Lenov Open Cloud Network Reference Architecture

Feng Xia

Contents

List of Tables

List of Figures

Title: Setting Up Lenovo Open Cloud Networks Date: 2018-09-14 09:35 Tags: reference architecture Slug: ibb network design Author: Feng Xia Status: Draft

1 Abstract

The purpose of this document is to provide guidelines and considerations for setting up Lenovo Open Cloud (aka. LOC) networks.

2 Executive Summary

Lenovo Open Cloud consists of a list of physical servers (aka. nodes) and virtual machines (VMs). This reference environment provides a comprehensive example demonstrating how to set up networks to connect these servers and VMs together.

Lenovo Open Cloud is a highly configurable system. From the point of view of physical servers, LOC can be deployed on a 6-server or 9-server configuration. In this document we will use a 9-server configuration as example.

3 Architecture Overview

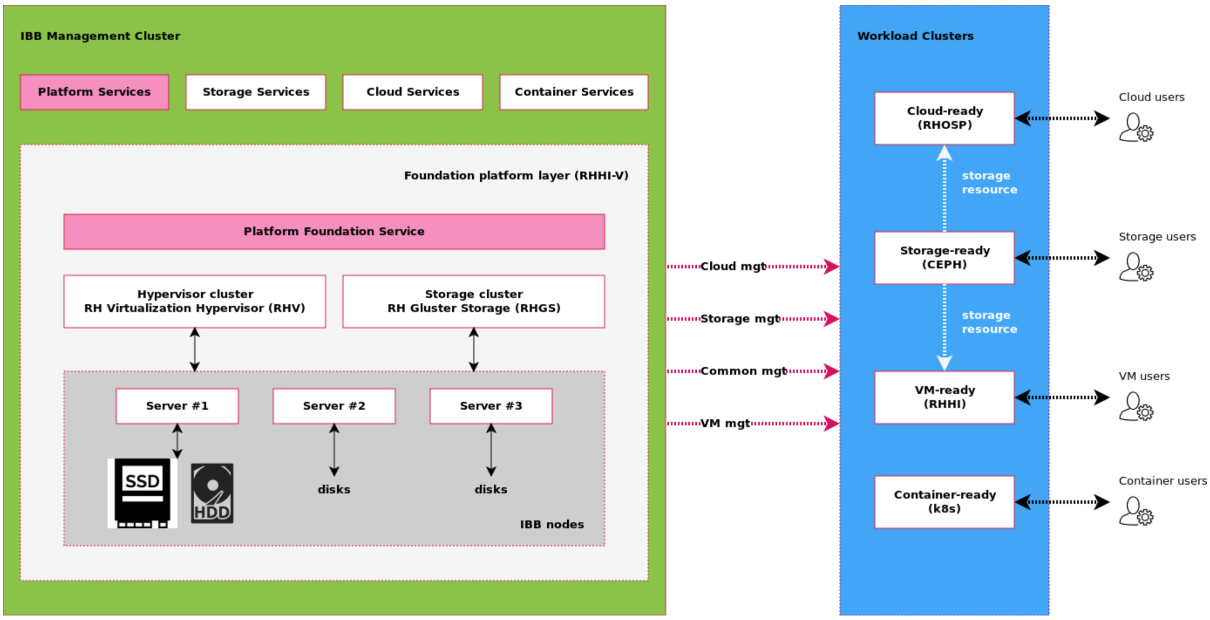


Figure 1: Lenovo Open Cloud Architecture

Lenovo Open Cloud has two sets of clusters: `management cluster` and `workload cluster`. Workload cluster refers to applications directly interfacing with end user. Management cluster refers to applications that manage and provide Open Cloud services.

4 Hardware

4.1 Platform servers

Figure 2: Lenovo ThinkSystem SR650

The Lenovo ThinkSystem SR650 server 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. Plus, storage can be tiered for greater application performance, to provide the most cost-effective solution.

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. All these advanced features make the server ideal to run data and bandwidth intensive VNF workload and storage functions of NFVI platform.

For more information, see product guide.

4.1.1 Memory

There are 24 slots in total for memory in the SR650 server. The maximum memory for each slot is 128 GB. So, the maximum memory of one SR650 server can reach `128GB * 24 slots = 3TB`.

For each SR650 server, it is recommended to use 64GB memory for a small deployment, 128G memory for a medium deployment, and 256GB for a large deployment.[1]

4.1.2 RAID controller

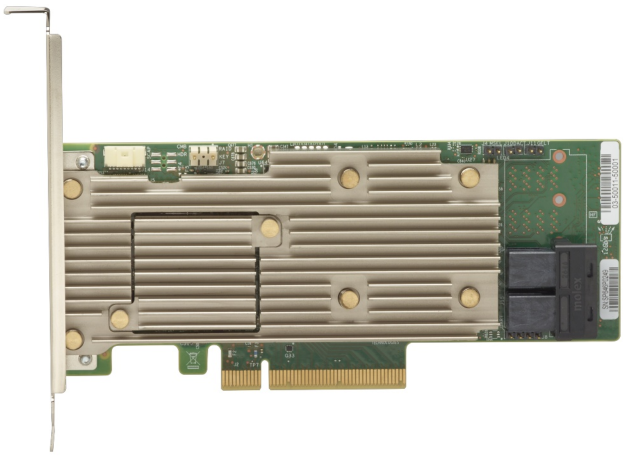


Figure 3: ThinkSystem RAID 930-8i RAID Controller

The ThinkSystem RAID 930 family of internal 12 Gbps SAS RAID controllers are high-performance RAID-on-chip (ROC) adapters. These adapters support RAID levels 0/1/10/5/50/6/60 as well as JBOD, and include an extensive list of RAS and management features.

The family is comprised of four adapters:

1. The ThinkSystem RAID 930-4i supports up to four internal SAS and SATA drives
2. The ThinkSystem RAID 930-8i supports up to eight internal SAS and SATA drives
3. The ThinkSystem RAID 930-16i supports up to 16 internal SAS and SATA drives
4. The ThinkSystem RAID 930-24i supports up to 24 internal SAS and SATA drives

For more information, see product guide.

[1] See Red Hat DEPLOYING RED HAT HYPERCONVERGED INFRASTRUCTURE for details.

4.1.3 Disk configurations

Disk configuration is important to achieve high performance. We recommend the following in each SR650 server:

Type	Position	Number	Size	RAID	Purpose
SSD	Front backplane	2	800GB	RAID1	Operating system
SSD	Front backplane	2	128GB	RAID1	LVM cache
SAS HDD	Rear backplane	8	2TB	RAID6	Glusterfs

Total Size for Gluster: `3 * 8 * 2TB = 48TB`.

4.2 Network Switches

The following sections describe the Top-of-Rack (ToR) switches used in this reference architecture. The Networking Operating System software features of these Lenovo switches deliver seamless, standards-based integration into upstream switches. Two 10 Gb switches and two 1Gb switches are used in this architecture.

4.2.1 Lenovo RackSwitch G8272 (10Gb)

Figure 4: Lenovo RackSwitch G8272

The Lenovo RackSwitch G8272 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 13) is ideal for latency sensitive applications, such as high-performance computing clusters, financial applications and NFV deployments. In addition to 10 Gb Ethernet (GbE) and 40 GbE connections, the G8272 can use 1 GbE connections.

For more information, see product guide.

4.2.2 Lenovo RackSwitch G8052 (1Gb)

The Lenovo RackSwitch™ G8052 (as shown in the following figure) is a top-of-rack data center switch that delivers unmatched line-rate Layer 2/3 performance at an attractive price. It has 48x 10/100/1000BASE-T

Figure 5: Lenovo RackSwitch G8052

RJ-45 ports and four 10 Gigabit Ethernet SFP+ ports (it also supports 1 GbE SFP transceivers), and includes hot-swap redundant power supplies and fans as standard, which minimizes your configuration requirements. Unlike most rack equipment that cools from side-to-side, the G8052 has rear-to-front or front-to-rear airflow that matches server airflow.

For more information, see product guide.

5 Software Services

Lenovo Open Cloud software can be viewed in three categories based on the services they provide:

1. **Platform services**: Platform services are built upon Red Hat Hyperconverged Infrastructure (RHHI). It provides LOC core services each deployed in one or more virtual machines.
2. **Storage services**: Storage services are built upon Ceph. It provides capability to manage Ceph cluster up to xx.
3. **Cloud services**: Cloud is built upon Red Hat Openstack.

5.1 Platform services

Platform services provide administrative functions to support operation of the Open Cloud. This includes management of software life cycle, automation, list of artifacts such as ISO images and qcow images, and new server discovery.

5.1.1 Runtime service

Built upon Red Hat Hyperconverged Infrastructure (RHHI-V). It supports virtual machine users out of box, and is the foundation of other Lenovo Open Cloud services.

> RHHI integrates Red Hat Virtualization (RHV) and Red Hat Gluster Storage (RHGS). RHHI for Virtualization provides open-source, centrally administered, and cost-effective integrated compute and storage in a compact footprint for remote sites.

See product guide for details.

5.1.2 Software repository & life cycle management service

Built upon Red Hat Satellite. All Open Cloud servers are registered to this service, who then is responsible to manage life cycle of:

1. RHEL and Red Hat software products that are deployed in the Open Cloud.
2. Release, update, patch of Lenovo software products.
3. `.iso` and `qcow2` images, which are used by VM creation and server provisioning.

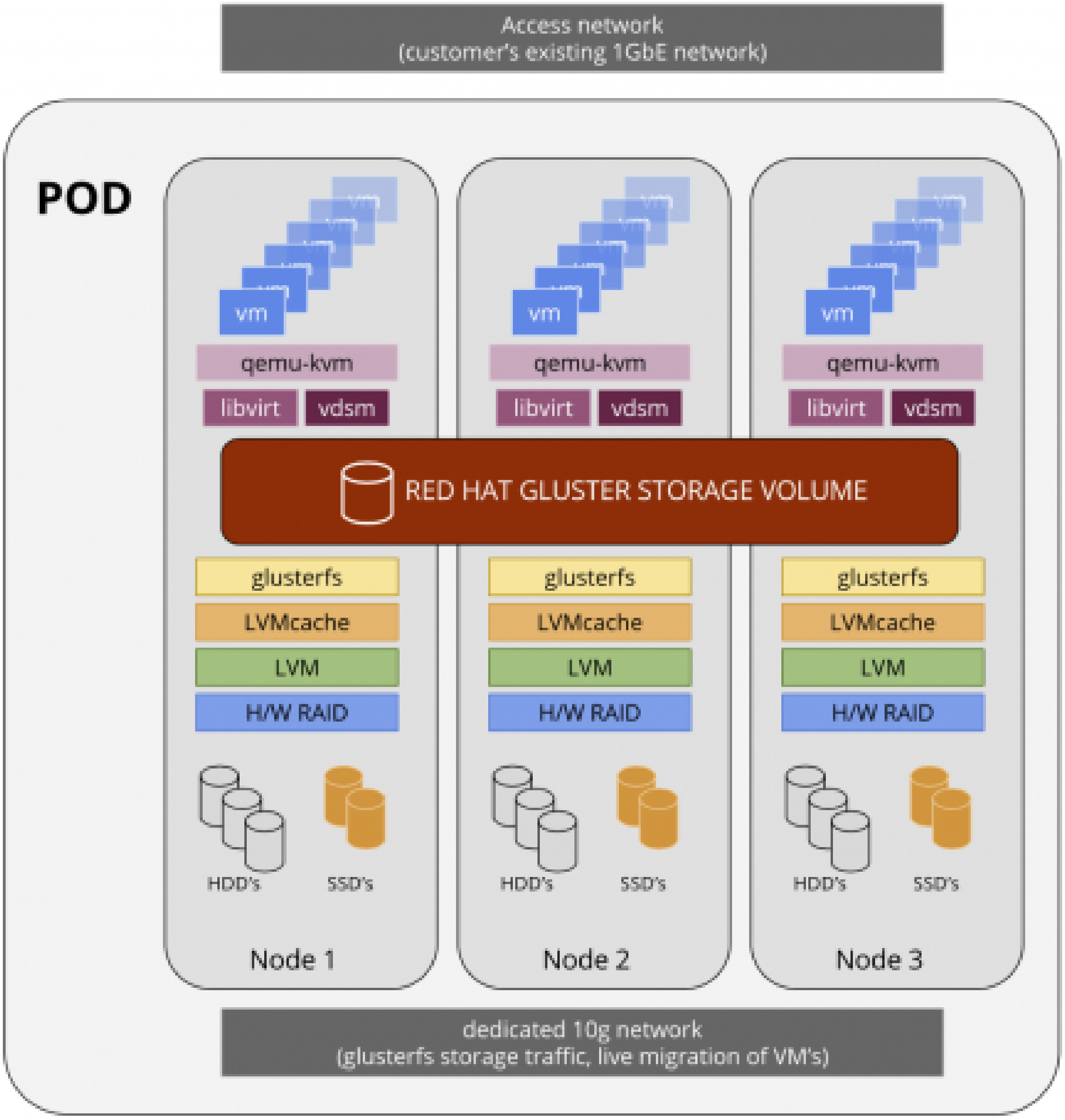


Figure 6: RHHI Single Cluster Architecture

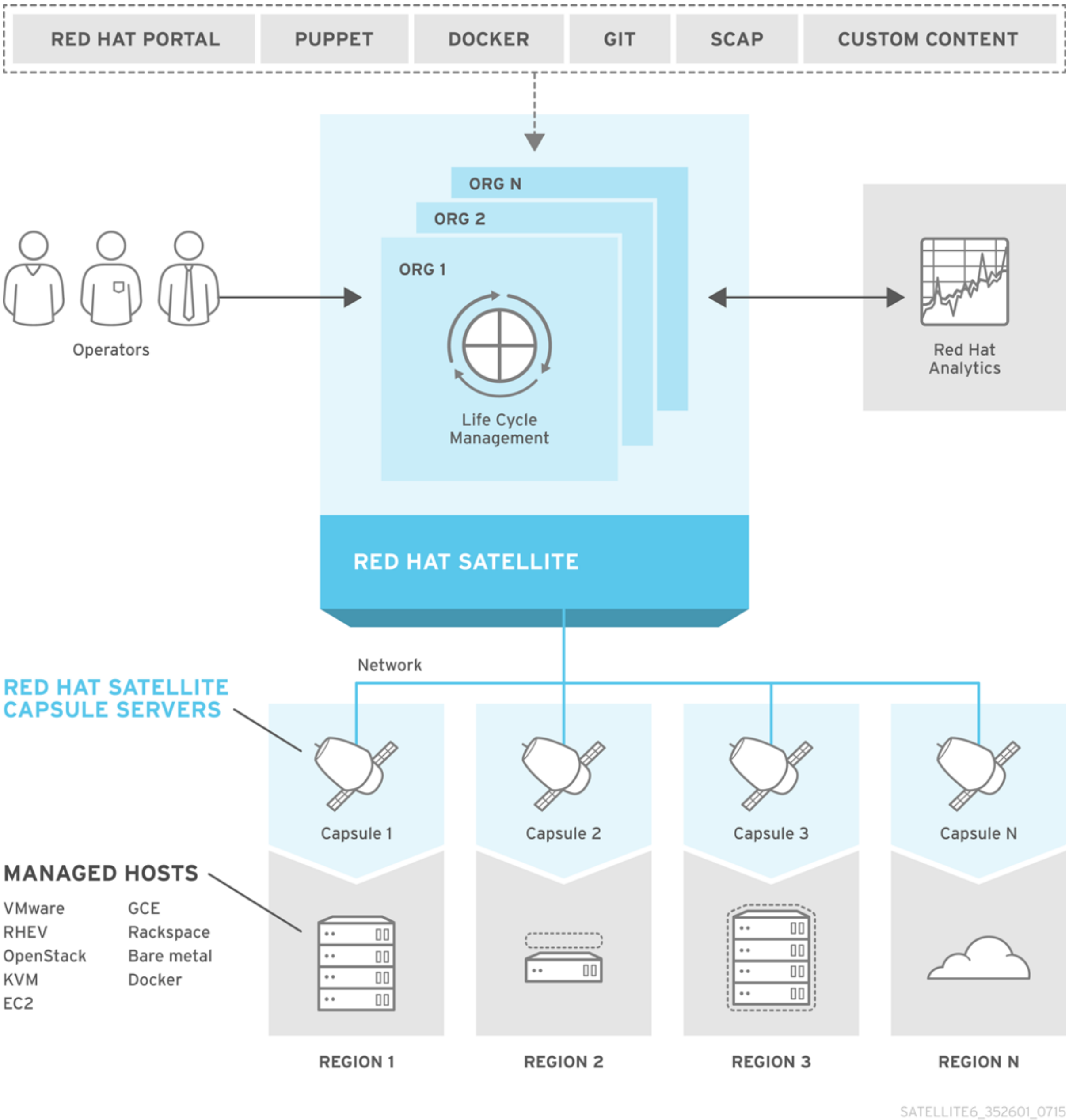


Figure 7: Red Hat Satellite Architecture

> Satellite is an on-premise alternative to trying to download all of your content from the Red Hat content delivery network or managing your subscriptions through the Customer Portal. From a performance side, it reduces hits to your network bandwidth because local systems can download everything they need locally; from a security side, it can limit the risks of malicious content or access, even enabling entirely disconnected environments.
>
> Satellite is composed of a centralized Satellite Server. Depending on your data center setup, organization design, and geographic locations, you can have local Capsule Servers, which are proxies that locally manage content and obtain subscription, registration, and content from the central Satellite Server.

See product guide for details.

5.1.3 Automation service

Build upon Red Hat Ansible Tower. It is the single point of contact to manage servers and VMs using ansible playbooks.

Lenovo Open Cloud is shipped with a list of pre-defined automations that makes managing the infrastructure easy and efficient.

See product guide for details.

5.1.4 Discovery service

Build upon Lenovo Confluent. It continuously monitors network for new Lenovo server and switch. Once identified, the new hardware can be enlisted by other Open Cloud services, such as extending Ceph cluster or adding an Openstack compute node.

See product guide for details.

5.1.5 Inventory planning service

5.1.6 Server config & OS deployment service

5.1.7 OS image service

5.1.8 Configure & Automation repository service

5.2 Storage services

5.2.1 Ceph

5.3 Cloud services

5.3.1 Openstack

6 Network Design

LOC networks can be viewed in three groups whereas:

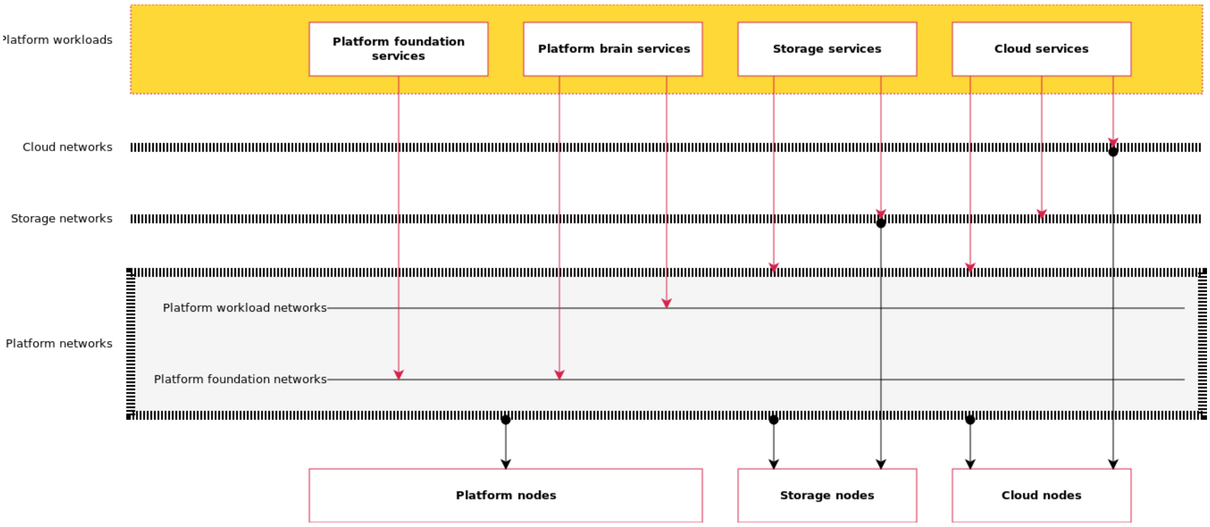


Figure 8: Lenovo Open Cloud Network Overview

1. **platform network**: to support platform services.
2. **storage network**: built on top of platform network with added networks to handle Ceph data storage traffic and Ceph management functions.
3. **cloud network**:

6.1 Conventions

Hardware can break. It is important to keep this in mind when designing a network connection. In this architecture we have followed these conventions:

1. Inter switch connections are paired.
2. Except BMC connection, server to switch connections are paired.
    1. Each pair connect to separate NICs on the server at north bound, and separate switch at south bound.

This then requires matching configuration on the switch using LACP, and on the server using **active-active** `bonding`.

6.2 Connection To Upstream

Showing switch topology within IBB as well as how it is connected to upstream → what is required from upstream, eg. dhcp, dns, gateway, access to RH CDN.

6.3 Platform networks

6.3.1 Platform VLANs

In order to support LOC functions, a list of VLANs are defined as shown below. VLAN index/schema are examples of a test deployment. Your environment can be different. In the following sections we will go over these logical networks in detail, and in section "Implementation Worksheet" we also provide a tool to map these VLANs to your environment.

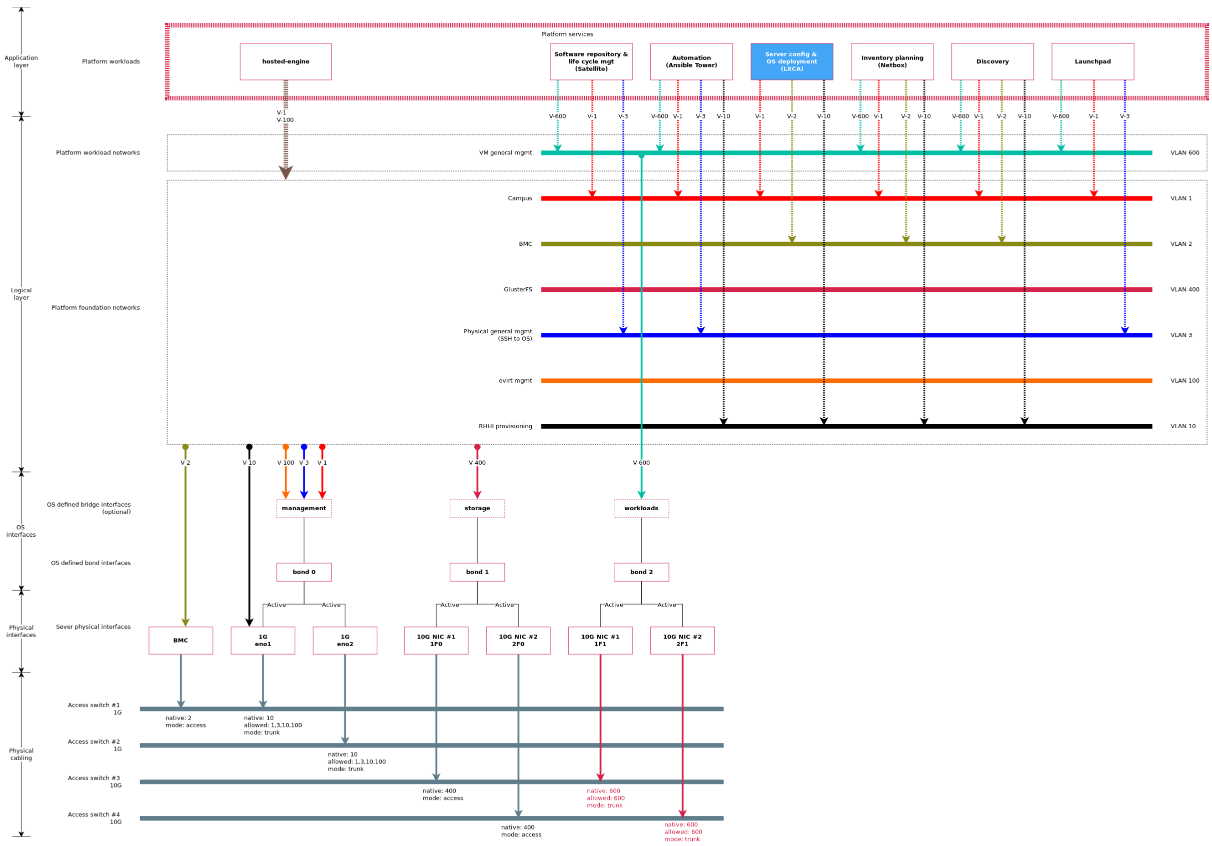


Figure 9: Lenovo Open Cloud Platform Networks

Campus `campus` network is a name for public access. This is the network that Open Cloud user uses to access its services, eg. Ansible Tower's web UI.

BMC also called `Out-of-Band` (OOB) management network. It connects to a dedicated management port on physical server that is separated from data ports.

ovirt management is a private network linking RHHI management console to RHHI clusters. Except Platform admin, other users should not have direct access to it.

glusterFS is a private network used by `gluster` clusters. Gluster cluster is the storage backend of Open Cloud Platform. For example, in a 6-server configuration, three platform servers will form a 3-node gluster cluster.

Physical server management is to support traffic of `In-Band` managerial tasks, eg. `ssh` to a server.

OS provisioning is to support data traffic of installing OS on a physical server. Separating this to its own network is a best practice because operating system image can be large, thus its loading to server can have negative impact on shared traffics.

VM management is to access RHHI virtual machines. This supports both the Open Cloud services and VM workloads. Later we will see that it's also advised to dedicate a NIC for this same purpose.

Table 2: Platform Network VLANs

Network	VLAN	Subnet	Addresses	Mask	Static / DHCP	Gateway
Campus	1	10.240.x.x[2]	10		static	10.240.x.1
BMC	2	192.168.2.x	254	/24	static	192.168.2.1
Physical server management	3	192.168.3.x	254	/24	static	192.168.3.1

[2] This is an example subnet.

Network	VLAN	Subnet	Addresses	Mask	Static / DHCP	Gateway
OS provisioning	10	192.168.10.x	3/6/9	/24	static	192.168.10.1
OVIRT management	100	192.168.100.x	3/6/9	/29	static	192.168.100.1
glusterFS	400	192.168.40.x	3/6/9	/29	static	192.168.40.1
VM management	600	192.168.60.x	11	/28	static	192.168.60.1

6.3.2 Platform services to VLAN mapping

Table 3: Platform Services to VLANs mapping

Platform Services	1	2	3	10	100	400	600
Runtime	x				x		
Software repository & life cycle management	x		x				x
Automation	x		x	x			x
Server config & OS deployment	x	x		x			
Inventory planning	x	x		x			x
Discovery	x	x		x			x
OS image							
Configure & Automation repository service							

6.3.3 Platform server's NIC to VLAN mapping

Beside BMC port, each server has minimal two 1Gb ports and four 10Gb ports. Interfaces are paired to form an active-active bonding interface on Platform server. Optionally, we can also create a network bridge on top of a bonding interface. A sample configuration is shown below:

Table 4: Platform server NIC to VLAN mapping

Network	VLAN	BMC	2 x 1G	2 x 10G	2 x 10G	Bond	Bridge
Campus	1		x			0	management
BMC	2	x				n/a	n/a
Physical server management	3		x			0	management
OS provisioning	10		x[3]			n/a	
OVIRT management	100		x			0	management
glusterFS	400			x		1	storage
VM management	600				x	2	workloads

[3] Bonding provisioning network is optional because loading an operating system is not a frequent event.

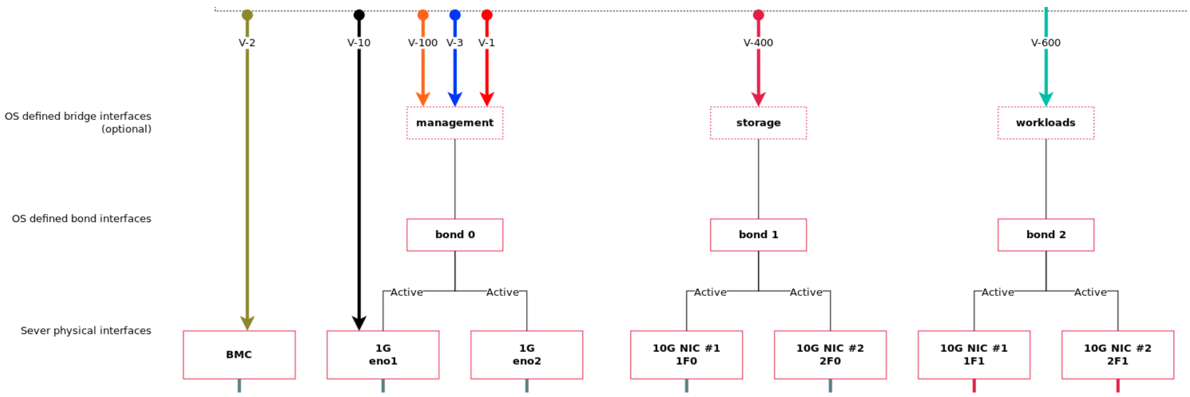
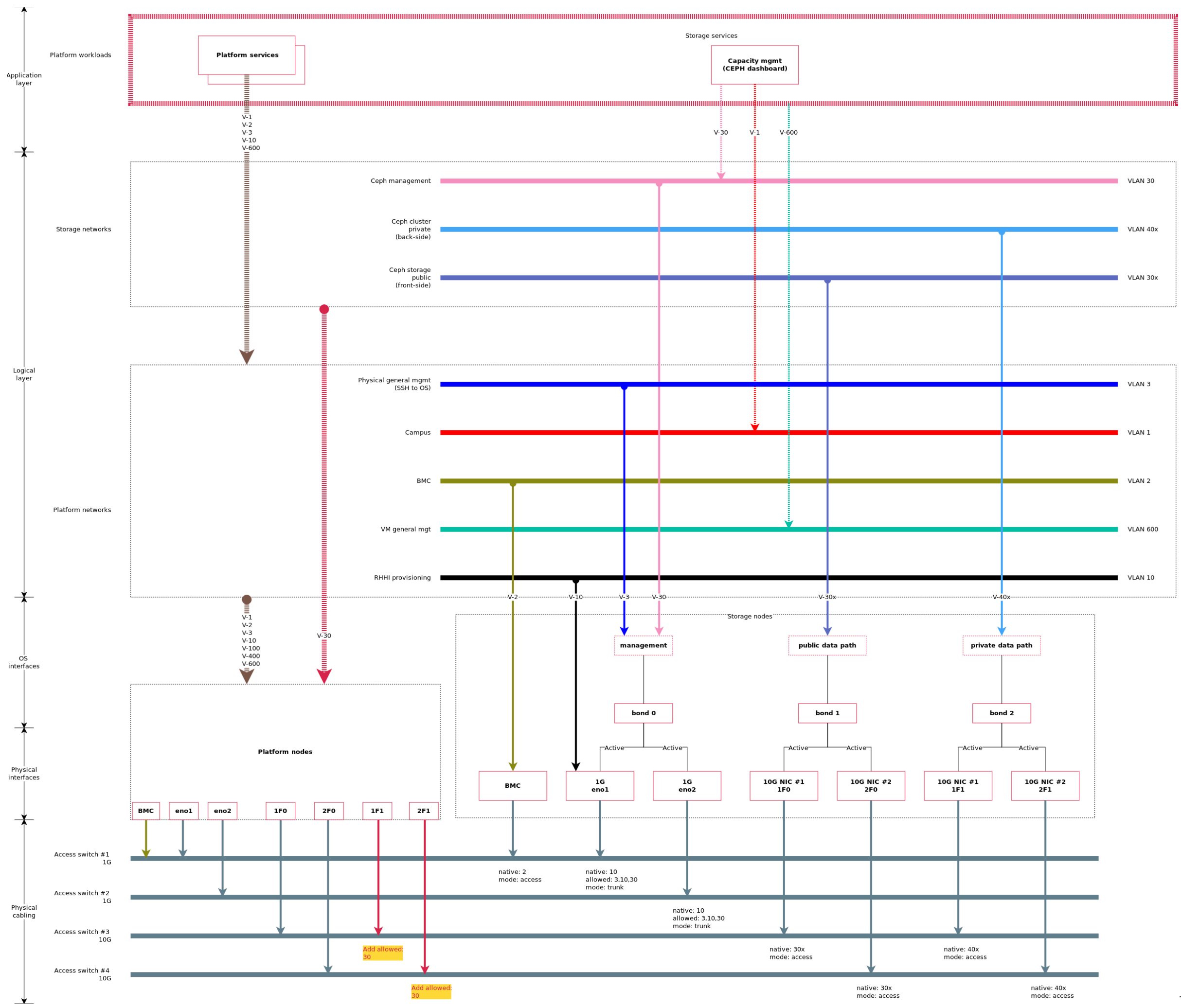


Figure 10: Lenovo Open Cloud Plaform Server Network Interfaces

6.4 Storage networks



{.col .s12}

6.4.1 Storage VLANs

Ceph management is communication between Ceph dashboard and Ceph nodes, e.g. RPC, transferring zabbix monitoring data.

Ceph storage public Data transferring between OpenStack nodes (both controller nodes and compute nodes) and Ceph nodes.

Ceph cluster private Ceph private data transferring, e.g. rebalancing.

Table; Storage Network VLANs

Network	VLAN	Subnet	Addresses	Mask	Static / DHCP	Gateway
Ceph management	30	192.168.30.x	254	/24	static	192.168.30.1
Ceph storage public	30x	192.168.3.x	254	/24	static	192.168.3.1
Ceph cluster private	40x	192.168.4.x	254	/24	static	192.168.4.1

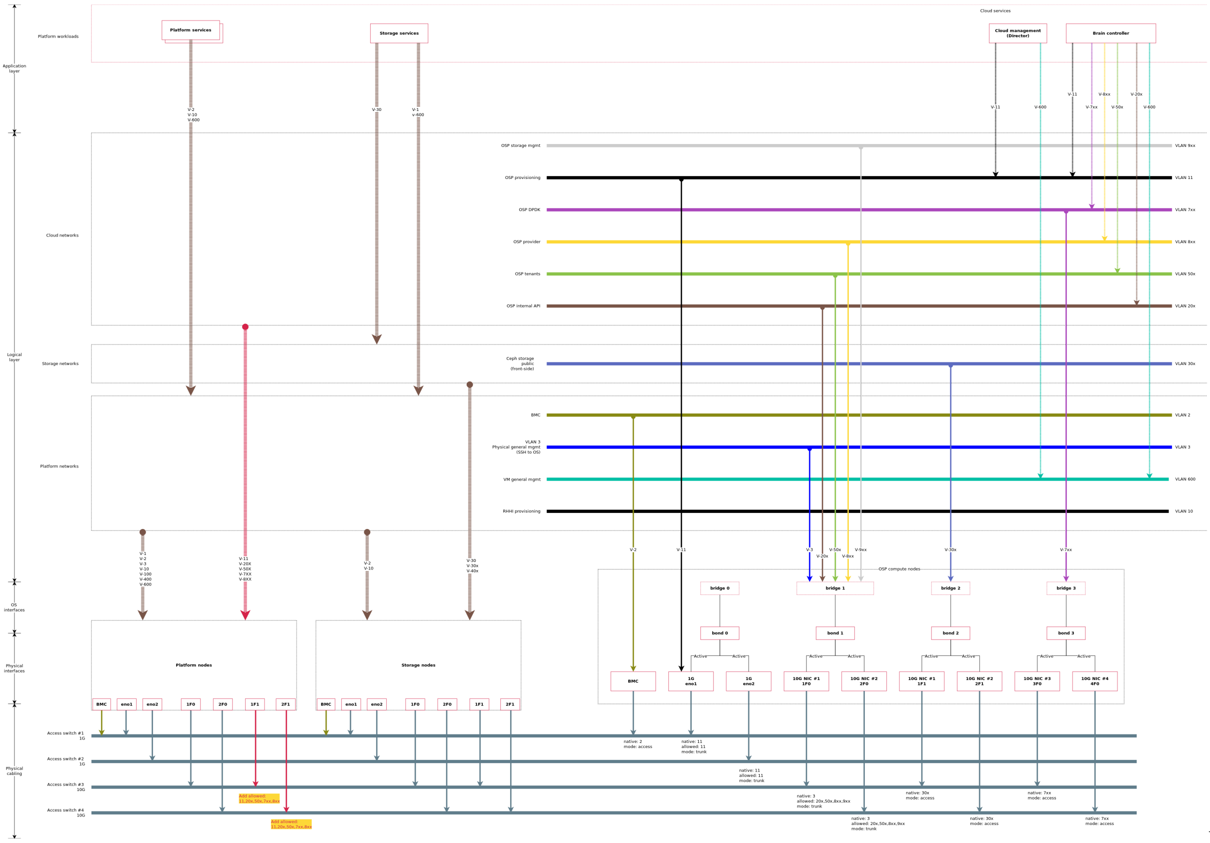
6.4.2 Storage server's NIC to VLAN mapping

Table 6: Storage server NIC to VLAN mapping

Network	VLAN	BMC	2 x 1G	2 x 10G	2 x 10G
BMC	2	x			
Ceph management	30		x		
Ceph storage public	30x			x	
Ceph cluster private	40x				x

6.4.3 Configure Storage server network interfaces

6.5 Cloud networks



{.col .s12}

6.5.1 Cloud VLANs

6.5.2 Cloud server's NIC to VLAN mapping

6.5.3 Configure Cloud server network interfaces

7 Configure Switches

There are two aspects of switch configurations:

1. **inter switch connections**: switch are connected to form a topology allowing data traffic between Lenovo Open Cloud environment and its host environments, and between LOC switches within the LOC itself. All switches are paired for high availability.
2. **server connections**: are connections between server and switch. Except out-of-band connection which has only one connection between a server and a switch, thus does not have redundancy, all other server to switch connections are in pairs.

7.1 switch to switch

7.1.1 cable schema

7.1.2 port configurations

7.2 switch to server

7.2.1 Platform servers

7.2.1.1 Platform server to switch cable schema

Each environment is different. Here we present an example cable schema following the network designs laid out in previous sections. In the following sections we will use this schema to demonstrate switch port configurations.

Port	1G switch	1G switch	10G switch	10G switch
1	server 1 BMC		server 1 bond 1	server 1 bond 1
2	server 2 BMC		server 2 bond 1	server 2 bond 1
3	server 3 BMC		server 3 bond 1	server 3 bond 1
17	server 1 bond 0	server 1 bond 0		
18	server 2 bond 0	server 2 bond 0		
19	server 3 bond 0	server 3 bond 0		

7.2.1.2 Platform server switch port configurations

7.2.2 Storage servers

7.2.2.1 Storage server to switch cable schema

7.2.2.2 Storage server switch port configurations

7.2.3 Cloud servers

7.2.3.1 Cloud server to switch cable schema

7.2.3.2 Cloud server switch port configurations

8 Configure Virtual Machines network interfaces

9 Appendix

9.1 Implementation Worksheet (questioinnaire)

9.2 Hardware BOM

Simplified version covers server & switch at high level should be fine.

9.3 Software BOM

BOM matrix without $$.