enterprise. Figure 4 shows the architecture of the Red Hat Gluster storage using two servers as an example. This picture is available from Red Hat documentation.
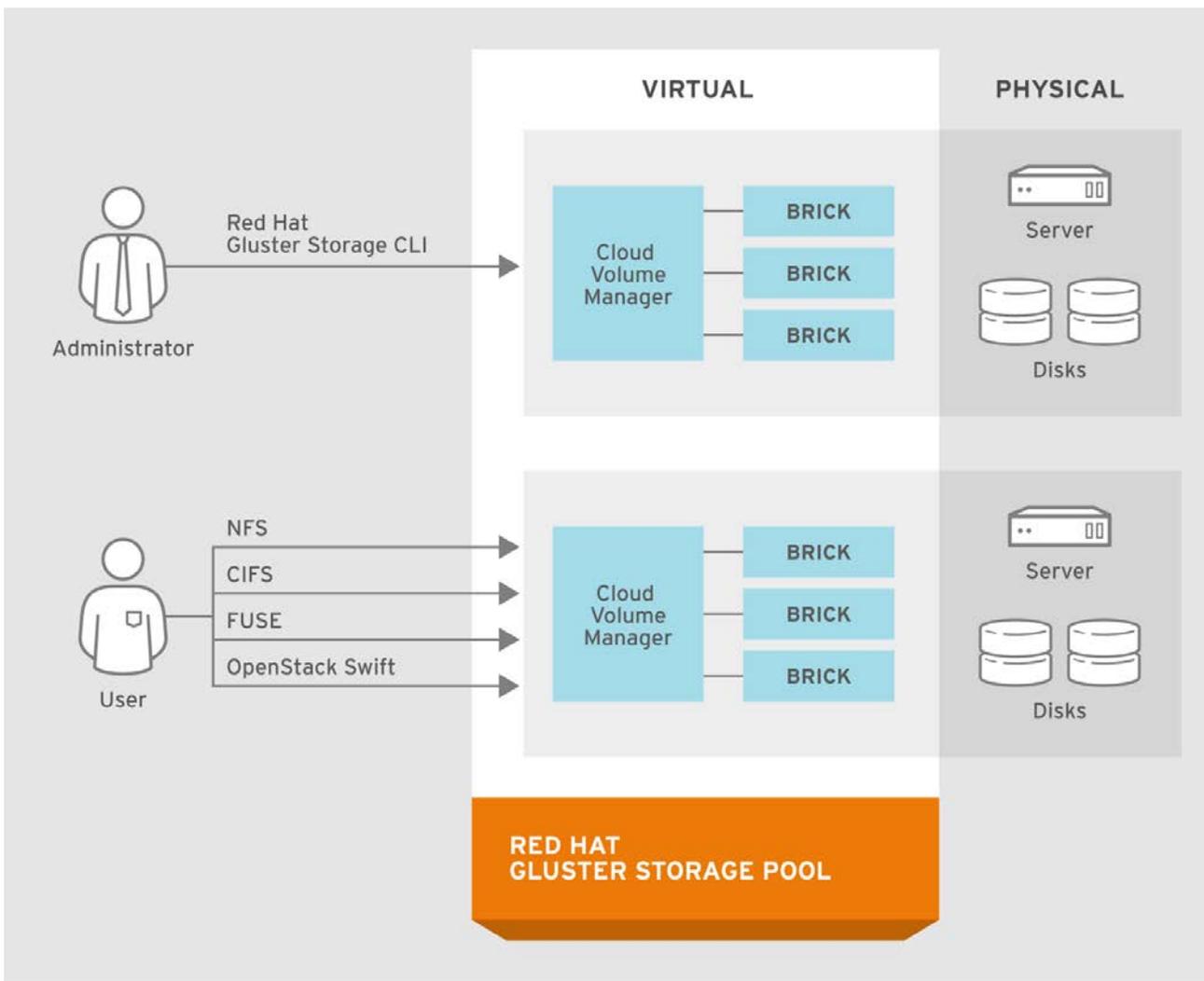
Figure 4: Red Hat Gluster Storage architecture

5.2.1 Bricks

A brick is the GlusterFS basic unit of storage. Each server has three bricks corresponding to a logical volume created on the host:

- engine brick: space for hosted-engine VM storage only.
- vmstore brick: space for all VM disk images.
- data brick: space for static files such as an operating system ISO image that is used to create VMs.

Figure 5 shows the three Gluster bricks in context of logical volumes and disks on a node.
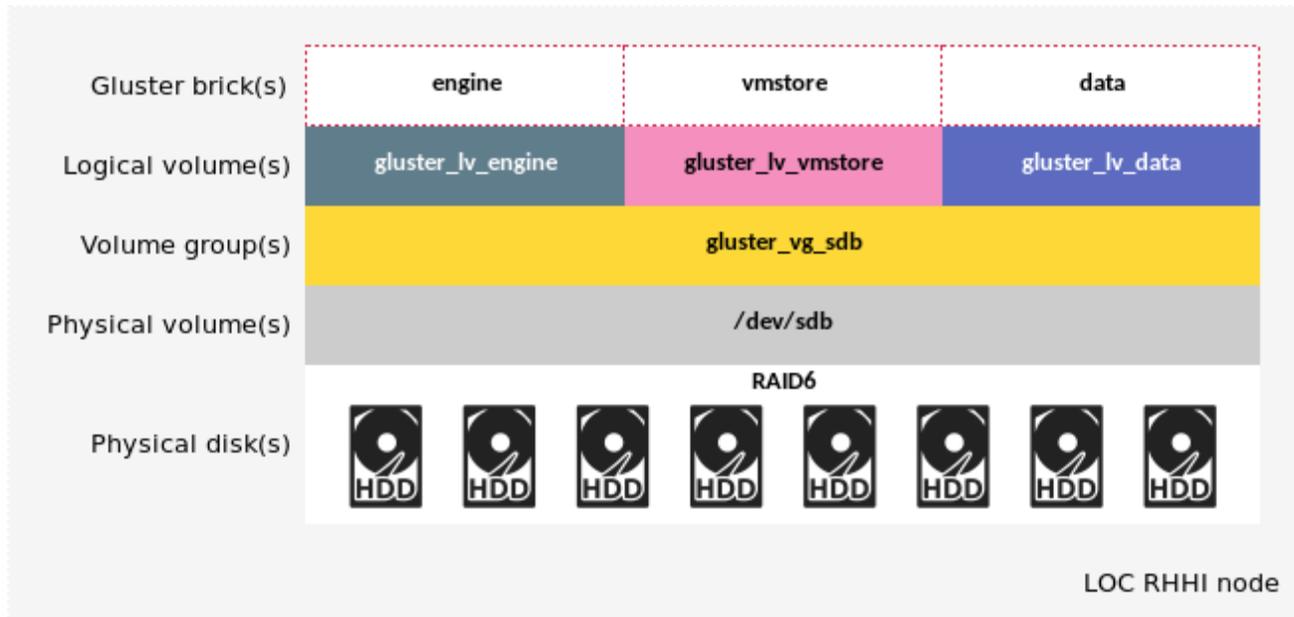

Figure 5: Red Hat Gluster Logical Volumes and Gluster Bricks

An export directory on a node in the trusted storage pool represents a brick. The format of the export directory is "SERVER:EXPORT". Table 2 lists these mounting points and the logical volume type.

Table 2. Gluster bricks and mounting points

Index	LV Name	File System	Mounting Point	LV Type
1	gluster_lv_engine	XFS	/gluster_bricks/engine	thick
2	gluster_lv_vmstore	XFS	/gluster_bricks/vmstore	thin
3	gluster_lv_data	XFS	/gluster_bricks/data	thin

5.2.2 Gluster Volumes

A brick is a local storage to a node. The power of the GlusterFS is to build a cluster of storage over bricks on multiple nodes. A Gluster volume is logical grouping of Gluster bricks. There are three Gluster volumes:

- engine: collection of all engine bricks
- vmstore: collection of all vmstore bricks
- data: collection of all data bricks

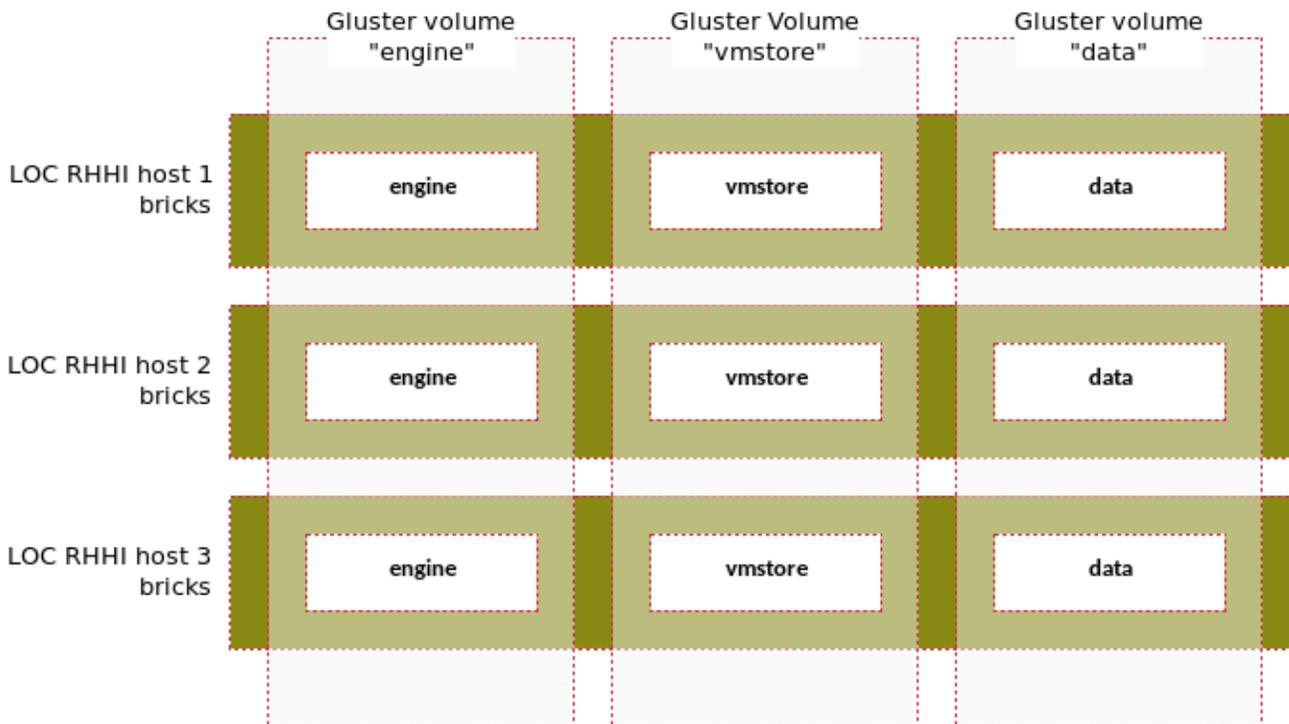Figure 6 shows the Gluster volumes and bricks across three nodes.

Figure 6: Gluster volumes on three nodes

5.2.3 Volume Types

Gluster supports many different types of volumes but in the case of Red Hat Hyperconverged Infrastructure for Virtualization, the only three supported volume types are:

- Replicated — Files are replicated across bricks in the volume. Three copies of the same data are stored on three bricks across three nodes. This type is useful for high-availability and high-reliability is important.

- Arbitrated Replicated — An Arbiter volume is a special subset of a replicated Gluster volume that prevents split-brains. It provides the same consistency guarantees as a replicated volume but two copies of the data are stored on two bricks and one arbiter brick contains the metadata.

- Distributed — Files are distributed across bricks in the volume with no replication. This type is useful for storing large amounts of data but assumes that redundancy is not important or provided elsewhere in the system.

All replicated and arbitrated-replicated volumes must span exactly three nodes. Note that arbiter bricks store only file names, structure, and metadata. This means that a three-way arbitrated replicated volume requires about 75% of the storage space of a three-way replicated volume but is only equivalent to a two-way replication because the arbiter brick only stores metadata.

For more information on Arbiter volumes, see:
https://docs.gluster.org/en/latest/Administrator%20Guide/arbiter-volumes-and-quorum/.

For more information on Gluster volume types, see
docs.gluster.org/en/latest/Administrator%20Guide/Setting%20Up%20Volumes/.

The default volume type is "Replicated" which provides the best comprise of high-availability and performance. However, the cluster will not grow linearly by adding more storage servers to the trusted pool because the size of the storage is limited to the size of the smallest brick in the Gluster volume. Lenovo recommends increasing the Gluster capacity by either:

- Using larger capacity drives
- Using RAID 10 instead of RAID 6 for the volume

5.2.4 LVM cache

Performance of a block device can be improved by using a LVM Cache logical volume. This is done by attaching a smaller and much faster device such as a solid state drive (SSD) to act as a data acceleration layer. The Linux kernel attempts to keep 'hot' blocks in the fast cache layer resulting in better I/O performance for many workloads. Lenovo recommends using SSDs with high endurance for the cache SSDs because they sustain a large number of writes.

For more details on using a LVM cache with Gluster storage, see access.redhat.com/documentation/en-us/red_hat_gluster_storage/3.4/html/administration_guide/sect-lvm_cache.

6 Operational model

This chapter describes the options for mapping the logical components of Red Hat Hyperconverged Infrastructure for Virtualization onto Lenovo ThinkSystem servers and network switches.

Red Hat Hyperconverged Infrastructure for Virtualization has two deployment modes: standalone or self-hosted engine. This reference architecture concentrates on the self-hosted engine mode.

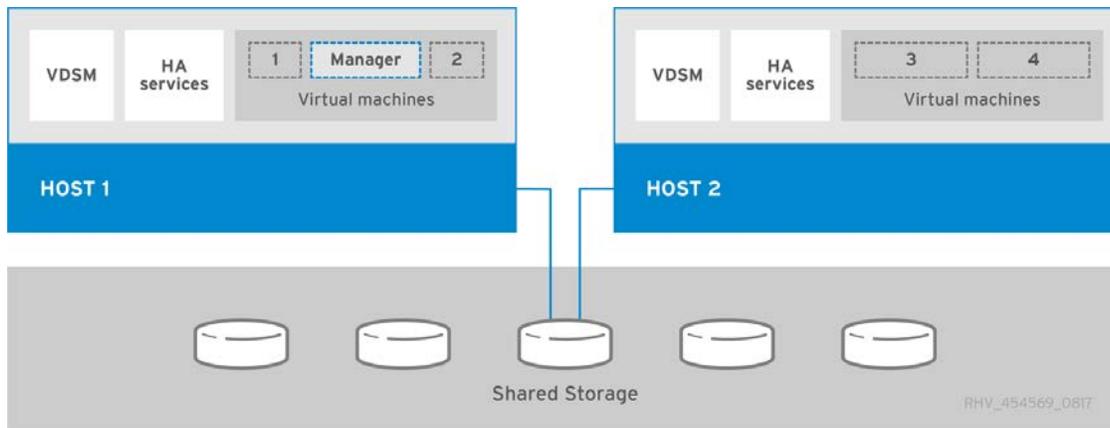Figure 7 shows an overview of the self-host engine architecture. This picture is available from Red Hat documentation.

Figure 7: Self-Hosted Engine Architecture

The Red Hat Virtualization Manager runs as a VM on self-hosted engine nodes (specialized nodes) in the same environment it manages. A self-hosted engine environment requires two but its deployment needs at least three nodes. In this case, the Red Hat Virtualization Manager requires more administrative overhead to deploy and manage but is highly available without needing external HA management. If the node on which Virtualization Manager VM is currently residing goes down, e.g. for maintenance or due to hardware failure, the VM automatically migrates to a different node.

6.1 Hardware components

A Red Hat Hyperconverged Infrastructure for Virtualization cluster contains servers (Red Hat Hyperconverged Infrastructure nodes) and switches.

6.1.1 Servers

Different form factor servers can be used such as the 2U Lenovo ThinkSystem SR650.

Lenovo ThinkSystem SR650

Lenovo ThinkSystem SR650 is the recommended server for nodes because of its overall cost effectiveness, compute, and storage capacity.

The Lenovo ThinkSystem SR650 server (as shown in Figure 8 and Figure 9) is an enterprise class 2U two-socket versatile server that incorporates outstanding reliability, availability, and serviceability (RAS), security, and high efficiency for business-critical applications and cloud deployments. Unique Lenovo AnyBay

technology provides the flexibility to mix-and-match SAS/SATA HDDs/SSDs and NVMe SSDs in the same drive bays. Four direct-connect NVMe ports on the motherboard provide ultra-fast read/writes with NVMe drives and reduce costs by eliminating PCIe switch adapters. In addition, storage can be tiered for greater application performance, to provide the most cost-effective solution.

Figure 8: Lenovo ThinkSystem SR650 (with 24 x 2.5-inch disk bays)

Figure 9: Lenovo ThinkSystem SR650 (with 12 x 3.5-inch disk bays)

Combined with the Intel® Xeon® Scalable processors product family, the Lenovo ThinkSystem SR650 server offers a high-density combination of workloads and performance. Its flexible, pay-as-you-grow design and great expansion capabilities solidify dependability for any kind of virtualized workload, with minimal downtime. Additionally, it supports two 300W high-performance GPUs and ML2 NIC adapters with shared management.

The Lenovo ThinkSystem SR650 server provides internal storage density of up to 100 TB (with up to 26 x 2.5-inch drives) in a 2U form factor with its impressive array of workload-optimized storage configurations. The ThinkSystem SR650 offers easy management and saves floor space and power consumption for the most demanding storage virtualization use cases by consolidating the storage and server into one system. The Lenovo ThinkSystem SR650 server supports up to twenty-four 2.5-inch or fourteen 3.5-inch hot-swappable SAS/SATA HDDs or SSDs together with up to eight on-board NVMe PCIe ports that allow direct connections to the U.2 NVMe PCIe SSDs. The ThinkSystem SR650 server also supports up to two NVIDIA GRID cards for AI or media processing acceleration.

The SR650 server supports up to two processors, each with up to 28-core or 56 threads with hyper-threading enabled, up to 38.5 MB of last level cache (LLC), up to 2666 MHz memory speeds and up to 3 TB of memory capacity. The SR650 also support up to 6 x PCIe slots. Its on-board Ethernet solution provides 2/4 standard embedded Gigabit Ethernet ports and 2/4 optional embedded 10 Gigabit Ethernet ports without occupying PCIe slots.

For more information, see the following website: [ThinkSystem SR650 Product Guide.](#)

6.1.2 Network switches

The following Top-of-Rack (ToR) switches are recommended for the Lenovo Red Hat OpenStack Platform Solution:

- Lenovo ThinkSystem NE0152T RackSwitch
- Lenovo ThinkSystem NE1032 RackSwitch or Lenovo ThinkSystem NE1032T RackSwitch
- Lenovo RackSwitch G8272 or Lenovo ThinkSystem NE1072T RackSwitch

Lenovo ThinkSystem NE0152T Gigabit Ethernet switch

The Lenovo ThinkSystem NE0152T RackSwitch (as shown in Figure 10) is a 1U rack-mount Gigabit Ethernet switch that delivers line-rate performance with feature-rich design that supports virtualization, high availability, and enterprise class Layer 2 and Layer 3 functionality in a cloud management environment.

The NE0152T RackSwitch has 48x RJ-45 Gigabit Ethernet fixed ports and 4x SFP+ ports that support 1 GbE and 10 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables.

The NE0152T RackSwitch runs the Lenovo Cloud Networking Operating System (CNOS) that provides a simple, open and programmable network infrastructure with cloud-scale performance. It supports the Open Network Install Environment (ONIE), which is an open, standards-based boot code that provides a deployment environment for loading certified ONIE networking operating systems onto networking devices.

Figure 10: Lenovo ThinkSystem NE0152T Gigabit Ethernet switch

For more information, see the ThinkSystem NE0152T Product Guide

Lenovo ThinkSystem NE1032/NE1032T/NE1072T RackSwitch family

The Lenovo ThinkSystem NE1032/NE1032T/NE1072T RackSwitch family is a 1U rack-mount 10 Gb Ethernet switch that delivers lossless, low-latency performance with feature-rich design that supports virtualization, Converged Enhanced Ethernet (CEE), high availability, and enterprise class Layer 2 and Layer 3 functionality. The hot-swap redundant power supplies and fans (along with numerous high-availability features) help provide high availability for business sensitive traffic. These switches deliver line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data.

The NE1032 RackSwitch (as shown in Figure 11) has 32x SFP+ ports that support 1 GbE and 10 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables.

Figure 11: *Lenovo ThinkSystem NE1032 RackSwitch*

For more information, see the ThinkSystem NE1032 Product Guide

The NE1032T RackSwitch (as shown in Figure 12) has 24x 1/10 Gb Ethernet (RJ-45) fixed ports and 8x SFP+ ports that support 1 GbE and 10 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables.

Figure 12: *Lenovo ThinkSystem NE1032T RackSwitch*

 For more information, see the [ThinkSystem NE1032T Product Guide](#)

The NE1072T RackSwitch (as shown in Figure 13) has 48x 1/10 Gb Ethernet (RJ-45) fixed ports and 6x QSFP+ ports that support 40 GbE optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. The QSFP+ ports can also be split out into four 10 GbE ports by using QSFP+ to 4x SFP+ DAC or active optical breakout cables.

Figure 13: *Lenovo ThinkSystem NE1072T RackSwitch*

For more information, see the [ThinkSystem NE1072T Product Guide](#)

Lenovo RackSwitch G8272

The Lenovo RackSwitch G8272 (as shown in Figure 14) uses 10Gb SFP+ and 40Gb QSFP+ Ethernet technology and is specifically designed for the data center. It is an enterprise class Layer 2 and Layer 3 full featured switch that delivers line-rate, high-bandwidth, low latency switching, filtering, and traffic queuing without delaying data. Large data center-grade buffers help keep traffic moving, while the hot-swap redundant power supplies and fans (along with numerous high-availability features) help provide high availability for business sensitive traffic.

The RackSwitch G8272 (as shown in Figure 14) is ideal for latency sensitive applications, such as high-performance computing clusters, financial applications and NFV deployments. In addition to 10Gb Ethernet (GbE) and 40 GbE connections, the G8272 can use 1 GbE connections.

Figure 14: Lenovo RackSwitch G8272

For more information, see the [RackSwitch G8272 Product Guide](#)

6.2 Red Hat Hyperconverged Infrastructure node

This section describes the recommended configuration for the Red Hat Hyperconverged Infrastructure node. The minimum number of nodes in a cluster is three. Lenovo recommends using at least four nodes so that a complete node failure still results in a highly available cluster and a second failure is not catastrophic.

The processor and memory configuration depends on the types of workload VMs that are going to be used. Generally, Xeon Scalable Processors Gold series CPUs paired with 384 GB or 768 GB of system memory will give an excellent density of VMs. Less capable configurations are still good but may require more servers.

A recommended configuration to execute 150 or more VMs is as follows:

- Lenovo ThinkSystem SR650
    - 2 x Intel Xeon Gold 6130 16C 125W 2.1GHz Processor
    - 768 GB of system memory
    - 2 x ThinkSystem RAID 930-16i 8GB Flash PCIe 12Gb Adapter (for >16 drives)
    - Drives depending on disk configuration (see "Disk configurations" on page 17)

Lenovo recommends installing Red Hat Enterprise Linux 7.5 or later on the M.2 drives and configuring two partitions with the XFS filesystem as follows:

- 100GB mounted as /home
- 380GB mounted as "/"

6.2.1 RAID controller configuration

Lenovo recommends using the 930-16i 8GB flash RAID adapter because the large cache improves performance. The Lenovo XClarity Controller (XCC) can be used to configure the RAID controller for optimal performance. Lenovo recommends the following settings:

- Read policy set to "Adaptive Read Ahead"
- Write policy set to "Write Back" (uses the cache with battery to store the cache if there is a power failure)
- Each drive is configured as a separate RAID 0 drive
- Stripe size of 256 KB
- I/O Policy set to direct I/O
- Access policy set to read write.
- Background initialization is disabled

6.2.2 Disk configurations

The correct disk configuration is important to achieving high performance. In general, SSDs will yield better performance than hard disk drives (HDDs).

Table 3 lists a cost-effective, capacity oriented disk configuration. The SSD-based LVM cache is optional.

Table 3. Disk configuration for capacity

Type	Position	Number	Size	RAID	Purpose
SATA SSD	Front backplane	2	480GB	RAID 1	Operating system
SAS SSD	Front backplane	2	400GB	RAID 1	LVM cache
SAS HDD	Front backplane	18	2.4TB	RAID 10	GlusterFS

Table 4 lists a performance-oriented disk configuration with all SSDs.

Table 4. Disk configuration for performance

Type	Position	Number	Size	RAID	Purpose
SATA SSD	Front backplane	2	480GB	RAID 1	Operating system
SAS SSD	Front backplane	2	800GB	RAID 1	LVM cache
SATA SDD	Front backplane	18	960GB	RAID 10	GlusterFS

In addition, Red Hat Hyperconverged Infrastructure for Virtualization supports alternative network storage such as NFS, for example, used for a SWIFT object store. Disk configuration will thus require two additional disks for the purpose:

Table 5 lists disk configuration that has additional support for file-based storage.

Table 5. Disk configuration with additional file-based storage

Type	Position	Number	Size	RAID	Purpose
SATA SSD	Front backplane	2	480GB	RAID 1	Operating system
SATA SSD	Front backplane	2	400GB	RAID 1	LVM cache
SAS HDD	Front backplane	16	2.4TB	RAID 10	GlusterFS
SAS HDD	Front backplane	2	2.4TB	RAID 1	Affinitized VMs

6.2.3 TPM 2.0 and Secure Boot

Trusted Platform Module (TPM) is an international standard for a secure crypto processor, a dedicated microcontroller designed to secure hardware through integrated cryptographic keys. TPM technology is designed to provide hardware-based, security-related functions. There is no additional cost to enable TPM 2.0 on Lenovo ThinkSystem servers.

6.2.4 ThinkSystem XClarity Enterprise

The Lenovo XClarity™ Controller (aka. XCC) is the next generation management controller that replaces the baseboard management controller (BMC) for Lenovo ThinkSystem servers. Although the XCC Standard Level includes many important manageability features, Lenovo recommends upgrading to the XCC Enterprise Level of functionality. This enhanced set of features includes Virtual Console (out of band browser-based remote control), Virtual Media mounting, and other remote management capabilities.

6.3 Networking

A critical part of a successful Red Hat Hyperconverged Infrastructure for Virtualization deployment is the networking. This section describes the recommended networking design while taking advantage of Lenovo ThinkSystem switches to provide a robust and flexible design.

While there are different approaches to network design, and each data center is different, Lenovo recommends following these industry best practices:

- Data center switches follow a left-spine topology.
- Switches are always paired using inter-switch-links (ISL) for redundancy.
- Except for management connections like the XCC, server to switch connections are always paired each pair should connect to separate NICs on the server side, and to separate switches.
- Separate management traffic from data traffic whenever possible.
- Improve network isolation by assigning private IP space to internal only traffic whenever possible.

Figure 15 shows the network connections for the servers and switches. This requires matching configuration on the switch side, and on the server using active-active or active-backup bonding. For more information, see Red Hat Enterprise Linux deployment guide on channel bonding interfaces.
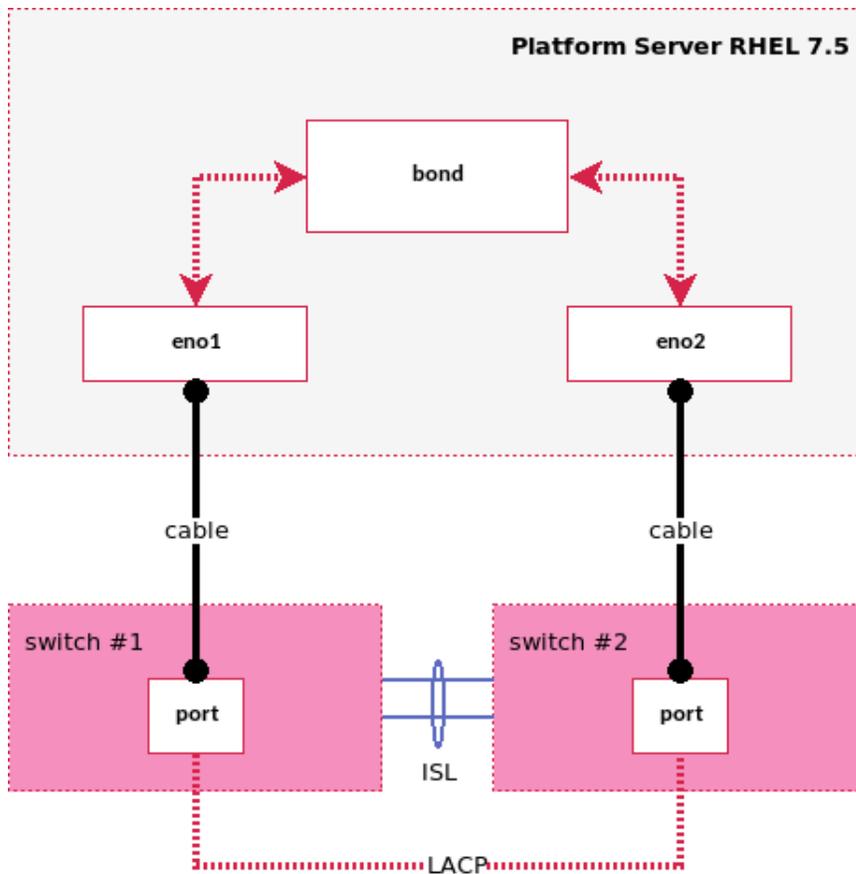
Figure 15: Network connection for servers and switches.

The networks used by Red Hat Hyperconverged Infrastructure for Virtualization can be categorized into three groups:

- management network
- storage network
- workloads network

This is an over-simplistic view as in practice, there is also traffic to manage bare-metal servers, to access the server XCC, and workloads may have their specific pattern of consuming and congesting a shared NIC.

Lenovo recommends the following approach in order to build a well-planned network.

1. **Define logical networks by their function**: this defines the purpose of this network, thus clarifying its characteristics such as load, latency, space, etc.
2. **Assign VLANs to logical networks**: each logical network is assigned a unique VLAN. Lenovo recommends establishing a consistent VLAN naming convention that all VLANs should follow.
3. **Map network/VLANs to server's network interface**: this defines how server side interfaces are configured to support these networks.
4. **Map server's networking interface to switch**: this defines the connection between server and switch including the switch side configuration such as port mode, vLAG, native VLAN, and untagged VLANS.
5. **Cabling schema**: this defines switch port assignments, which server port is connected to which switch port.
6. **Configure switches**: apply switch side configurations to Lenovo switches.
7. **Configure server network interfaces**: Use configuration files to create network interfaces on the server running RHEL.
8. **Map workload network to VLAN and RHV network**: this defines the networks that are needed to support workloads.

6.3.1 Network Topology

Figure 16 shows the overall topology of Red Hat Hyperconverged Infrastructure for Virtualization networking.
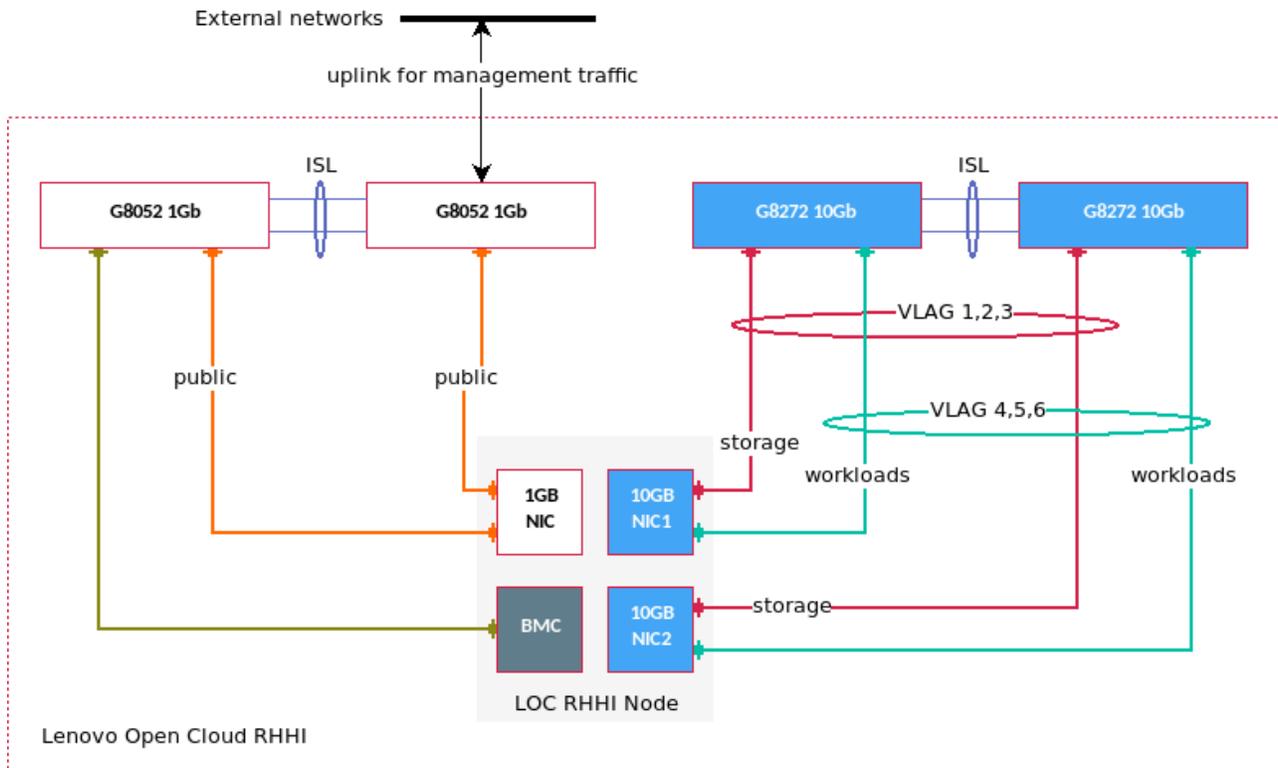
Figure 16: Network Topology

The diagram above shows:

- One 1 GbE switch is connected to uplink to allow management access from external network.
- Switches are paired via inter-switch-link (ISL).
- Except XCC connection, all server to switch links are in pairs on switch side.
- 10 GbE connections are teamed in vLAG; 1 GbE connections are not teamed.

6.3.2 Logical networks

Red Hat Hyperconverged Infrastructure for Virtualization networks are designed to offer performance and high availability. Defining logical networks allows each different network to be independently managed. They serve to distinguish operations and data in a Red Hat Hyperconverged Infrastructure for Virtualization deployment, and this distinction allows an administrator to understand their impact and implication, and to help make decisions on the best way to manage and operate the networks.

Figure 17 shows the logical networking needed for Red Hat Hyperconverged Infrastructure for Virtualization.
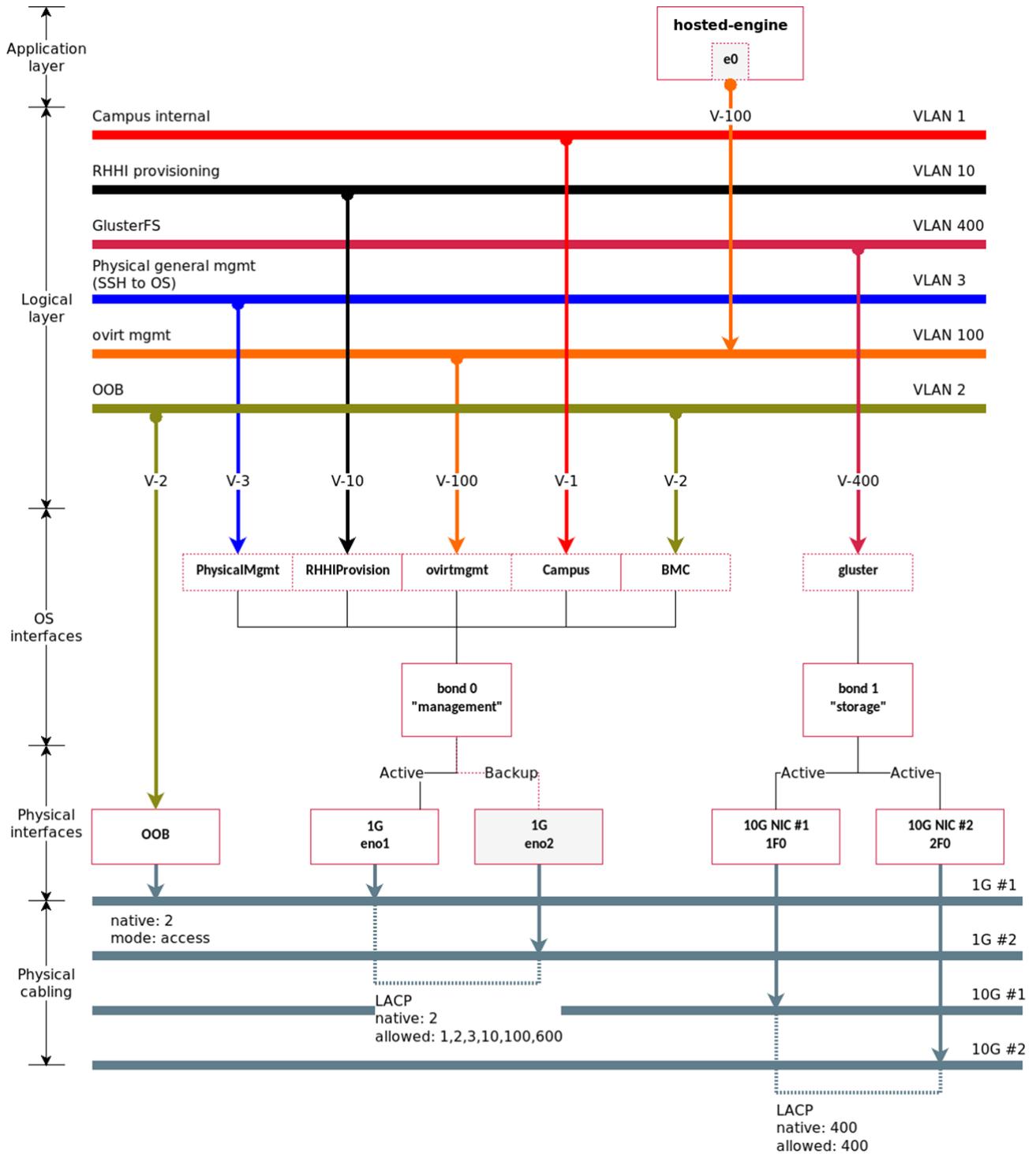
Figure 17: Logical Networks and VLANs

The logical networks are:

ovirt management This network links management console to clusters. This is the default network serving admin portal and ovirt API.

Campus internal The campus (or public network) is for a general audience to access web applications or other user interfaces either within the customer internal network or from the Internet.

In a typical deployment, the ovirt management network serves the administrator user interface. Therefore, it seems redundant to have another network for such purpose. However, following our design philosophy of making the deployment as much self-contained, the ovirt management network is treated as an internal network, while the Campus network represents public access. In this way, the administrator gains more control. A proxy (or jump host) could provide the implementation.

OOB The OOB (or out-of-band) management network connects to a dedicated management port on the physical servers. Often this is replaced by the Campus network so administrator can remotely dial into the XCC and send commands such as rebooting a server. This approach, however, is risking security for convenience. Lenovo recommends a dedicated network for OOB access that can be independently managed and secured.

GlusterFS GlusterFS needs a private network for Gluster data storage as the Gluster cluster is highly sensitive to network interruption. Lenovo recommends dedicating two 10G ports on each server for this network. See "Map Server NIC to Switch" which explains server and switch configuration changes that improve both data throughput and redundancy.

Server management The physical server management network supports in-band management tasks including using SSH to access a server. Often this is replaced by the Campus network for convenience but Lenovo does not recommend this insecure practice.

Provisioning The provisioning network supports installing the OS on a physical server. This is highly depending on the technology used. The most common method is Pre-Boot Execution Environment (PXE). PXE is a broadcast protocol, thus isolating it to this network avoids cross talk.

Lenovo's XClarity Administrator management software can provision a bare-metal server by accessing XCC and mounting a remote disk image, making this network unnecessary.

Once the logical networks are determined, they need to be mapped to VLANs. Table 6 lists the VLAN mapping.

Table 6. VLAN mapping of logical networks

Logical Network	VLAN
Campus internal	1
OOB	2
Physical server management	3
Provisioning	10
OVIRT management	100
GlusterFS	400

6.3.3 Subnets and IP addresses

To keep the network self-contained, Lenovo recommends using a private subnetwork for all of the internal IP addresses in the network as defined by RFC1918. The examples in this RA use the 16-bit block of 192.168.0.0/16. Proxies or routing can be used to grant access to broader audience. Table 7 lists the blocks of IP addresses assigned to each logical network.

Table 7. Logical network subnets

Logical Network	VLAN	Subnet	Addresses	Mask	Gateway
Campus internal	1	192.168.1.x	254	/24	192.168.1.1
OOB	2	192.168.2.x	254	/24	192.168.2.1
Physical server management	3	192.168.3.x	254	/24	192.168.3.1
Provisioning	10	192.168.10.x	3/6/9	/24	192.168.10.1
OVIRT management	100	192.168.100.x	3/6/9	/29	192.168.100.1
GlusterFS	400	192.168.40.x	3/6/9	/29	192.168.40.1

6.3.4 Mapping VLAN to Server NICs with bonding and bridging

This section give an example of mapping VLANs to physical NICs on a server and bonding of those NICs.

As stated earlier, each connection to the 1G or 10G switches is paired but only the 10G connections are bonded (teamed). Table 8 lists the NICs used for each logical network.

Table 8. Logical network mapping to physical NICs

Logical Network	VLAN	XCC	2 x 1G	2 x 10G	Switch Port LACP
Campus internal	1		x		No
OOB	2	x			No
Physical server management	3		x		No
Provisioning	10		x		No
OVIRT management	100		x		No
GlusterFS	400			x	Yes

Note that the OOB logical network is connected to the XCC port on the Lenovo ThinkSystem server. The OOB logical network is also exposed as a Red Hat Virtualization network "OOB" which is through 1G network so VMs can have access to XCC servers.

The 1G connections form an active-backup bonding interface (bond 0 for management), and 10G connections form an active-active bonding interfaces (bond 1 for Gluster storage). Figure 18 shows the VLAN mapping, bonding, and physical NICs.
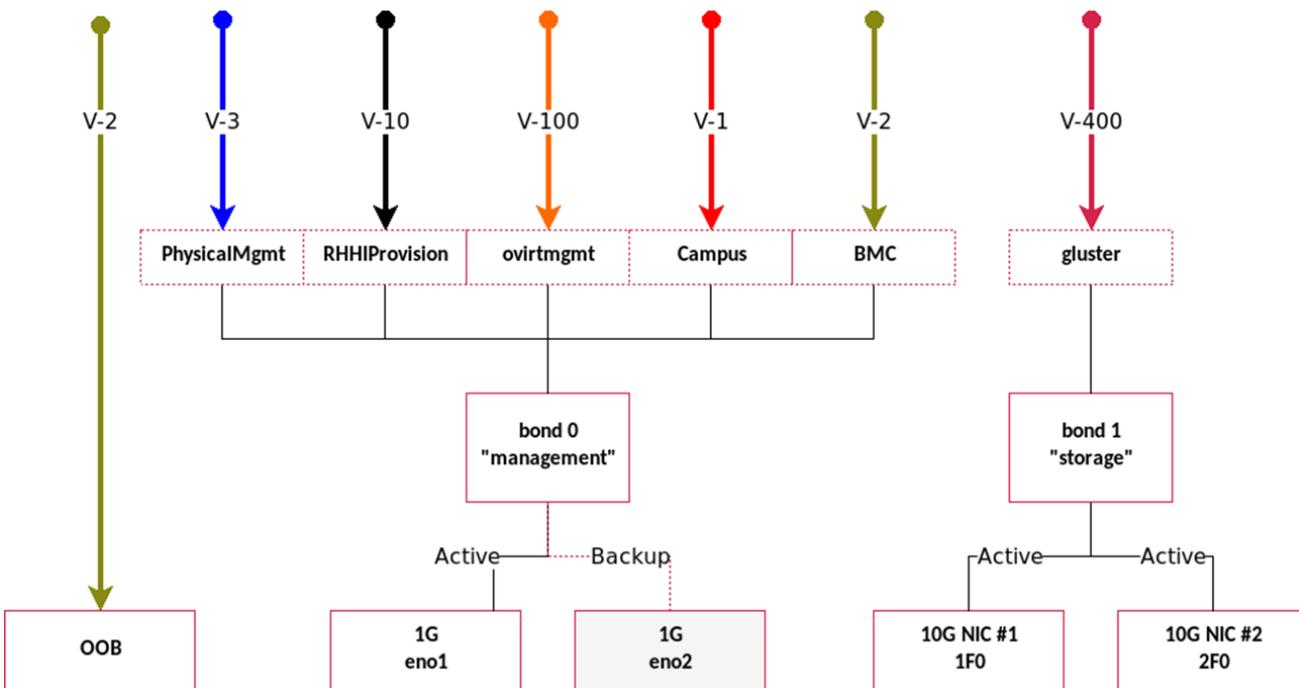
Figure 18: Bonding of VLAN mapping to physical NICs

Virtual bridge interfaces are created on top of a bonding interface to define a one-to-one mapping between a VLAN and a Red Hat Virtualization network. Table 9 shows the VLAN mapping to bonds and bridges in a tabular format.

Table 9. VLAN mapping to bonding interfaces

Logical Network	VLAN	Bonding Interfaces	Bridge Interface (RHV network)
Campus internal	1	bond 0 (management)	CampusInternal
OOB	2	bond 0 (management)	OOB
Physical server management	3	bond 0 (management)	PhysicalMgmt
Provisioning	10	bond 0 (management)	RHHIProvision
OVIRT management	100	bond 0 (management)	ovirtmgmt
GlusterFS	400	bond 1 (storage)	gluster

6.3.5 Map Server NIC to Switch

Lenovo recommends using two 1 GbE switches and two 10 GbE switches for redundancy. Table 10 lists the physical connections between a server and the switch.

Table 10. Server to switch physical connections

Server Port	Switch	Port Redundancy	Mode	Native VLAN	Allowed VLANs
OOB	1 GbE	none	access	–	2
eno1, eno2	1 GbE	pair w/o LACP	trunk	2	1,2,3,10,100
1F0, 2F0	10 GbE	pair w/ LACP	trunk	3999	400,3999
1F1, 2F1	10 GbE	pair w/ LACP	trunk	3999	3999

Note that 3999 is a special VLAN that represents a "not-in-use" network in this Reference Architecture. The two 10 Gb ports of 1F1 and 2F1 are used to support Red Hat Virtualization workloads such as Red Hat OpenStack. Therefore, the port configuration is determined by the workload requirements

6.4 Other Storage

Each node in the cluster provides the primary storage for Red Hat Hyperconverged Infrastructure for Virtualization. In addition, local drives on the servers can be used to provide NFS or cache storage.

6.4.1 NFS

The Gluster-based storage in a Red Hat Hyperconverged Infrastructure for Virtualization could be based on NFS. However, this section is about using NFS outside of a Red Hat Hyperconverged Infrastructure for Virtualization.

Outside of Gluster, NFS does not have built-in redundancy and may only be available to the local server. One advantage is that NFS provides better "write" performance. Lenovo recommends using RAID 1 on each node to provide NFS storage and the storage is mounted in the file system at /mnt/vm.

6.4.2 LVM cache

The LVM cache is local to each node. Two high-endurance SSDs configured as a RAID1 array can be used for the cache. Figure 19 shows an example of the configuration.
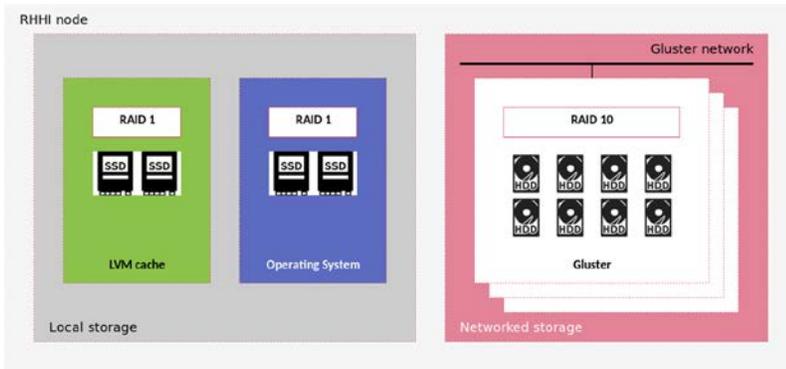
Figure 19: Disk Configuration with the LVM Cache

For deploying, the LVM cache, add the following to the deployment template.

- Set storage/lvm/size to the size of cache in GB. It should be an integer value <=800.
- Set storage/lvm/mounting_point, e.g. sdb.

Leaving the value of storage/lvm blank is to skip setting up the cache on all hosts in deployment even though disks are in place.

6.5 Systems management

Lenovo XClarity™ Administrator is a centralized resource management solution that reduces complexity, speeds up response, and enhances the availability of Lenovo® server systems and solutions.

The Lenovo XClarity Administrator provides agent-free hardware management for Lenovo's ThinkSystem® rack servers, ThinkAgile VX servers, System x® rack servers, and Flex System™ compute nodes and components, including the Chassis Management Module (CMM) and Flex System I/O modules. Figure 20 shows the Lenovo XClarity administrator interface, in which Flex System components and rack servers are managed and are seen on the dashboard. Lenovo XClarity Administrator is a virtual appliance that is quickly imported into a virtualized environment server configuration.
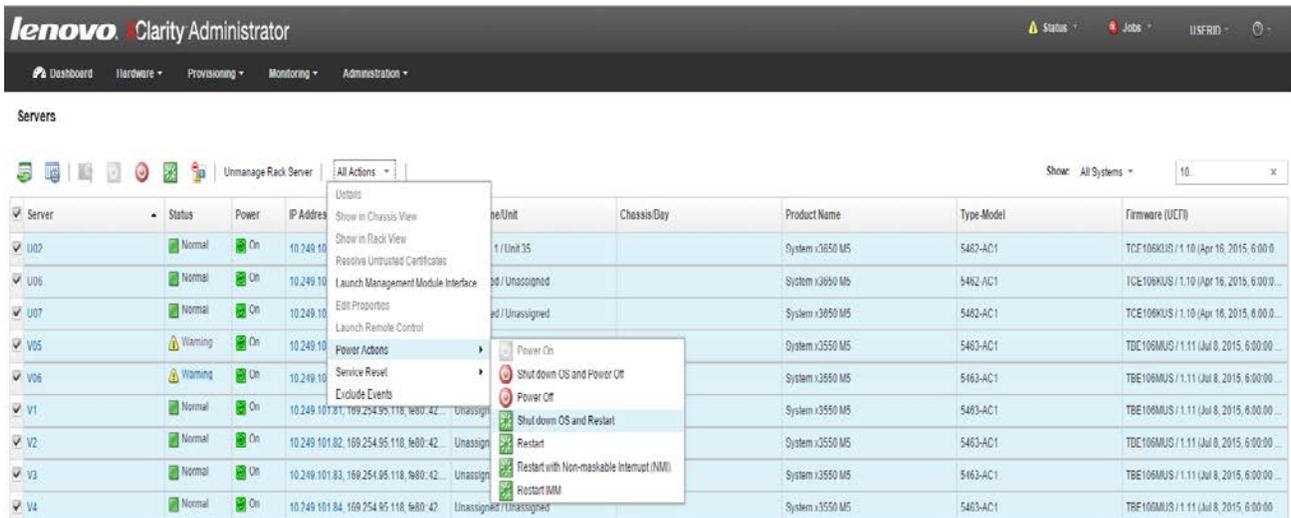
Figure 20: XClarity Administrator interface

7 Deployment

Traditionally Red Hat Hyperconverged Infrastructure for Virtualization can be deployed using either an ISO image or installing RHEL and using the "YUM" package manager. This chapter gives an overview of the Lenovo developed a bootstrap appliance which fully automates the deployment of Red Hat Hyperconverged Infrastructure for Virtualization.

The following conditions should be met before starting the automatic deployment process:

- Servers are cabled
- Switch ports are configured

7.1 Bootstrap Services

The Lenovo bootstrap appliance can execute either as a physical server or as a virtual appliance. Table 11 lists the core services that facilitate deployment.

Table 11. Bootstrap Appliance Core Services

Index	Service	Description
1	Software repo	Source of software installation
2	Automation	Deployment orchestration
3	OS deployment	Out-of-band (OOB) management of server
4	Configuration repo	Source of deployment templates & instance configurations
5	Validation	Deployment validation
6	OS Image	Source of OS images and VM images
7	Launchpad	Automation runtime worker environment

Table 12 lists the add-on services that extend automation such as discovering server on a monitored network:

Table 12. Bootstrap Appliance Add-On Services

Index	Service	Description
1	Discovery	Hardware detection and configuration
2	Inventory	Infrastructure planning and role-based hardware commission
3	Single pane of glass	Unified administration interface

7.2 Bootstrap networks

The bootstrap appliance must have proper networking connections the nodes that will become the Red Hat Hyperconverged Infrastructure for Virtualization cluster. Figure 21 shows the bootstrap services and three required networks: Campus internal, OOB, and ovirt management. These are the same logical networks described in "Logical networks" on page 21. The other networks are not used in the bootstrap process.
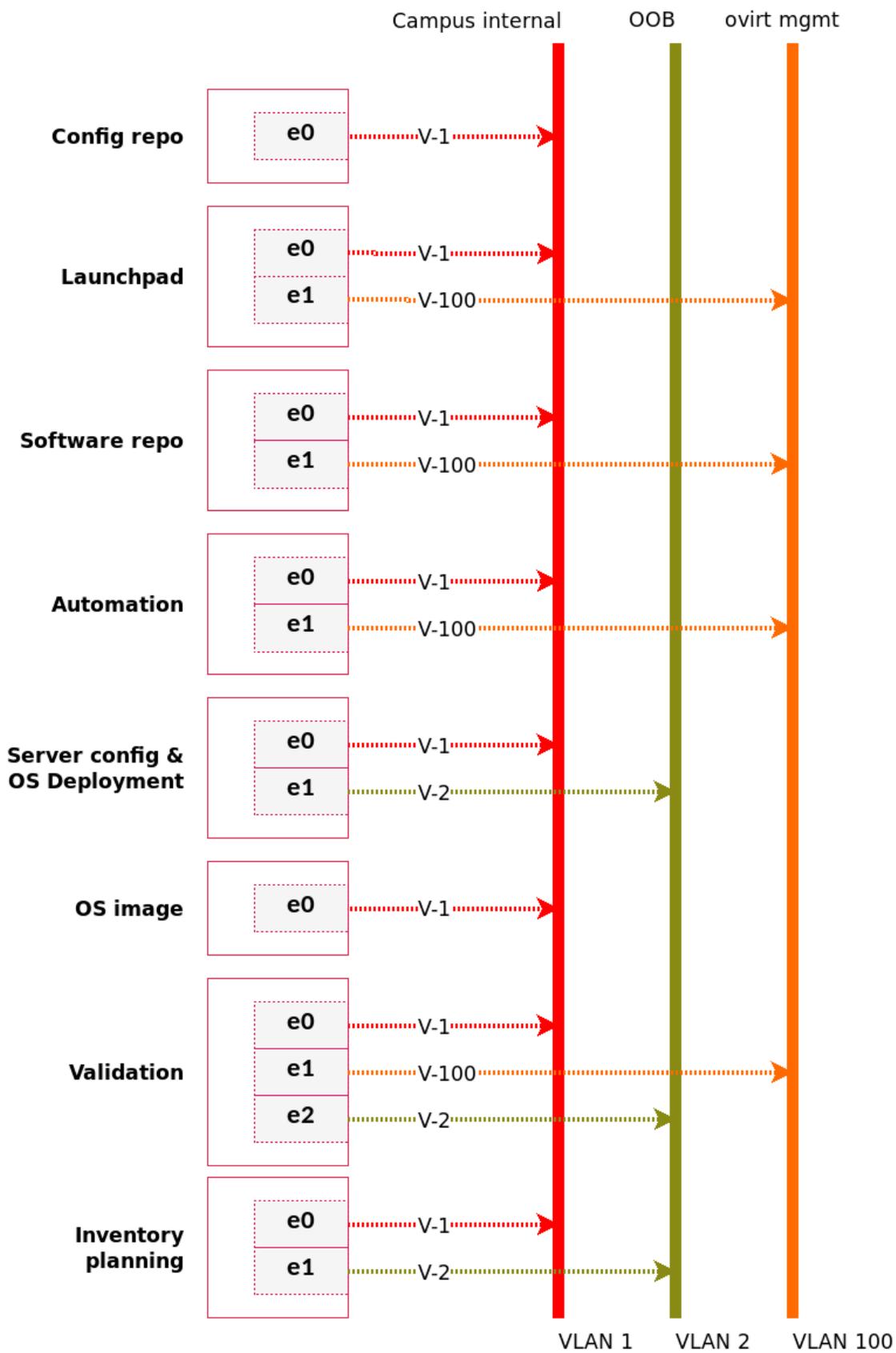
Figure 21: Bootstrap Appliance Networks

8 Appendix A: Lenovo bill of materials

This appendix contains the Bill of Materials (BOMs) for the different hardware configurations supported for the Red Hat Hyperconverged Infrastructure for Virtualization deployments.

8.1 Server BOM

The following section contains the BOM for Lenovo ThinkSystem Servers. The tables below are generated using the online Lenovo Data Center Solution Configurator. Each table represents one single node. To order N nodes to construct the cluster, just multiply by N with the value in the "Quantity" column.

Table 13. Configuration for capacity node

Code	Description	Quantity
7X06CTO1WW	Storage-development-mode-1-2.5" : ThinkSystem SR650 - 3yr Warranty	1
AUVV	ThinkSystem SR650 2.5" Chassis with 8, 16 or 24 bays	1
AX6D	Intel Xeon Gold 6130 16C 125W 2.1GHz Processor	2
AUND	ThinkSystem 32GB TruDDR4 2666 MHz (2Rx4 1.2V) RDIMM	24
AUR5	ThinkSystem 2U/Twr 2.5" AnyBay 8-Bay Backplane	1
AURA	ThinkSystem 2U/Twr 2.5" SATA/SAS 8-Bay Backplane	2
5977	Select Storage devices - no configured RAID required	1
B31E	ThinkSystem RAID 930-16i 8GB Flash PCIe 12Gb Adapter	2
B49M	ThinkSystem 2.5" Intel S4610 480GB Mainstream SATA 6Gb Hot Swap SSD	2
B4Y4	ThinkSystem 2.5" SS530 400GB Performance SAS 12Gb Hot Swap SSD	2
B0YS	ThinkSystem 2.5" 2.4TB 10K SAS 12Gb Hot Swap 512e HDD	18
AURC	ThinkSystem SR550/SR590/SR650 (x16/x8)/(x16/x16) PCIe FH Riser 2 Kit	1
AUKH	ThinkSystem 1Gb 4-port RJ45 LOM	1
AUKX	ThinkSystem Intel X710-DA2 PCIe 10Gb 2-Port SFP+ Ethernet Adapter	2
5053	SFP+ SR Transceiver	4
AVWF	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
AUPW	ThinkSystem XClarity Controller Standard to Enterprise Upgrade	1
B0MJ	Feature Enable TPM 1.2	1
AUQB	Lenovo ThinkSystem Mainstream MB - 2U	1

Table 14. Configuration for performance node

Code	Description	Quantity
7X06CTO1WW	Storage-development-mode-1-2.5" : ThinkSystem SR650 - 3yr Warranty	1
AUVV	ThinkSystem SR650 2.5" Chassis with 8, 16 or 24 bays	1
AX6D	Intel Xeon Gold 6130 16C 125W 2.1GHz Processor	2
AUND	ThinkSystem 32GB TruDDR4 2666 MHz (2Rx4 1.2V) RDIMM	24
AUR5	ThinkSystem 2U/Twr 2.5" AnyBay 8-Bay Backplane	1
AURA	ThinkSystem 2U/Twr 2.5" SATA/SAS 8-Bay Backplane	2
5977	Select Storage devices - no configured RAID required	1
B31E	ThinkSystem RAID 930-16i 8GB Flash PCIe 12Gb Adapter	2
B49M	ThinkSystem 2.5" Intel S4610 480GB Mainstream SATA 6Gb Hot Swap SSD	2
B4Y5	ThinkSystem 2.5" SS530 800GB Performance SAS 12Gb Hot Swap SSD	2
B49N	ThinkSystem 2.5" Intel S4610 960GB Mainstream SATA 6Gb Hot Swap SSD	18
AURC	ThinkSystem SR550/SR590/SR650 (x16/x8)/(x16/x16) PCIe FH Riser 2 Kit	1
AUKH	ThinkSystem 1Gb 4-port RJ45 LOM	1
AUKX	ThinkSystem Intel X710-DA2 PCIe 10Gb 2-Port SFP+ Ethernet Adapter	2
5053	SFP+ SR Transceiver	4
AVWF	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
AUPW	ThinkSystem XClarity Controller Standard to Enterprise Upgrade	1
B0MJ	Feature Enable TPM 1.2	1
AUQB	Lenovo ThinkSystem Mainstream MB - 2U	1

8.2 Networking BOM

This section contains the BOM for different types of supported networking switches.

8.2.1 NE0152T 1 GbE Switch

Table 15. BOM of NE0152T

Code	Description	Quantity
7Y810011WW	Lenovo ThinkSystem NE0152T RackSwitch (Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

8.2.2 NE1032 10 GbE Switch

Table 16.BOM of NE1032

Code	Description	Quantity
AU3A	Lenovo NE1032 RackSwitch (Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

8.2.3 NE1032T 10 GbE Switch

Table 17. BOM of NE1032T

Code	Description	Quantity
AU38	Lenovo NE1032T RackSwitch(Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

8.2.4 NE1072T 10 GbE Switch

Table 18. BOM of NE1072T

Code	Description	Quantity
AU36	Lenovo NE1072T RackSwitch(Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

8.3 Rack BOM

Table 19. BOM of Rack

Code	Description	Quantity
93634PX	42U 1100mm Enterprise V2 Dynamic Rack	1
00YJ780	0U 20 C13/4 C19 Switched and Monitored 32A 1 Phase PDU	2

8.4 Red Hat Subscription options

This configuration is based on a 3-node deployment with three year premium service support.

Table 20. BOM of Red Hat subscriptions

Code	Description	Quantity
RS00139F3	Red Hat Hyperconverged Infrastructure for Virtualization	1
RH00031F3	Smart Management	3
MCT3475F3	Red Hat Insights	1
00MT202	Lenovo XClarity Enterprise	3

Resources

For more information about the topics in this document, see the following resources:

- Red Hat Hyperconverged Infrastructure:

  access.redhat.com/products/red-hat-hyperconverged-infrastructure

- Product documentation for Red Hat Hyperconverged infrastructure for Virtualization 1.5:

  access.redhat.com/documentation/en-
  us/red_hat_hyperconverged_infrastructure_for_virtualization/1.5/

- Product documentation for Red Hat Virtualization:

  access.redhat.com/documentation/en-us/red_hat_virtualization/4.2/html/product_guide/index

- Product documentation for Red Hat Gluster Storage:

  access.redhat.com/documentation/en-us/red_hat_gluster_storage/3.4/

- Lenovo Data Center Solution Configurator:

  dcsc.lenovo.com

Acknowledgments

Trademarks and special notices