OpenStack Reference Architecture for Service Providers

Last update: 13 June 2019

Version 2.1

Describes the reference architecture for telco service providers based on Red Hat OpenStack 13

Describes Lenovo ThinkSystem servers, networking, and systems management software

Describes OVS-DPDK and SR-IOV configurations and benchmarks for a high performance solution

Includes validated and tested deployment and sizing guidance

Lijun Gu
Bin Zhou
Brahmanand Gorti
Miroslav Halas
Mike Perks
Shuang Yang

Table of Contents

1 Introduction

OpenStack continues to gain significant traction in the industry because of the growing adoption of cloud usage and the flexibility OpenStack offers as an open source product. The scope of OpenStack has grown in the past few years to span Communications Service Providers (CSPs).

Traditional networking is undergoing a major technological change where high-volume server platforms running virtual machines known as Virtual Network Functions (VNFs) are rapidly replacing purpose-built network appliances. This is called Network Functions Virtualization (NFV) and telecom operators and large enterprises stand to benefit from this trend. OpenStack is proving to be the virtualization platform of choice when service providers deploy NFV.

Comparing with traditional IT data centers, the NFV architecture requires high network bandwidth and compute capacity to deliver maximum packet throughput and low latencies. The Data Plane Development Kit (DPDK) and Single-Root Input/Output Virtualization (SR-IOV) are two of the key network acceleration techniques used by NFV.

This document describes the Lenovo NFVI Reference Architecture (RA) for CSPs by using Red Hat OpenStack Platform and DPDK and SR-IOV enabled on Lenovo hardware including industry leading servers, storage, networking, and Physical Infrastructure Management (PIM) tools from Lenovo.

Lenovo and Red Hat have collaborated to promote best practices and validate reference architecture for deploying private cloud infrastructure by leveraging the Red Hat OpenStack Platform 13. Red Hat OpenStack Platform 13 is offered with 3 years of production support with the option to purchase an extended lifecycle support (ELS) for 4th and 5th years.

The Lenovo NFVI platform provides an ideal infrastructure solution for NFV deployments. Lenovo servers provide the full range of form factors, features and functions that are necessary to meet the needs of small operators all the way up to large service providers. Lenovo uses industry standards in systems management on all these server platforms and enables seamless integration into cloud management tools such as OpenStack. Lenovo also provides data center network switches that are designed specifically for robust, scale-out server configurations and converged storage interconnect fabrics.

The Lenovo XClarity™ Administrator solution consolidates systems management across multiple Lenovo servers that span the data center. XClarity, which serves as PIM in an NFV deployment, enables automation of firmware updates on servers via compliance policies, patterns for system configuration settings, hardware inventory, bare-metal OS and hypervisor provisioning, and continuous hardware monitoring. XClarity easily extends via the published REST API to integrate into other management tools.

The intended audience for this document is IT and networking professionals, solution architects, sales engineers, and consultants. Readers should have a basic knowledge of Red Hat Enterprise Linux and OpenStack.

2 Business problem and business value

This chapter outlines the value proposition of Lenovo NFVI solution for communication service providers (CSPs).

2.1 Business problem

CSPs are looking to transform their infrastructure to make it better suited for supporting 5G and IoT (Internet of Things) deployments. Data traffic over communications networks is growing rapidly with the trend expected to continue for the foreseeable future. As a result, CSPs are evaluating and migrating to NFV at the data center and extending it to the central office and the network edge. Selecting next-generation infrastructure that enables CSP workloads to run in a performance-optimized, secure and cost-effective manner is a challenge.

2.2 Business value

This reference architecture provides a blueprint for accelerating the design, piloting, and deployment of NFV by outlining a validated Lenovo NFVI configuration that scales and delivers enterprise-level reliability across servers, storage and networking. This solution presented in this Reference Architecture provides the following benefits:

- Consolidated and fully integrated hardware resources with balanced workloads for compute, network, and storage.
- Guidelines to configure OVS-DPDK and SR-IOV as distributed in Red Hat OpenStack Platform 13.
- Elimination of single points of failure in every layer by delivering continuous access to virtual machines (VMs).
- Hardware redundancy and full utilization.
- Rapid OpenStack cloud deployment, including updates, patches, security, and usability enhancements with enterprise-level support from Red Hat and Lenovo.
- Unified management and monitoring for VMs.

3 Requirements

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

3.1 Functional requirements

Table 1 lists the functional requirements for a cloud solution implementation.

Table 1. Functional requirements

Requirement	Description	Supported by
Mobility	Workload is not tied to any physical location	• Enabled VM is booted from distributed storage and runs on different hosts • Live migration of running VMs • Rescue mode support for host maintenance
Resource provisioning	Physical servers, virtual machines, virtual storage, and virtual network can be provisioned on demand	• OpenStack compute service • OpenStack block storage service • OpenStack network service • OpenStack bare-metal provisioning service
Management portal	Web-based dashboard for workloads management	• OpenStack dashboard (Horizon) for most routine management operations
Multi-tenancy	Resources are segmented based on tenancy	• Built-in segmentation and multi-tenancy in OpenStack
Metering	Collect measurements of used resources to allow billing	• OpenStack metering service (Ceilometer)

3.2 Non-functional requirements

Table 2 lists the non-functional requirements for a service provider cloud implementation.

Table 2. Non-functional requirements

Requirement	Description	Supported by
OpenStack environment	Supports the current OpenStack edition	• OpenStack Queens release through Red Hat OpenStack Platform 13
Scalability	Solution components can scale for growth	• Compute nodes and storage nodes can be scaled independently within a rack or across racks without service downtime

Requirement	Description	Supported by
Load balancing	Workload is distributed evenly across servers	• Use of OpenStack scheduler for balancing compute and storage resources • Data blocks are distributed across storage nodes and can be rebalanced on node failure
High availability	Single component failure will not lead to whole system unavailability	• Hardware architecture ensures that computing service, storage service, and network service are automatically switched to remaining components • Network Interfaces are bonded to prevent single point of failures in the data path and control path • Controller node, and storage node are redundant • Data is stored on multiple servers and accessible from any one of them; therefore, no single server failure can cause loss of data • Virtual machines are persistent on shared storage service
Mobility	VM can be migrated or evacuated to different hosting server	• VM migration • VM evacuation
Physical footprint	Compact solution	• Lenovo ThinkSystem server, network devices, and software are integrated into one rack with validated performance and reliability • Provides 1U compute node option
Ease of installation	Reduced complexity for solution deployment	• A dedicated deployment server with web-based deployment tool and rich command line provide greater flexibility and control over how you deploy OpenStack in your cloud • Optional deployment services

Requirement	Description	Supported by
Support	Available vendor support	• Hardware warranty and software support are included with component products • Standard or Premium support from Red Hat included with Red Hat OpenStack Platform and Optional Red Hat Ceph Storage subscription
Flexibility	Solution supports variable deployment methodologies	• Hardware and software components can be modified or customized to meet various unique customer requirements • Provides local and shared storage for workload
Robustness	Solution continuously works without routine supervision	• Red Hat OpenStack Platform 13 is integrated and validated on Red Hat Enterprise Linux 7.5 • Integration tests on hardware and software components
Security	Solution provides means to secure customer infrastructure	• Security is integrated in the Lenovo ThinkSystem hardware with ThinkSystem Trusted Platform Assurance, an exclusive set of industry-leading security features and practices • SELinux is enabled and in enforcing mode by default in Red Hat OpenStack Platform 10 and above versions. • Network isolation using virtual LAN (VLAN) and virtual extensible LAN (VXLAN).
High performance	Solution components are high-performance	• Provides 40 - 90 average workloads (2 vCPU, 8 GB vRAM, 80 GB disk) per host. • OVS-DPDK or SR-IOV can be enabled on NFV compute nodes to provide high-throughput and low latency.

4 NFV overview and key technologies

The Lenovo NFVI solution is engineered and validated with Red Hat OpenStack Platform 13. It encompasses software, hardware, along with integration of the solution components. The core software includes the Red Hat OpenStack Platform, Red Hat Ceph Storage, Lenovo XClarity Administrator, and Red Hat CloudForms. This specific software stack has been verified with the Lenovo ThinkSystem SR650 and SR630 servers and Lenovo RackSwitch G8052 and NE2572 top of rack switches.

4.1 NFV architecture

The Lenovo NFVI solution follows the European Telecommunications Standards Institute (ETSI) NFV architectural framework. The NFV Infrastructure (NFVI) provides a stable platform with hardware and software components optimized for the upper layer VNF ecosystem. The NFVI provides a multi-tenant infrastructure that leverages standard virtualization technologies that support multiple use cases and applications simultaneously.

The NFV-MANO layer provides the capability of NFV management and orchestration (MANO) so that software implementation of network functions can be decoupled from the compute, storage, and network resources provided via NFVI. The NFV Orchestrator (NFVO) performs orchestration functions of NFVI resources across multiple Virtual Infrastructure Managers (VIMs) and lifecycle management of network services. The VNF Manager (VNFM) performs orchestration and management functions of VNFs. The VIM performs orchestration and management functions of NFVI resources within a domain.

Figure 1 shows the architecture of the Lenovo NFVI solution along with other components that make up the entire NFV stack.

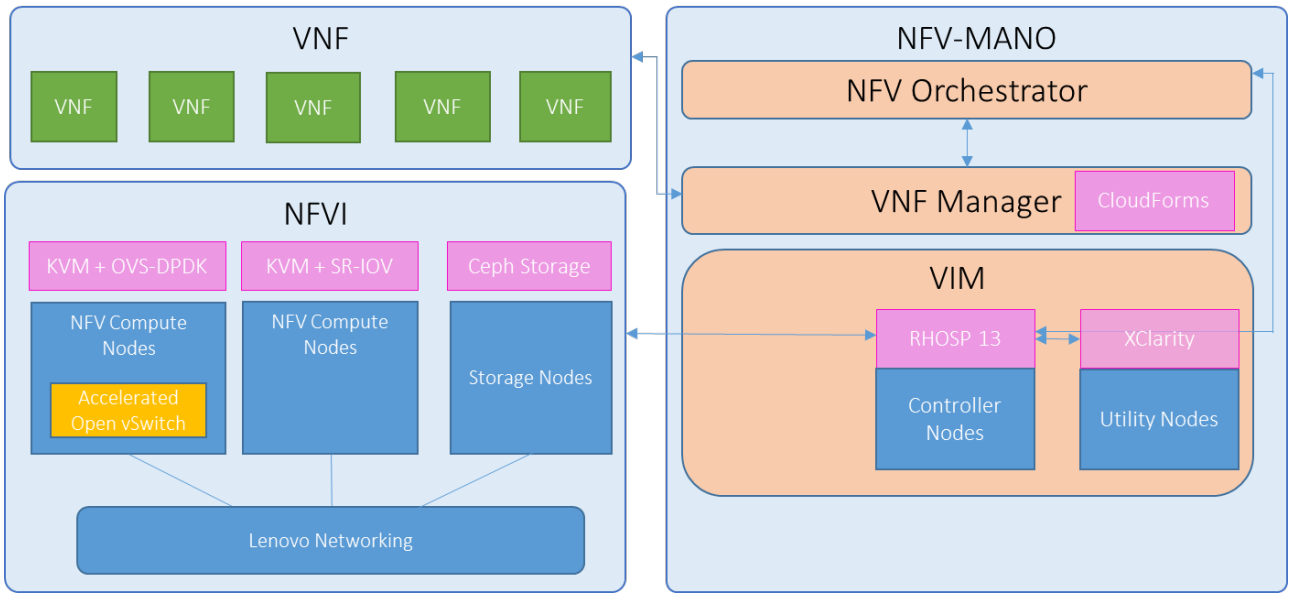


Figure 1. Overview of Lenovo NFV Platform

The Lenovo NFVI solution is composed of various components, which CSPs can deploy according to their specific use-case requirements.

The NFVI layer is based on several hardware components provided by Lenovo. The Compute Nodes for NFV usually enable accelerated software IO technologies such as OVS-DPDK, SR-IOV, CPU pinning and NUMA awareness to provide advanced IO performance for VNF applications when combined with Lenovo supported hardware NICs such as the Intel XXV710 25GbE adapters.

The Lenovo NFVI solution supports various options to provide highly available persistent storage for the NFV cloud. The storage solutions include:

- Deployment without a Ceph Cluster
- Deployment with a Red Hat Ceph Storage Cluster
- Deployment utilizing a Hyperconverged Ceph Cluster

The option to deploy without a Ceph Cluster is targeted towards users who do not have a need for persistent storage backing their VM workloads and users that do not wish to use compute resources for a Ceph Cluster. It leverages additional physical disks on OpenStack Controller Nodes to build a Swift cluster to provide backing storage for OpenStack images, metrics and backup data while user workloads run using the ephemeral storage on the compute nodes.

A typical selection of storage solution is to deploy a dedicated Ceph Cluster as the storage backend of the NFV cloud. The Ceph Cluster is highly scalable and provides unified storage interfaces, which makes it suitable for various user scenarios for block, object, and file storage.

The Hyperconverged Ceph storage option is targeted at users who have a need for persistent data storage for their VMs but do not want or have space for additional storage nodes in their infrastructure. Additional disks from the compute nodes is used to provide highly available, persistent storage for workloads and the rest of the OpenStack deployment. The deployment collocates the Ceph on the compute nodes and configures the nodes for optimal utilization of resources. But for compute nodes that will have demanding NFV workloads running on it, it is not recommended to adopt a hyperconverged infrastructure.

In this Lenovo NFVI architecture, Red Hat Ceph Storage is used to provide block, image, and snapshots as well as object storage service via Storage Nodes, which is optimized for disk capacity to support up to twenty-four drives combining NVMe, SSD and HDD backed storage tiers.

4.2 Key NFVI technologies

The Lenovo NFVI solution can be configured and tuned to take advantage of high performance network capabilities provided by its individual software and hardware components. This results in optimal value and performance for demanding VNF workloads. Below are some key technologies adopted in Lenovo NFVI solution.

4.2.1 SR-IOV

Single root I/O virtualization (SR-IOV) is an extension to the PCI Express (PCIe) specification. SR-IOV enables a single PCIe device to appear as multiple separate PCI devices. SR-IOV enabled devices have the ability to dedicate isolated access to its resources, e.g. “Virtual Functions” (VF). These VFs are later assigned to the virtual machines (VMs), which allow direct memory access (DMA) to the network data. VM guests can gain the performance advantage of direct PCI device assignment, while only using single VF on the physical NIC. Comparing with traditional virtualized environment without SR-IOV, where a packet has to go through an

extra layer of the hypervisor, SR-IOV eliminates extra multiple CPU interrupts per packet at hypervisor layer as well as virtual switch layer, As a result, guest VMs connected to virtual functions are able to achieve low latency and near-line wire speed.

4.2.2 OVS-DPDK

Open vSwitch (OVS) is an open source software switch that is commonly used as a virtual switch within a virtualized server environment. OVS supports the capabilities of a regular L2-L3 switch and provides support to the SDN protocols such as OpenFlow to create user-defined overlay networks. OVS uses Linux kernel networking to switch packets between virtual machines and across hosts using physical NIC.

DPDK is a set of data plane libraries and NIC drivers for fast packet processing. DPDK applications rely on the poll mode driver (PMD) which runs a dedicated user space process (or thread) to poll the queue of the NIC. When packets arrive, the PMD will receive them continuously. If there are no packets, the PMD will just poll in an endless loop, which usually results in high power consumption. In this way, the packets processing traverses the hypervisor's kernel and IP (Internet Protocol) stack.

Open vSwitch can be bundled with DPDK for better performance, resulting in a DPDK-accelerated OVS (OVS-DPDK). Red Hat OpenStack Platform supports OVS-DPDK deployment on compute nodes so that NFV workloads can benefit from the DPDK data path at an application level. Recent advanced features of OVS-DPDK, such as the vHost multi-queue provides further improvement in the network throughput of guest VMs. At a high level, OVS-DPDK replaces the standard OVS kernel data path with a DPDK based data path, and creates a user-space vSwitch on the host, which is using DPDK internally for its packet forwarding. The architecture is mostly transparent to users since the basic OVS features as well as the interfaces (such as OpenFlow, OVSDB, and command line) remain mostly the same.

Figure 2 shows the diagram of IO data path for VM applications running on Red Hat OpenStack Platform.

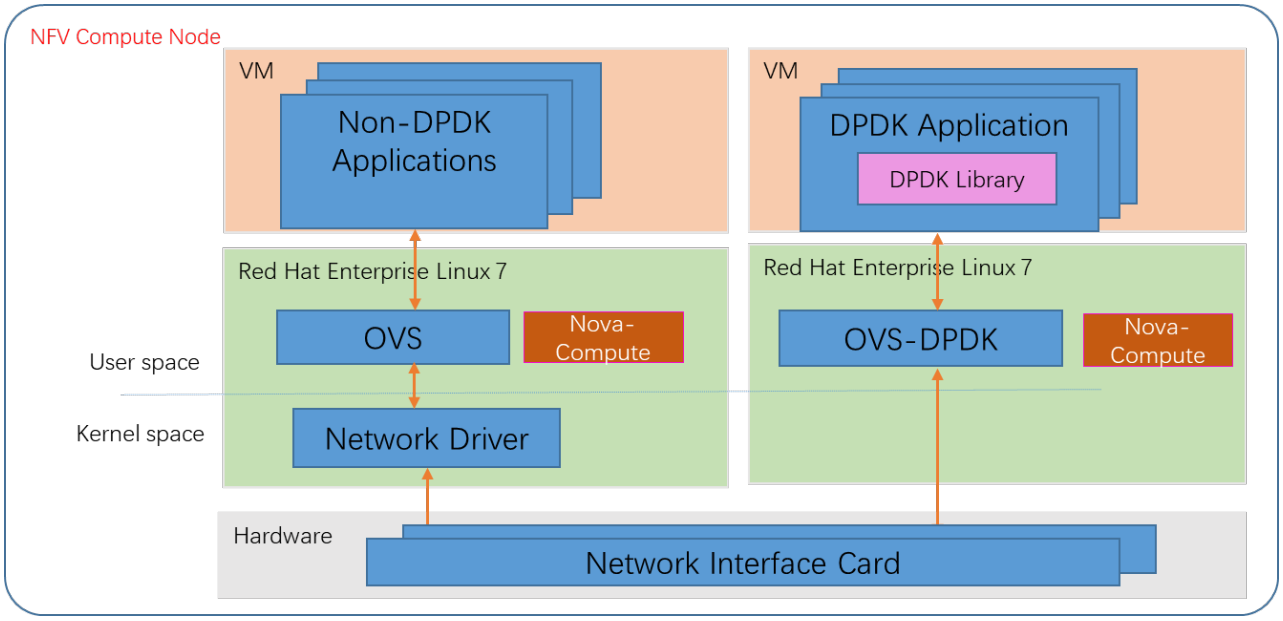


Figure 2. IO Data Path: OVS-DPDK vs. Non-DPDK

OVS-DPDK requires one or more cores dedicated to the poll mode thread and thus isolates the allocated CPU cores from being scheduled for other tasks. This will avoid context switching and reduce the cache miss

rate. OVS-DPDK runs in user space and allows applications to access raw traffic data from NIC without interacting with Linux kernel and kernel IP stack thus providing high I/O performance for VNF workloads..

4.2.3 NUMA awareness

NUMA, or Non-Uniform Memory Access, is a shared memory architecture that describes the placement of main memory modules with respect to processors in a multiprocessor system. When planning NFVI deployment, the Cloud Administrator needs to understand the NUMA topology of Compute node to partition the CPU and memory resources for optimum performance. For example, the following NUMA topology information should be collected at hardware introspection:

- RAM (in kilobytes)
- Physical CPU cores and their sibling threads
- NICs associated with the NUMA node

To achieve high performance in NFVI environment, Cloud Admin need to partition the resources between the host and the guest. For SR-IOV Compute nodes and OVS-DPDK Compute nodes, the partition of resources will be different although it has to follow the same principle: the packets traversing through entire data path should only go through the same NUMA node. The VNFs should use NICs associated with the same NUMA node that they use for memory and CPU pinning.

Figure 3 shows an example of the Lenovo recommended NUMA partitioning on a dual socket Compute nodes. In the case of OVS-DPDK deployment, OVS-DPDK performance depends on reserving a block of memory and CPUs pinned for the PMDS local to the NUMA node. Three VNFs use NICs associated with the same NUMA node. Also it is recommended that both interfaces in a bond are from NICs on the same NUMA node. In the case of SR-IOV deployment, no memory and CPUs are required for PMD and thus leave more VCPUs available in the pool for other VNFs.

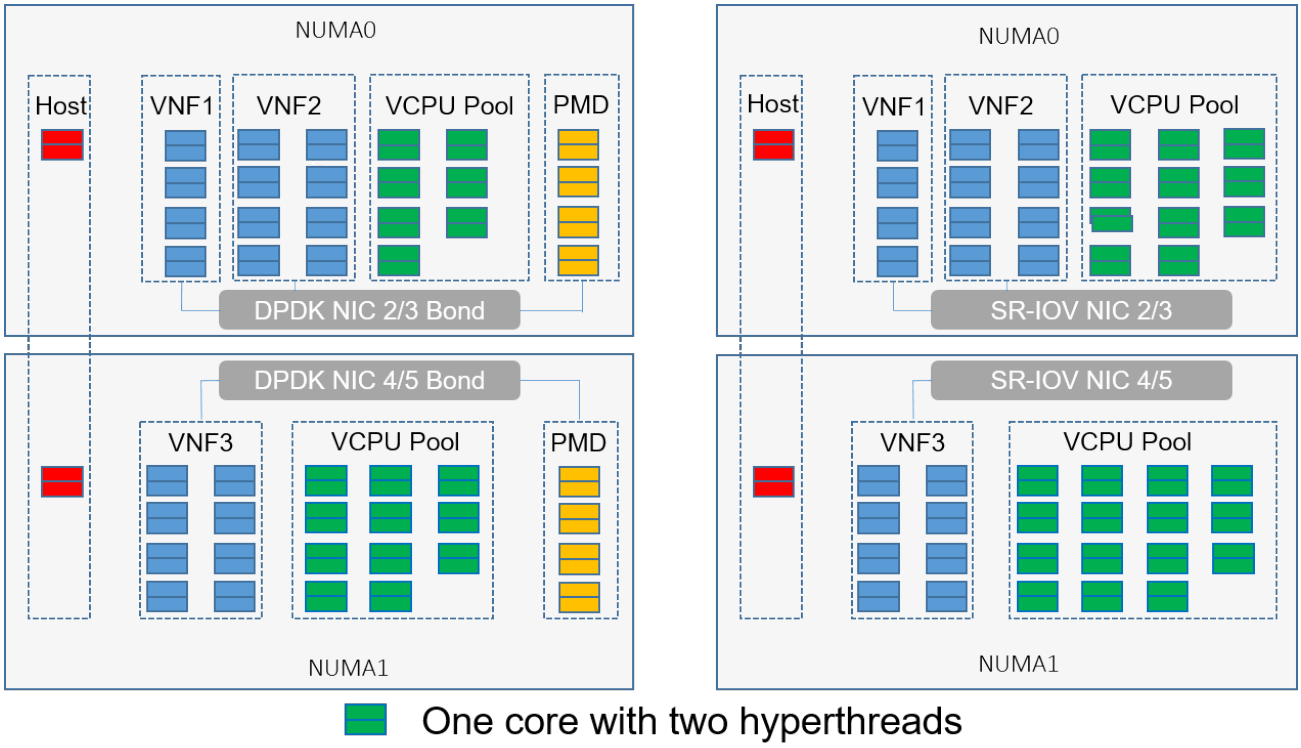


Figure 3. NUMA Nodes Partitioning for DPDK and SR-IOV Deployment

4.2.4 CPU pinning

CPU pinning refers to reserving physical cores for specific VNF guest instances or host processes. The configuration is implemented in two parts: ensuring that virtual guests can only run on dedicated cores; ensuring that common host processes do not run on those cores.

CPU Pinning is often used together with DPDK applications and PMD. The DPDK applications and PMD use the thread affinity feature of the Linux kernel to bind the thread to a specific core in order to avoid context switching.

Figure 4 shows an example of CPU Pinning for both host processes and guest vCPUs. The Kernel configuration parameter *isolcpus* specifies a list of physical CPU cores (in orange color) isolated from the kernel scheduler. Host processes scheduled by kernel (in red color) do not run on isolated CPU cores. In the case of SR-IOV deployment, no host processes are required for cpu pinning. For OVS-DPDK deployment, DPDK PMDs should be pinned to the isolated physical CPU cores (in orange color). In general, all VNFs should use CPUs exclusively to ensure the performance SLA.

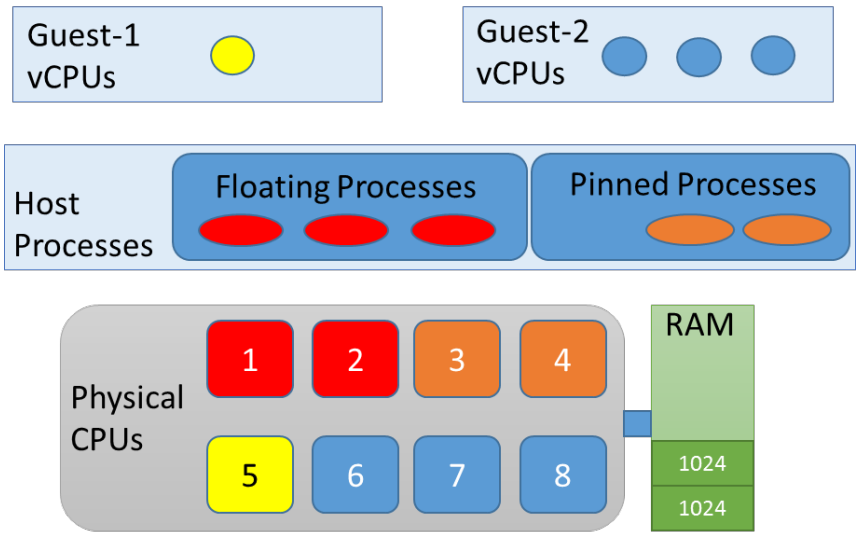


Figure 4. Example of CPU Pinning

4.2.5 Huge pages

In general, the CPU allocates RAM by pages, e.g. chunks of 4K bytes. Modern CPU architectures support a much larger page size. In an NFVI environment, *hugepages* support is required for the large memory pool allocation used for data packet buffers. By using *hugepages* allocations, performance is increased since fewer pages are needed, and therefore fewer Translation Lookaside Buffers (TLB) lookups. This in turn reduces the time it takes to translate a virtual page address to a physical page address. Without *hugepages* enabled as kernel parameter high TLB miss rates would occur thereby slowing performance.

Huge pages are typically enabled together with DPDK and NUMA CPU pinning to provide accelerated high data path performance.

5 Component model

The section describes the components used in OpenStack specifically the Red Hat OpenStack Platform distribution.

5.1 Core Red Hat OpenStack Platform components

Figure 5 shows the core components of Red Hat OpenStack Platform. It does not include some optional add-on components listed in the “Third-party components” section on page 13.

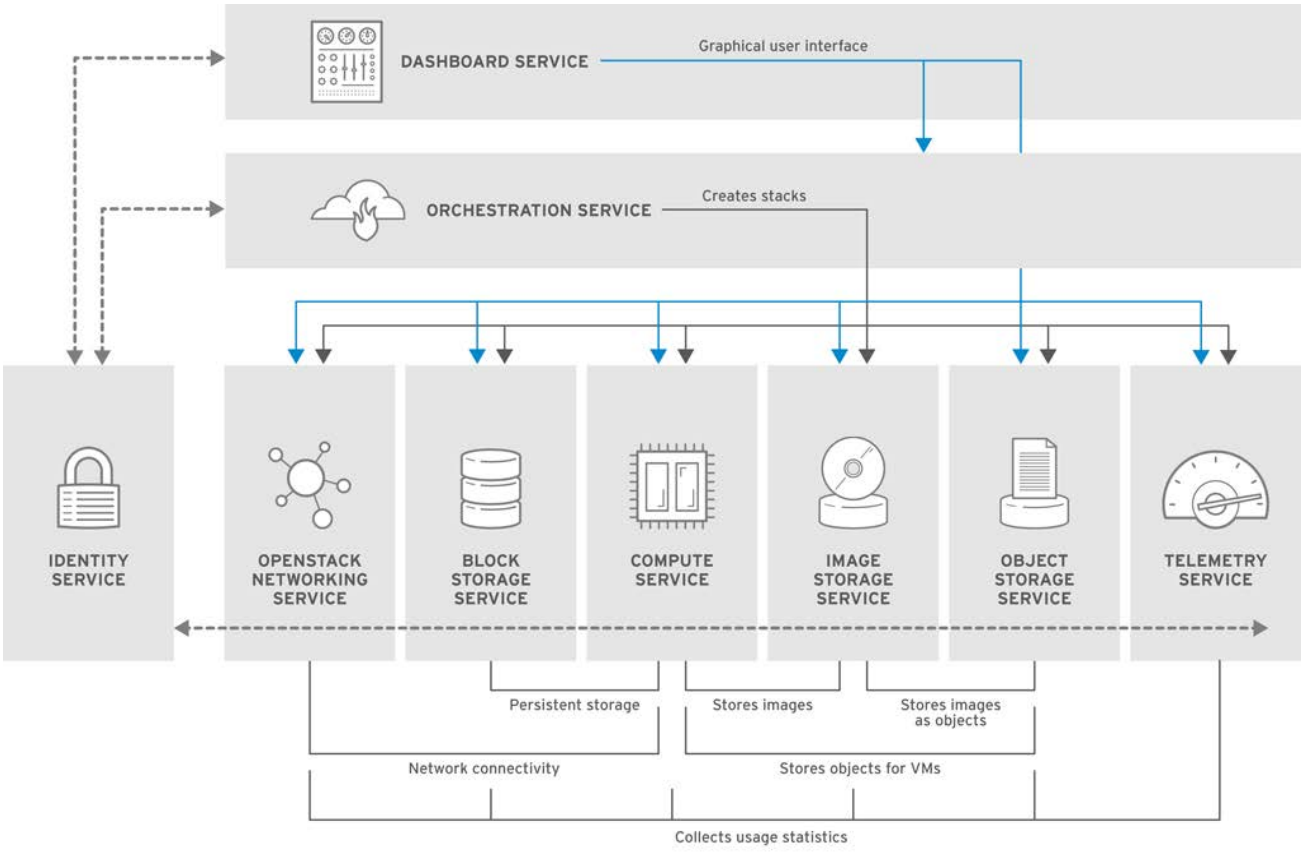


Figure 5. Components of Red Hat OpenStack Platform

Table 3 lists the core components of Red Hat OpenStack Platform as shown in Figure 5.

Table 3. Core components

Component	Code name	Description
Compute service	Nova	Provisions and manages VMs, which creates a redundant and horizontally scalable cloud-computing platform. It is hardware and hypervisor independent and has a distributed and asynchronous architecture that provides HA and tenant-based isolation.

Component	Code name	Description
Block storage service	Cinder	Provides persistent block storage for VM instances. The ephemeral storage of deployed instances is non-persistent; therefore, any data generated by the instance is destroyed after the instance terminates. Cinder uses persistent volumes attached to instances for data longevity, and instances can boot from a Cinder volume rather than from a local image.
Networking service	Neutron	OpenStack Networking is a pluggable “networking as a service” framework for managing networks and IP addresses. This framework supports several flexible network models, including Dynamic Host Configuration Protocol (DHCP) and VLAN.
Image service	Glance	Provides discovery, registration, and delivery services for virtual disk images. The images can be stored on multiple back-end storage units and cached locally to reduce image staging time.
Object storage service	Swift	Provides cloud storage software built for scale and optimized for durability, availability, and concurrency across the entire data set. It can store and retrieve data with a simple API, and is ideal for storing unstructured data that can grow without bound. (Red Hat Ceph Storage is used in this reference architecture, instead of Swift, to provide the object storage service)
Identity service	Keystone	Centralized service for authentication and authorization of OpenStack service and for managing users, projects and roles. Identity supports multiple authentication mechanisms, including user.
Telemetry service	Ceilometer	Provides infrastructure to collect measurements within OpenStack. Delivers a unique point of contact for billing systems to acquire all of the needed measurements to establish customer billing across all current OpenStack core components. An administrator can configure the type of collected data to meet operating requirements. Gnocchi is a multi-tenant, metrics and resource database that is as ceilometer backend.
Dashboard service	Horizon	Dashboard provides a graphical user interface for users and administrator to perform operations such as creating and launching instances, managing networking, and setting access control.
Orchestration	Heat	Heat is an orchestration engine to start multiple composite cloud applications based on templates in the form of text files. AWS CloudFormation templates are one example.
File Share Service	Manila	A file share service that presents the management of file shares (for example, NFS and CIFS) as a core service to OpenStack.

Table 4 lists the optional components in the Red Hat OpenStack Platform release. Actual deployment use cases will determine when and how these components are used.

Table 4. Optional components

Component	Code name	Description
Bare-metal provisioning service	Ironic	OpenStack Bare Metal Provisioning enables the user to provision physical, or bare metal machines, for a variety of hardware vendors with hardware-specific drivers.
Data Processing	Sahara	Provides the provisioning and management of Hadoop cluster on OpenStack. Hadoop stores and analyzes large amounts of unstructured and structured data in clusters.

Table 5 lists the OpenStack concepts to help the administrator further manage the tenancy or segmentation in a cloud environment.

Table 5. OpenStack tenancy concepts

Name	Description
Tenant	The OpenStack system is designed to have multi-tenancy on a shared system. Tenants (also called projects) are isolated resources that consist of separate networks, volumes, instances, images, keys, and users. These resources can have quota controls applied on a per-tenant basis.
Availability Zone	In OpenStack, an availability zone allows a user to allocate new resources with defined placement. The “instance availability zone” defines the placement for allocation of VMs, and the “volume availability zone” defines the placement for allocation of virtual block storage devices.
Host Aggregate	A host aggregate further partitions an availability zone. It consists of key-value pairs assigned to groups of machines and used by the scheduler to enable advanced scheduling.
Region	Regions segregate the cloud into multiple compute deployments. Administrators use regions to divide a shared-infrastructure cloud into multiple sites each with separate API endpoints and without coordination between sites. Regions share the Keystone identity service, but each has a different API endpoint and a full Nova compute installation.

5.2 Third-party components

Table 6 lists the following third-party components can also be used:

Table 6. Third-party components

Name	Description
MariaDB	MariaDB is open source database software shipped with Red Hat Enterprise Linux as a replacement for MySQL. MariaDB Galera cluster is a synchronous multi-master cluster for MariaDB. It uses synchronous replication between every instance in the cluster to achieve an active-active multi-master topology, which means every instance can accept data retrieving and storing requests and the failed nodes do not affect the function of the cluster.
RabbitMQ	RabbitMQ is a robust open source messaging system based on the AMQP standard, and it is the default and recommended message broker in Red Hat OpenStack Platform.
Memcached	Memcached is high-performance, distributed in-memory a key-value store for small chunks of arbitrary data to speed up dynamic web applications by reducing the database load.
Redis	Redis is an open-source in memory database that provides the alternative solution to Memcached for web application performance optimization

5.3 Red Hat Ceph Storage component

Red Hat Ceph Storage is open source software from Red Hat that provides Exabyte-level scalable object, block, and file storage from a completely distributed computer cluster with self-healing and self-managing capabilities. Red Hat Ceph Storage virtualizes the pool of the block storage devices and stripes the virtual storage as objects across the servers.

Red Hat Ceph Storage is integrated with Red Hat OpenStack Platform. The OpenStack Cinder storage component and Glance image services can be implemented on top of the Ceph distributed storage.

OpenStack users and administrators can use the Horizon dashboard or the OpenStack command-line interface to request and use the storage resources without requiring knowledge of where the storage is deployed or how the block storage volume is allocated in a Ceph cluster.

The Nova, Cinder, Swift, and Glance services on the controller and compute nodes use the Ceph driver as the underlying implementation for storing the actual VM or image data. Ceph divides the data into placement groups to balance the workload of each storage device. Data blocks within a placement group are further distributed to logical storage units called Object Storage Devices (OSDs), which often are physical disks or drive partitions on a storage node.

The OpenStack services can use a Ceph cluster in the following ways:

- VM Images: OpenStack Glance manages images for VMs. The Glance service treats VM images as immutable binary blobs and can be uploaded to or downloaded from a Ceph cluster accordingly.
- Volumes: OpenStack Cinder manages volumes (that is, virtual block devices) attached to running VMs or used to boot VMs. Ceph serves as the back-end volume provider for Cinder.

- Object Storage: OpenStack Swift manages the unstructured data that can be stored and retrieved with a simple API. Ceph serves as the back-end volume provider for Swift.
- VM Disks: By default, when a VM boots, its drive appears as a file on the file system of the hypervisor. Alternatively, the VM disk can be in Ceph and the VM started using the boot-from-volume functionality of Cinder or directly started without the use of Cinder. The latter option is advantageous because it enables maintenance operations to be easily performed by using the live-migration process, but only if the VM uses the RAW disk format.

If the hypervisor fails, it is convenient to trigger the Nova `evacuate` function and almost seamlessly run the VM machine on another server. When the Ceph back end is enabled for both Glance and Nova, there is no need to cache an image from Glance to a local file, which saves time and local disk space. In addition, Ceph can implement a copy-on-write feature ensuring the start-up of an instance from a Glance image does not actually use any disk space.

5.4 Red Hat OpenStack Platform specific benefits

Red Hat OpenStack 13 offers a better and easier OpenStack based cloud management platform built on Red Hat Enterprise Linux that includes three year of production support. For more details, please see: Red Hat OpenStack Platform Life Cycle and Red Hat OpenStack Platform Director Life Cycle.

Listed below are the most important new features in the Red Hat OpenStack Platform 13. For more details, please see: Red Hat OpenStack Platform 13 Release Notes.

- Fast forward upgrade path supported by Red Hat OpenStack Director, specifically from Red Hat OpenStack Platform 10 to Red Hat OpenStack Platform 13. This enables easy upgrade to current long life version from previous LTS version for customers.
- Controller nodes deployed in Red Hat Virtualization is now supported by Director node provisioning. New driver(staging-ovirt) is included in Director Bare Metal (ironic) service.
- Fully containerized services are provided. All Red Hat OpenStack Platform services are deployed as containers.
- L3 routed spine-leaf network is supported for director to provision and introspect nodes with multiple networks. This feature, in conjunction with composable networks, allows users to provision and configure a complete L3 routed spine-leaf architecture for the overcloud.
- Red Hat Ceph Storage 3.0 is the default supported version of Ceph for Red Hat OpenStack Platform. Red Hat Ceph Storage 2.x is continuously compatible with the newer Ceph client as external Ceph Storage. Storage nodes deployed by Director will be version 3.0 by default. The new Red Hat Ceph Storage 3.0 also supports scale out Ceph Metadata server(MDS) and RADOS gateway nodes.
- The Shared File System service(manila) of Red Hat OpenStack Platform 13 now supports mounting shared file systems backed by Ceph File System(CephFS) via the NFSv4 protocol. Multi-tenancy is supported.

- Full support for Real time KVM(RT-KVM) integration into the Compute service(nova). User will be benefited with lower latency for system calls and interrupts. Furthermore, RT-KVM compute nodes now support NFV workloads for workloads with latency requirements.
- Director supports to deploy Instance HA. Instance HA allows Red Hat OpenStack Platform to automatically evacuate and re-spawn instances on a different Compute node when their host Compute node fails. From Red Hat OpenStack Platform 13 and later, Instance HA is deployed and configured with the director.
- OpenDaylight is fully supported and can be enabled as the SDN networking backend for OpenStack. The OpenDaylight project NetVirt support OpenStack neutron APIs.
- OpenStack Key Manager(barbican) is enhanced to support encrypted volumes for Block Storage (cinder), and signed image for Image Service(glance).

6 Operational model

This section describes the operational model of the Lenovo NFVI solution based on Red Hat OpenStack Platform 13 using Lenovo hardware and software.

6.1 Hardware components

The following section describes the hardware components that recommended for the Lenovo NFVI solution.

6.1.1 Rack servers introduction

Lenovo recommends the following servers:

- Lenovo ThinkSystem SR650
- Lenovo ThinkSystem SR630

Lenovo ThinkSystem SR650

The Lenovo ThinkSystem SR650 server (as shown in Figure 6 and Figure 7) 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.

Figure 6. Lenovo ThinkSystem SR650 (with 24 x 2.5-inch disk bays)

Figure 7. 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 of workloads and performance that is targeted to lower the total cost of ownership (TCO) per VM. 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 hyperthread 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 makes the server ideal to run data and bandwidth intensive VNF workload and storage functions of NFVI platform.

For more information, see the following website: ThinkSystem SR650 Product Guide

Lenovo ThinkSystem SR630

The Lenovo ThinkSystem SR630 server (as shown in Figure 8) is an ideal 2-socket 1U rack server for small businesses up to large enterprises that need industry-leading reliability, management, and security, as well as maximizing performance and flexibility for future growth. The SR630 server is designed to handle a wide range of workloads, such as databases, virtualization and cloud computing, virtual desktop infrastructure (VDI), infrastructure security, systems management, enterprise applications, collaboration/email, streaming media, web, and HPC. It improves productivity by supporting up to two processors, 56 cores, and 112 threads, and and up to 3 TB of memory capacity with memory speed of up to 2666 MHz, which is capable to host Red Hat OpenStack Platform controller services. The Key difference ThinkSystem SR630 offers up to twelve 2.5-inch hot-swappable SAS/SATA HDDs or SSDs together with up to four on-board NVMe PCIe ports that allow direct connections to the U.2 NVMe PCIe SSDs.

The Lenovo ThinkSystem SR630 is ideal for OpenStack Controller and Compute nodes and utility nodes.

Figure 8. Lenovo ThinkSystem SR630

For more information, see the following website: ThinkSystem SR630 Product Guide

6.1.2 Network switches introduction

The following Top-of-Rack (ToR) switches are recommended for the Lenovo NFVI solution:

- Lenovo RackSwitch G8052
- Lenovo ThinkSystem NE0152T RackSwitch
- Lenovo RackSwitch G8272

- Lenovo ThinkSystem NE1032 RackSwitch
- Lenovo ThinkSystem NE1032T RackSwitch
- Lenovo ThinkSystem NE1072T RackSwitch
- Lenovo ThinkSystem NE2572 RackSwitch

The 10Gb and 25Gb Ethernet switches are used for the internal and external network of Red Hat OpenStack Platform cluster, and 1Gb Ethernet switch is used for out-of-band server management. The Networking Operating System software features of these Lenovo switches deliver seamless, standards-based integration into upstream switches.

Lenovo RackSwitch G8052

The Lenovo RackSwitch G8052 (as shown in Figure 9) 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 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.

Figure 9. Lenovo Rackswitch G8052

For more information, see the Rackswitch G8052 Product Guide

Lenovo ThinkSystem NE0152T Gigabit Ethernet switch

The Lenovo ThinkSystem NE0152T RackSwitch 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 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 is ideal for latency sensitive applications, such as high-performance computing clusters, financial applications and NFV deployments. In addition to the 10 Gb Ethernet (GbE) and 40GbE connections, the G8272 can use 1GbE connections.

Figure 11. Lenovo RackSwitch G8272

For more information, see the RackSwitch G8272 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 has 32x 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 NE1032 RackSwitch

For more information, see the ThinkSystem NE1032 Product Guide

The NE1032T RackSwitch 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 13. Lenovo ThinkSystem NE1032T RackSwitch

For more information, see the ThinkSystem NE1032T Product Guide

The NE1072T RackSwitch 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 14. Lenovo ThinkSystem NE1072T RackSwitch

For more information, see the ThinkSystem NE1072T Product Guide

Lenovo ThinkSystem NE2572 RackSwitch family

The Lenovo ThinkSystem NE2572 RackSwitch is designed for the data center and provides 10 Gb/25 Gb Ethernet connectivity with 40 Gb/100 Gb Ethernet upstream links. It is ideal for big data, cloud, and enterprise workload solutions. It is an enterprise class Layer 2 and Layer 3 full featured switch that delivers line-rate, high-bandwidth 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 software features) help provide high availability for business sensitive traffic.

The NE2572 RackSwitch has 48x SFP28/SFP+ ports that support 10 GbE SFP+ and 25 GbE SFP28 optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. The switch also offers 6x QSFP28/QSFP+ ports that support 40 GbE QSFP+ and 100 GbE QSFP28 optical transceivers, active optical cables (AOCs), and direct attach copper (DAC) cables. The QSFP28/QSFP+ ports can also be split out into two 50 GbE (for 100 GbE QSFP28), or four 10 GbE (for 40 GbE QSFP+) or 25 GbE (for 100 GbE QSFP28) connections by using breakout cables.

Figure 15. Lenovo ThinkSystem NE2572 RackSwitch

The NE2572 runs Lenovo Cloud Network Operating System (CNOS) with robust Layer 2/Layer 3 performance. It is ideal for Lenovo datacenter solutions, hybrid cloud designs, hyperscale networks, or high performance computing.

For more information, see the ThinkSystem NE2572 Product Guide

6.2 Deployment mapping of OpenStack software components

Because Red Hat OpenStack provides a great deal of flexibility, the operational model for Red Hat OpenStack Platform normally depends upon a thorough understanding of the requirements and needs of the cloud users to design the best possible configuration to meet the requirements. For more information about Red Hat OpenStack Platform 13, please refer to the Red Hat OpenStack Platform 13 Product. For NFV product information, please refer to Red Hat Network Functions Virtualization.

Figure 16 shows how the different software components relate to each other and form an OpenStack cluster with integrated management functions.

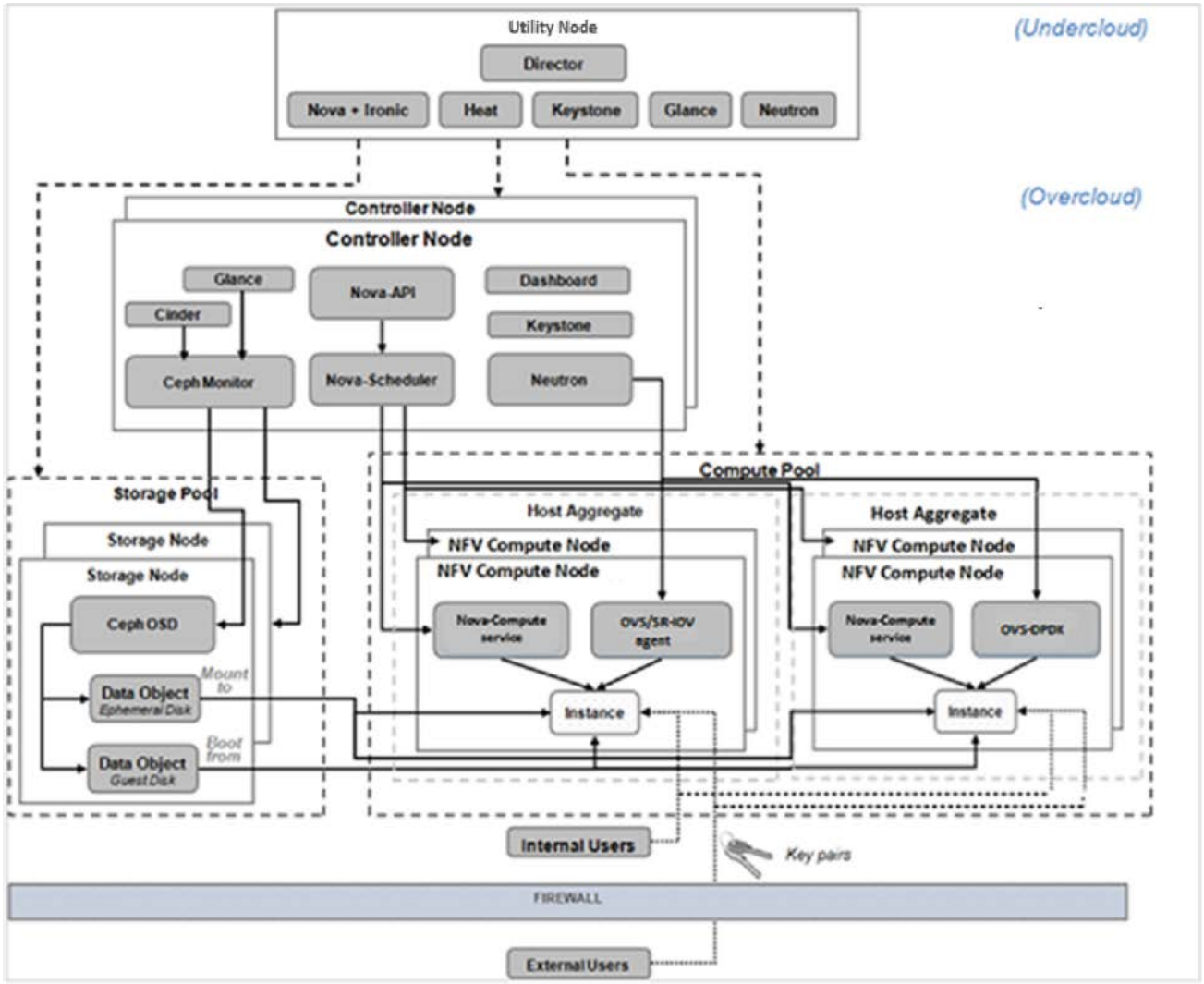


Figure 16. Deployment Model for Lenovo NFVI Solution Using Red Hat OpenStack Platform

A Red Hat OpenStack Platform consists of two clouds: an undercloud and an overcloud. The undercloud is a lifecycle management tool to provide environment planning, bare metal system control, images installation, configuration and deployment of the overcloud. The overcloud is the production and tenant facing cloud. The director sometimes is treated as synonymous to the undercloud; it bootstraps the undercloud OpenStack deployment and provides the necessary tooling to deploy an overcloud.

In this solution, the Nova-Compute, Open vSwitch agents and SR-IOV agents run on the compute nodes. Compute nodes have either OVS-DPDK or SR-IOV enabled to better satisfy the performance requirement upon the data path from VNFs. The agents receive instrumentation requests from the controller node via RabbitMQ messages to manage the compute and network virtualization of instances that are running on the compute nodes.

The compute nodes can be aggregated into pools of various sizes for better management, performance, or isolation. A Red Hat Ceph Storage cluster is created on the storage nodes. It is largely self-managed and supervised by the Ceph monitor installed on the controller node. The Red Hat Ceph Storage cluster provides block data storage for Glance image store and for VM instances via the Cinder service and Nova services as well as for Telemetry service.

6.2.1 Utility node

The utility node is responsible for the initial deployment of the controller nodes, compute nodes, and storage nodes by leveraging the OpenStack bare metal provisioning service, it is also capable of running the Lenovo system management tools like Lenovo XClarity Administrator.

The Lenovo NFVI solution for service providers uses Red Hat OpenStack Platform Director as the toolset for installing and managing a production OpenStack environment, e.g. overcloud. The Red Hat OpenStack Platform Director is based primarily on the OpenStack TripleO project and uses a minimal OpenStack installation to deploy an operational OpenStack environment, including controller nodes, compute nodes, and storage nodes as shown in the diagrams. Director can be installed directly on the bare metal server or can run as a guest VM on the utility node. The Ironic component enables bare metal server deployment and management. This tool simplifies the process of installing and configuring the Red Hat OpenStack Platform while providing a means to scale in the future.

The Lenovo NFVI solution also integrates hardware management and cloud management to enable users to manage the physical and virtual infrastructure efficiently. The additional services can run as guest VMs together with underCloud director on Utility Node in a small-scale deployment. They can be deployed on additional utility nodes in a large-scale cloud environment.

The recommended minimum configuration for the utility node is:

- 2 x Intel Xeon Gold 5118 12C 105W 2.3GHz Scalable Processor
- 192 GB of system memory
- 1 x ThinkSystem RAID 930-8i 2GB Flash PCIe 12Gb Adapter
- 8 x 2.5" 1.8TB 10K SAS HDD (in RAID 10 array for local storage)

6.2.2 Compute nodes

The compute pool consists of multiple compute nodes virtualized by OpenStack Nova to provide a scalable cloud-computing environment. Red Hat OpenStack Platform provides Nova drivers to enable virtualization on Lenovo ThinkSystem servers. Network addresses are automatically assigned to VM instances. With the composable role feature available on director, user can create custom deployment roles for heterogeneous hardware configuration or network topology. For example, user can create separate roles for OVS-DPDK compute node and SR-IOV compute node, and configure them with separate parameters and network mappings.

Compute nodes inside the compute pool can be grouped into one or more “host aggregates” according to business need. For example, hosts may be grouped based on hardware features, capabilities or performance characteristics (e.g. NICs and OVS-DPDK configuration for acceleration of data networks used by VNFs).

The recommended configuration for compute nodes is:

- 2 x Intel Xeon Gold 6152 22C 140W 2.1GHz Scalable Processor
- 384 GB of system memory
- 1 x ThinkSystem RAID 930-8i 2GB Flash PCIe 12Gb Adapter
- 2 x ThinkSystem S4610 480GB Mainstream SSDs (in RAID 1 for Operating System)
- 2 x ThinkSystem U.2 Intel P4510 2.0TB NVMe (for optional local storage)

6.2.3 Controller nodes

The controller nodes act as a central entrance for processing all internal and external cloud operation requests. The controller nodes manage the lifecycle of all VM instances that are running on the compute nodes and provide essential services, such as authentication and networking to the VM instances. The controller nodes rely on support services, such as DHCP, DNS, and NTP. Typically, the controller nodes are implemented as a cluster of three nodes to support High Availability.

The controller cluster also hosts proxy and message broker services for scheduling compute and storage pool resources and provides the data store for cloud environment settings. In addition, the controller cluster also provides virtual routers and some other networking functions for all the VMs.

The recommended configuration for controller nodes is:

- 2 x Intel Xeon Gold 6126 12C 125W 2.6GHz Scalable Processor
- 192 GB of system memory
- 1 x ThinkSystem RAID 930-16i 4GB Flash PCIe 12Gb Adapter
- 2 x ThinkSystem S4610 480GB Mainstream SSDs (in RAID 1 for Operating System)
- 8 x ThinkSystem 2.5" 1.8TB 10K SAS (for optional local storage)

6.3 Storage

The OpenStack cluster provides two types of storage technology. The first is local storage in each server and the second uses Red Hat Ceph Storage. Both have their own advantages and apply to different scenarios.

The local storage pool consists of local drives in a server. Therefore the configuration is easier and they provide high-speed data access. For workloads that have demanding requirement to storage performance, it is suitable to use local storage pools. However, this approach lacks high-availability across servers and affects the ability to migrate workloads, and is usually limited by the storage capacity of the local disks.

The Red Hat Ceph Storage pool consists of multiple storage nodes that provide persistent storage resources from their local drives. In this reference architecture, all cloud data are stored in a single Ceph cluster for simplicity and ease of management. Figure 17 shows the details of the integration between the Red Hat OpenStack Platform and Red Hat Ceph Storage.

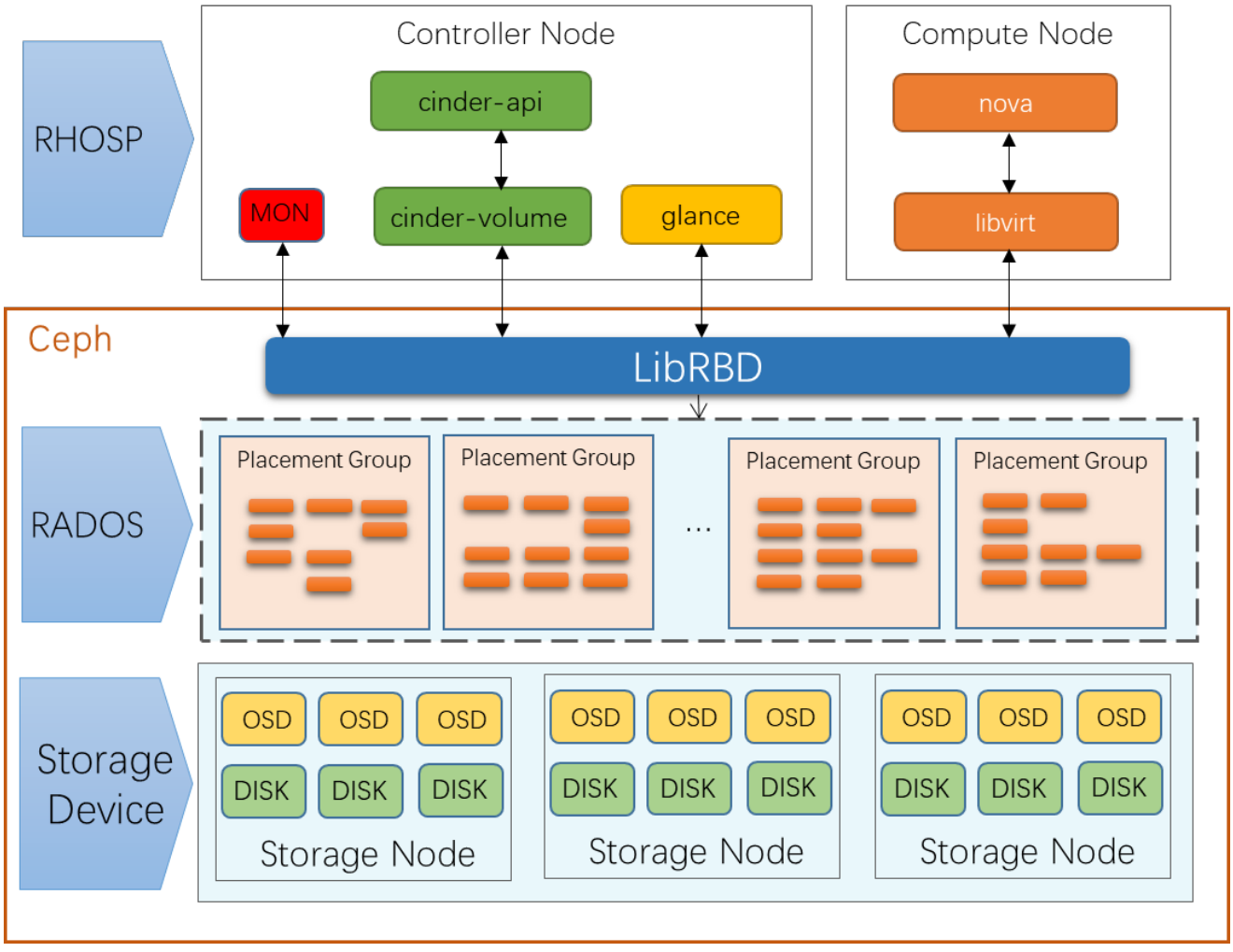


Figure 17. Red Hat Ceph Storage and OpenStack Services Integration

Ceph uses a write-ahead mode for local operations; a write operation hits the file system journal first and from there copied to the backing file store. To achieve optimal performance, SSDs are recommended for the operating system and the Ceph journal data. Please refer to Red Hat Ceph Storage Configuration Guide for Ceph OSD and SSD journal configuration details.

The recommended configuration for each Ceph storage node is:

- 2 x Intel Xeon Silver 4116 12C 85W 2.1GHz Scalable Processor
- 192 GB of system memory
- 1 x ThinkSystem RAID 930-16i 4GB Flash PCIe 12Gb Adapter
- 2 x ThinkSystem S4610 480GB Mainstream SSDs (in RAID 1 for Operating System)
- 12 x ThinkSystem 2.5” 1.8TB 10K SAS
- 2 x ThinkSystem 2.5" SS530 400GB Performance SAS SSD (for Ceph Journal)

6.4 Networking

In a typical NFV environment, network traffic needs to be isolated to fulfil the security and quality of service (QoS) requirements. To fulfil the OpenStack network requirement, a set of logical networks is created accordingly, some meant for internal traffic, such as the storage network or OpenStack internal API traffic, and others for external and tenant traffic. VLAN technology provides a simple way for logical grouping and isolation of the various networks.

Table 7 lists the Lenovo recommended VLANs.

Table 7. VLANs

Network	Description
Provisioning	Provides DHCP and PXE boot functions to help discover bare metal systems for use in the OpenStack installer to provision the system.
Tenant	This is the subnet for allocating the VM private IP addresses. Through this network, the VM instances can talk to each other
External	This is the subnet for allocating the VM floating IP addresses. It is the only network where external users can access their VM instances.
Storage	The front-side storage network where Ceph clients (through Glance API, Cinder API, or Ceph CLI) access the Ceph cluster. Ceph Monitors operate on this network.
Storage Management	The backside storage network to which Ceph routes its heartbeat, object replication, and recovery traffic.
Internal API	The Internal API network is used for communication between the OpenStack services using API communication, RPC messages, and database communication.
NFV data	NFV Data Network is used for NFV data plane traffic for VM applications that require high performance.

Lenovo recommends using one 1GbE switch, e.g. Lenovo RackSwitch G8052, to provide networking for the server BMC (Baseboard Management Controller) through the dedicated 1GbE port on each server and for the cloud provisioning through on-board 1GbE ports.

For the logical networks on OpenStack controller nodes, compute nodes, storage nodes, and utility server, it is recommended to use two 10Gb or 25Gb switches, e.g. NE2572, in pairs to provide redundant networking. The Virtual Link Aggregation Group (vLAG) feature on NE2572 switches is enabled to allow a pair of switches to act as a single endpoint for aggregation. It provides improved high availability compared to a single switch, enhances performance and increases the available bandwidth by splitting loads across the aggregated links and switches.

Red Hat OpenStack Platform provides IPv6 support to deploy and configure the overcloud. IPv6-native VLANs is supported as well for network isolation. For details, refer to IPv6 Networking for the Overcloud.

6.5 Systems Management

This section describes software for systems management of the solution.

6.5.1 Lenovo XClarity Administrator

Lenovo XClarity Administrator is a centralized resource management software that is aimed at reducing complexity, speeding up response, and enhancing the availability of Lenovo server systems and solutions. Lenovo XClarity Administrator provides agent-free hardware management for servers, storage and network switches.

Figure 18 shows the Lenovo XClarity Administrator interface, where Lenovo ThinkSystem servers are managed from the dashboard.

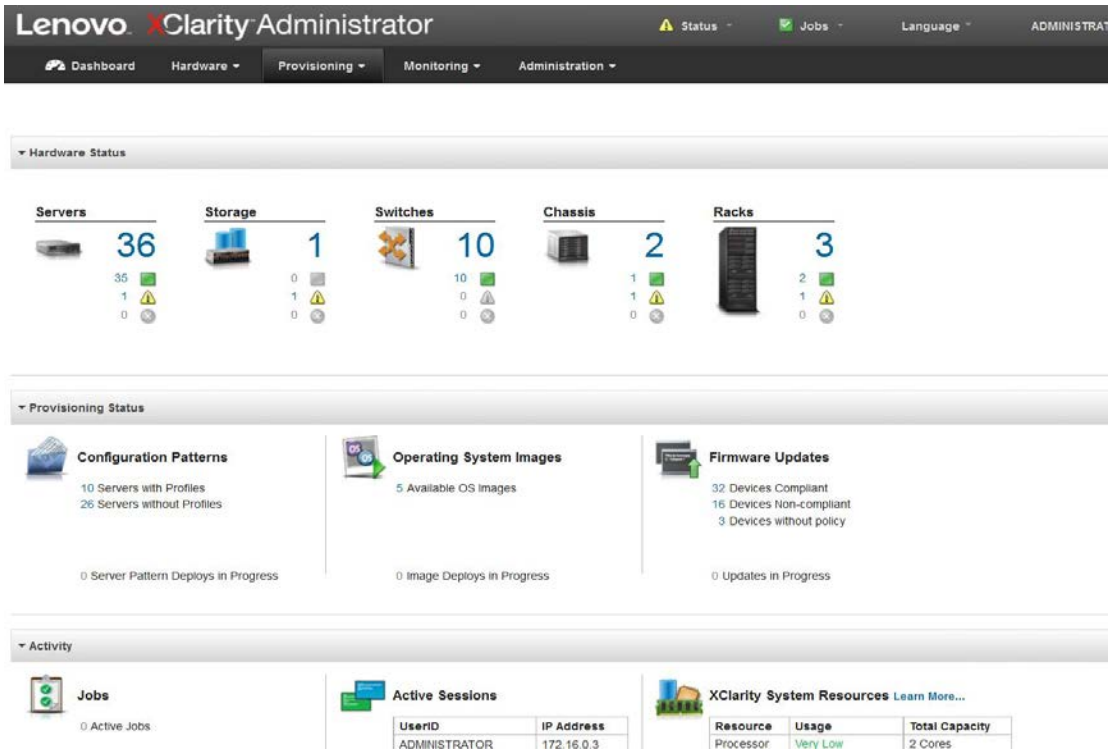


Figure 18. Lenovo XClarity Administrator Interface

XClarity provides the following features:

- Intuitive, Graphical User Interface
- Auto-Discovery and Inventory

- Firmware Updates and Compliance
- Configuration Patterns
- Bare Metal Deployment
- Security Management
- Upward Integration
- REST API , PowerShell, Python Scripts and Ansible module
- SNMP, SYSLOG and Email Forwarding

For more information, please refer to Lenovo XClarity.

6.5.2 Red Hat CloudForms

The Red Hat CloudForms Management Engine is another utility adopted in Lenovo NFVI solution to deliver insight, control and automation that enterprises need to address the challenges of managing physical and virtualized environments. This technology enables enterprises with existing IT infrastructures to improve visibility and control, and those starting virtualization deployments to build and operate a well-managed physical or virtualized infrastructure. Red Hat CloudForms has the following capabilities:

- Accelerate service delivery and reduce operational costs
- Improve operational visibility and control
- Ensure compliance and governance

The Lenovo Physical Infrastructure Provider for CloudForms 4.7 provides IT administrators the ability to integrate the management features of Lenovo XClarity Administrator with the hybrid-cloud management capabilities of Red Hat CloudForms. Lenovo expands physical-infrastructure management for on-premise cloud configurations extending the visibility of connected physical and virtual infrastructures. This facilitates the configuration, monitoring, event management, and power monitoring needed to reduce cost and complexity through server consolidation and simplified management across physical and virtual component boundaries.

Figure 19 shows the architecture and capabilities of Red Hat CloudForms. These features are designed to work together to provide robust management and maintenance of your virtual infrastructure.

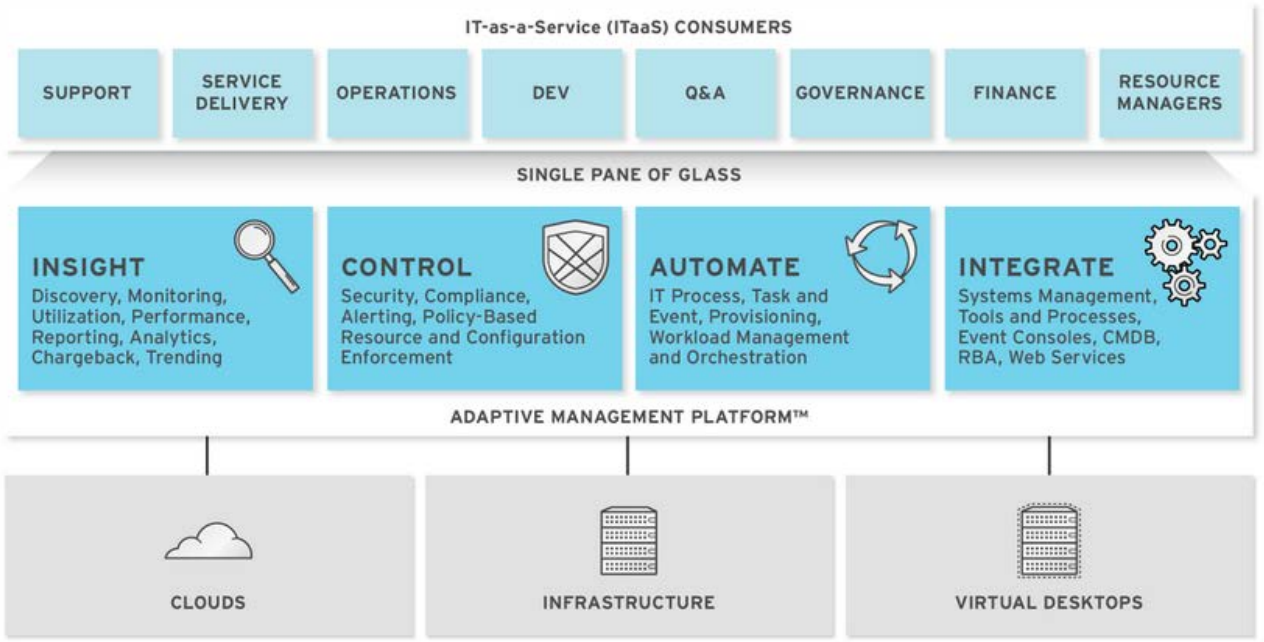


Figure 19. Red Hat CloudForms Architecture

Figure 20 demonstrates a physical infrastructure topology generated by Red Hat CloudForms and integrated with the Lenovo XClarity provider.

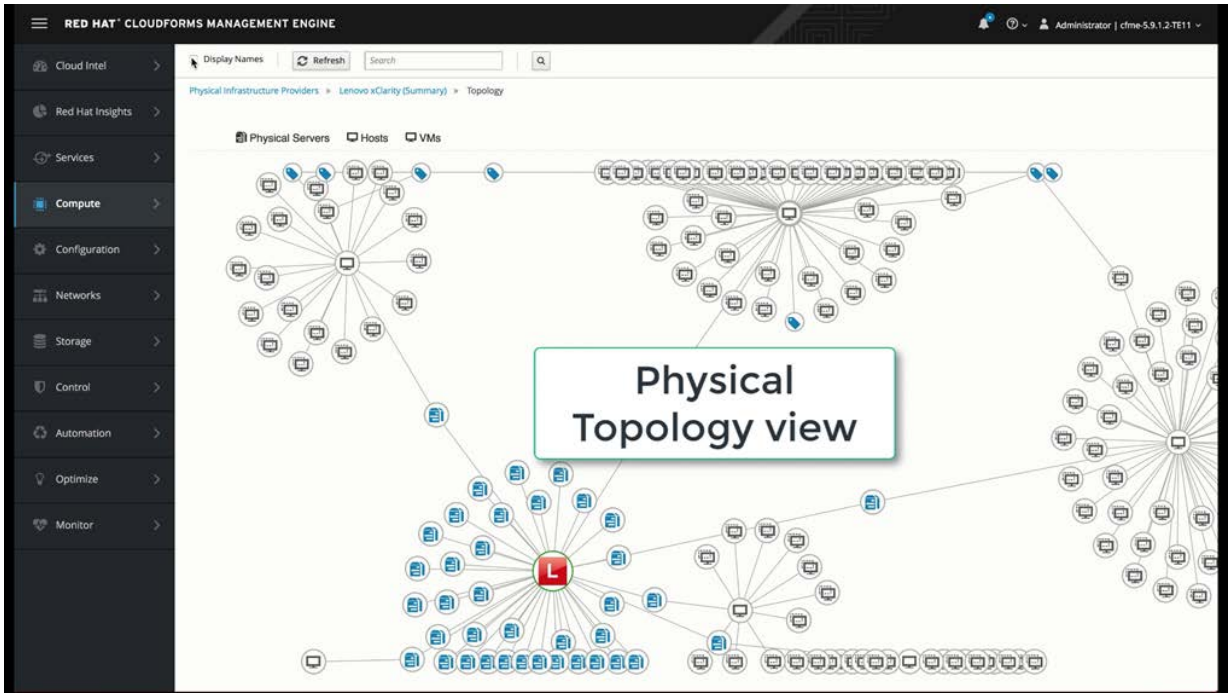


Figure 20 Topology View of Lenovo XClarity Providers Integrated in Red Hat CloudForms

To install Red Hat CloudForms 4.7 on Red Hat OpenStack platform, please refer to Installing CloudForms.

7 NFVI & VNF characterization and benchmarking

It has been a challenge in telco industry to define common standards and industry-accepted benchmarks for conformance to carrier-grade requirements. However, there has been recent advances in benchmarking methodologies from international standard organizations such as IETF, ETSI NFV, and Open Platform for NFV (OPNFV) community as well as from testing agencies such as EANTC. This chapter covers Lenovo's testing framework and practices for validating NFVI software and hardware.

7.1 Introduction to Network Services Benchmarking

Network Services Benchmarking (NSB) is part of OPNFV Yardstick project which provides a test framework for benchmarking virtual network functions (VNFs) in an NFVI environment. The general use case for NSB is to onboard a VNF, configure the source and sink traffic for the VNF and a traffic generator, and then measure the performance of the VNF with various provided KPI metrics. The NSB toolset provides industry-accepted benchmark KPIs for conformance to carrier-grade requirements.

7.2 Configuring the NSB testbed

7.2.1 Testing components and network topology

Figure 21 shows the networking topology of the NSB testbed for Lenovo NFVI solution.

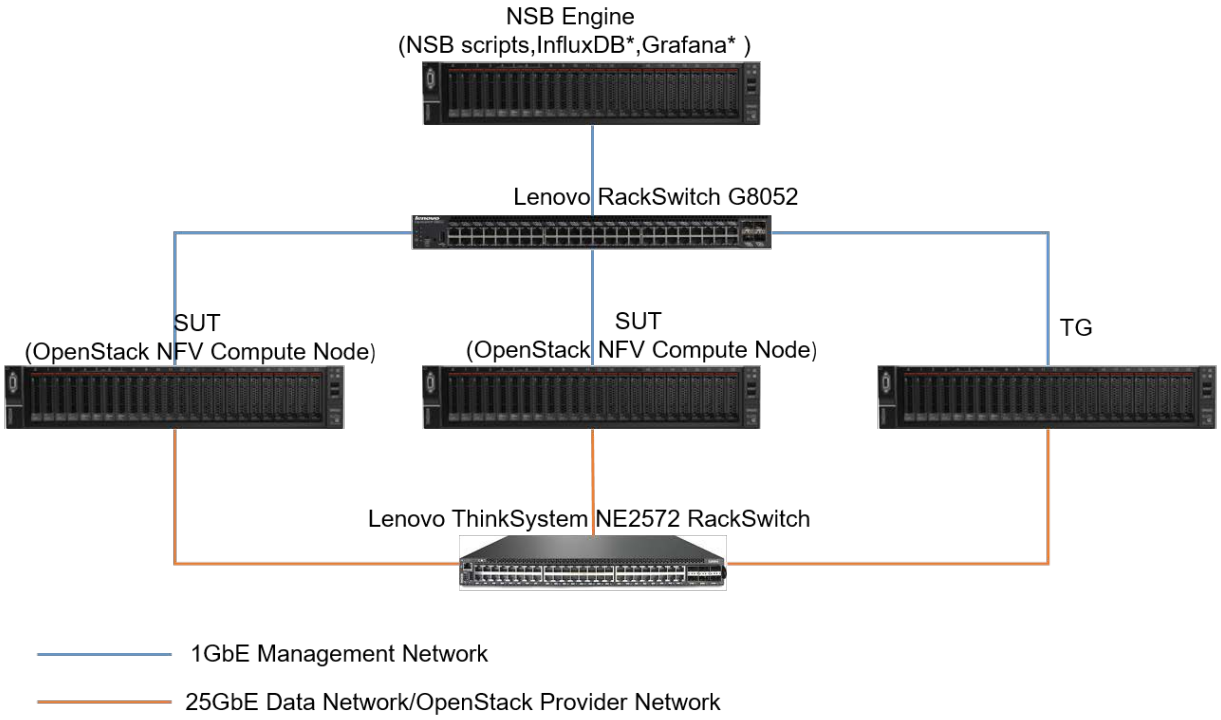


Figure 21. Network Topology of NSB Testbed

For performance benchmarking, it is critical the separate the VNF workload data traffic with all the other traffic including the management traffic, so that the performance KPIs are not polluted by other traffic sources. The Lenovo RackSwitch G8052 is used to aggregate the control and management traffic, and Lenovo

ThinkSystem NE2572 RackSwitch is used to carry the VNF workload data traffic at a bandwidth of up to 25Gbps per port.

The NSB Engine runs in a virtual machine which includes the orchestration, management and visualization components of NSB suite. The user can orchestrate a VNF workload in the targeted cloud environment using the management network. Once the selected benchmark tests are completed, the NSB engine automatically collects all the KPIs, saves the benchmark data in a database, and displays the results in a dashboard for review.

The System Under Test (SUT) is the NFVI compute node running VNFs which are orchestrated by NSB engine. The NSB engine generates the VNF configuration according to the test plan. Typical configuration parameters include NIC ports, number of RX/TX queues, cores for PMDs, etc. Due to the existing limitation of the NSB test suite, the bare-metal compute node and Red Hat OpenStack environment need to be configured properly to provide the matching configuration on NFVI. The detailed configurations can be found in sections 7.2.5 and 7.2.6.

The Traffic Generator (TG) is a bare-metal compute node running Open Source traffic generator software such as Pktgen or Trex to generate VNF data traffic at line rate and display real time metrics on the NIC ports. Similar to the SUT, TG can be configured at run-time by the NSB engine. The traffic generator compute node should use the same BIOS settings as the SUT.

7.2.2 Test setup

One of the main goals of the performance benchmarking is to provide a reproducible configuration and performance KPIs on the tested hardware and software environment. The performance KPIs and benchmarking results will provide a quantitative reference for the users who want to verify a setup in the production or for proof of concepts before making procurement decisions.

Lenovo choose the Open Source application L3fwd as the sample VNF. L3fwd is a simple application performing layer-3 packet forwarding using the DPDK. It is widely used and well accepted as a useful VNF for benchmarking. For more information, see dpdk.org/doc/guides/sample_app_ug/l3_forward.html.

The L3fwd application is executed in following three environments to provide a comprehensive comparison of typical NFVI deployments:

- Bare-metal
- OVS-DPDK compute
- SR-IOV compute

The best performance is expected when executing VNFs on bare-metal compute nodes because there is the least overhead in data traffic processing. Performance benchmarks obtained from bare-metal setups are often used to evaluate new hardware configurations and new software stacks. Executing VNFs in a SR-IOV environment usually provides sub-optimal throughput compared to the bare-metal setup, but is still able to achieve much higher throughput than OVS-DPDK because there is no hypervisor overhead.

Figure 22 shows the three test environments and the respective traffic flows.

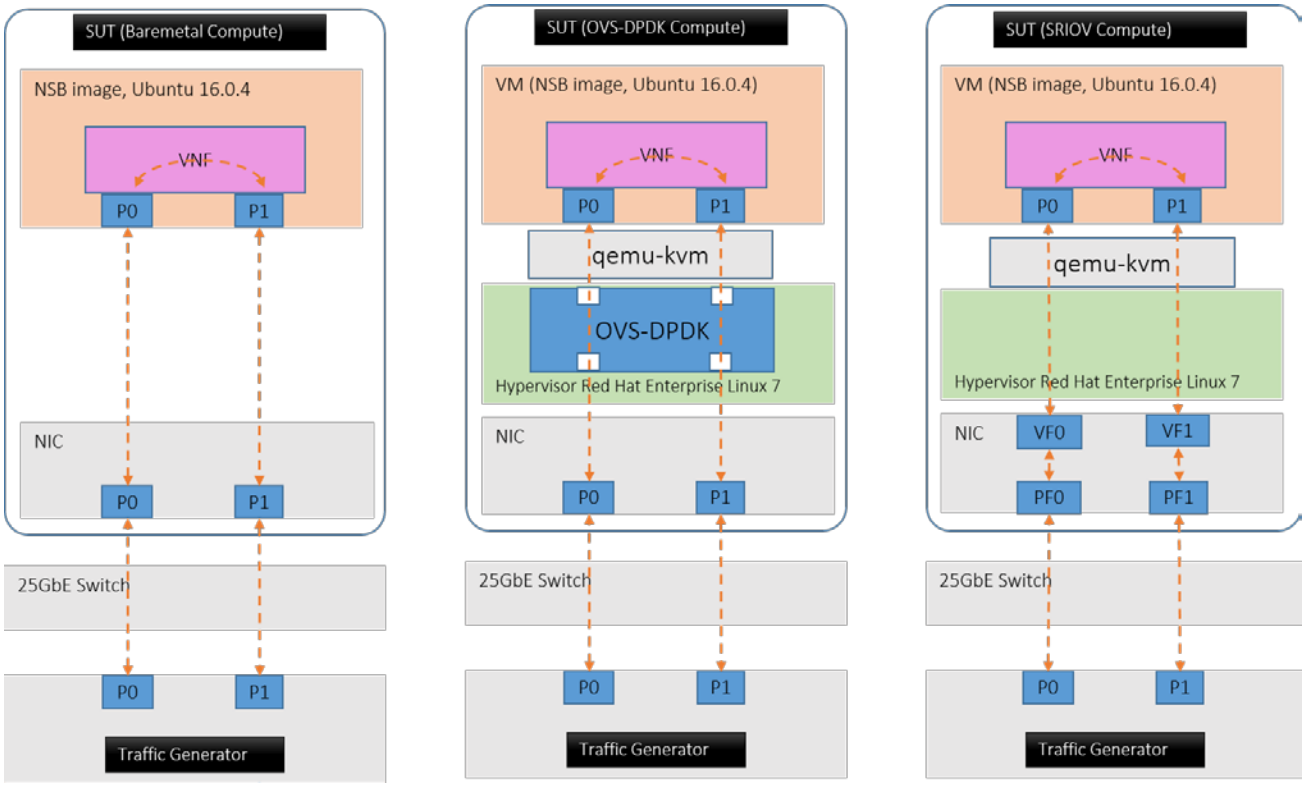


Figure 22. Bare metal, OVS-DPDK and SR-IOV NSB Test Configuration

NIC ports in the SUT are connected to the traffic generator ports through a Lenovo ThinkSystem NE2572 RackSwitch. Bi-directional traffic is sent from traffic generator and the aggregated throughputs at the receiving side of traffic generator are calculated to give the overall throughput for different packet sizes from 64 bytes to 1518 bytes.

The benchmarking methodology documented in the RFC2544 standard was adopted in the testing setup and benchmark data collection. In particular, the test cases are designed to check the maximum IO throughput for a single core. Two Intel XXV710 NIC cards are attached to the first processor of the compute node. Only the first port is used on each NIC for transporting the VNF traffic. Each port has one queue assigned and all the TX/RX queues are assigned to the same logical core.

7.2.3 Hardware components

Table 8 lists the hardware configuration of the ThinkSystem SR650 SUT.

Table 8. Hardware Configuration

Hardware	Description
Processor	2 x Intel(R) Xeon(R) Gold 6138T CPU @ 2.00GHz
Memory	12 x ThinkSystem 32GB TruDDR4 2666 MHz (2Rx4 1.2V) RDIMM
Storage	1 x ThinkSystem 480GB 6Gbps SATA 2.5" SSD
NICs	1 x Intel X722 LOM 1 x ThinkSystem Intel XXV710-DA2 PCIe 25Gb 2-Port 1 x ThinkSystem Intel X710-DA2 PCIe 10Gb 2-port
BMC	Lenovo XClarity Controller (XCC) Version 2.10
UEFI	Version 1.41 Build IVE126O (Build CDI328M)

7.2.4 Software components

Table 9 provides the detailed version for each software components of the testbed.

Table 9. Software Configuration

Software	Version
Red Hat OpenStack Platform	Red Hat OpenStack Platform 13
Host Operating System	Red Hat Enterprise Linux Server release 7.6 (Maipo) Kernel version: 3.10.0-957
VM Operating System	
KVM	qemu-kvm 2.12.0
Open vSwitch	Open vSwitch 2.9.0
DPDK	DPDK 17.11-13
NSB Test Suite	stable-gambia
DPDK for VNFs	17.05
NIC Driver	I40E 2.3.2-k
NIC Firmware	6.01 0x8000385c 1.1892.0

7.2.5 BIOS settings

Table 10 lists the key BIOS (UEFI) settings for the NFV compute nodes in order to achieve optimal performance.

Table 10. BIOS Configuration

Menu	Setting	Value
System Settings->Operating Modes	Choose Operation Mode	Maximum Performance
System Settings->Processors	Turbo Mode	Enable
System Settings->Processors	CPU P-state Control	None
System Settings->Processors	C-States	Disable
System Settings->Processors	C1 Enhanced Mode	Disable
System Settings->Processors	DCA	Enable
System Settings->Power	Power/Performance Bias	Platform Controlled
System Settings->Power	Platform Controlled Type	Maximum Performance
System Settings->Memory	Mirror Mode	Disable
System Settings->Memory	Memory Speed	Max Performance
System Settings->Memory	Memory Power Management	Disable
System Settings->Memory	Socket Interleave	NUMA
System Settings->Memory	Patrol Scrub	Disable

7.2.6 OpenStack settings

In order to achieve optimal performance for the OVS-DPDK configuration, the sibling logical CPU cores are not used for PMD cores . This ensures that the PMD core mapping is from only one logical CPU to each physical CPU core.

Below are the SR-IOV specific parameters and OVS-DPDK specific parameters for Red Hat OpenStack deployment.

```
 ComputeSriovParameters:
   NovaSchedulerDefaultFilters:
"RamFilter,ComputeFilter,ServerGroupAffinityFilter,ServerGroupAntiAffinityFilter,Availa
bilityZoneFilter,ComputeCapabilitiesFilter,ImagePropertiesFilter,PciPassthroughFilter,N
UMATopologyFilter"
   KernelArgs: "isolcpus=4-19,24-39,44-59,64-79 nohz_full=4-19,24-39,44-59,64-79
default_hugepagesz=1GB hugepagesz=1G hugepages=192 iommu=pt intel_iommu=on"
   SriovNeutronNetworkType: 'flat'
   NovaVcpuPinSet: "4-15,24-35,44-55,64-75"
   IsolCpusList: "4-19,24-39,44-59,64-79"
   NeutronSriovNumVFs:
      - enp47s0f1:16:switchdev
      - enp47s0f0:16:switchdev
   NeutronPhysicalDevMappings: "prov704:enp47s0f1,prov703:enp47s0f0"
   NovaPCIPassthrough:
```

```
      - devname: "enp47s0f1"
        physical_network: "prov704"
      - devname: "enp47s0f0"
        physical_network: "prov703"
    TunedProfileName: "cpu-partitioning"
    NeutronSupportedPCIVendorDevs: ['8086:158b']

  ComputeOvsDpdkParameters:
    NovaVcpuPinSet: "4-15,24-35,44-55,64-75"
    NovaSchedulerDefaultFilters:
"RamFilter,ComputeFilter,ServerGroupAffinityFilter,ServerGroupAntiAffinityFilter,Availa
bilityZoneFilter,ComputeCapabilitiesFilter,ImagePropertiesFilter,PciPassthroughFilter,N
UMATopologyFilter"
    KernelArgs: "isolcpus=4-19,24-39,44-59,64-79 nohz_full=4-19,24-39,44-59,64-79
default_hugepagesz=1GB hugepagesz=1G hugepages=192 iommu=pt intel_iommu=on"
    IsolCpusList: "4-19,24-39,44-59,64-79"
    OvsEnableDpdk: True
    TunedProfileName: "cpu-partitioning"
    NovaReservedHostMemory: 4096
    OvsDpdkSocketMemory: "2048,2048"
    OvsDpdkMemoryChannels: "4"
    OvsPmdCoreList: "16,17,18,19, 56,57,58,59"
    VhostuserSocketGroup: "hugetlbfs"
```

7.3 Performance results

Figure 23 and Figure 24 show the throughput and line rate benchmarking results on bare metal, SR-IOV and OVS-DPDK configurations. Full line rate (e.g. 25Gbps) can be achieved by all the three configurations with large packet sizes. The bare-metal configuration shows the best performance for small packet sizes. Throughput for the SR-IOV configuration is close to bare-metal and is significantly better than OIVS-DPDK.

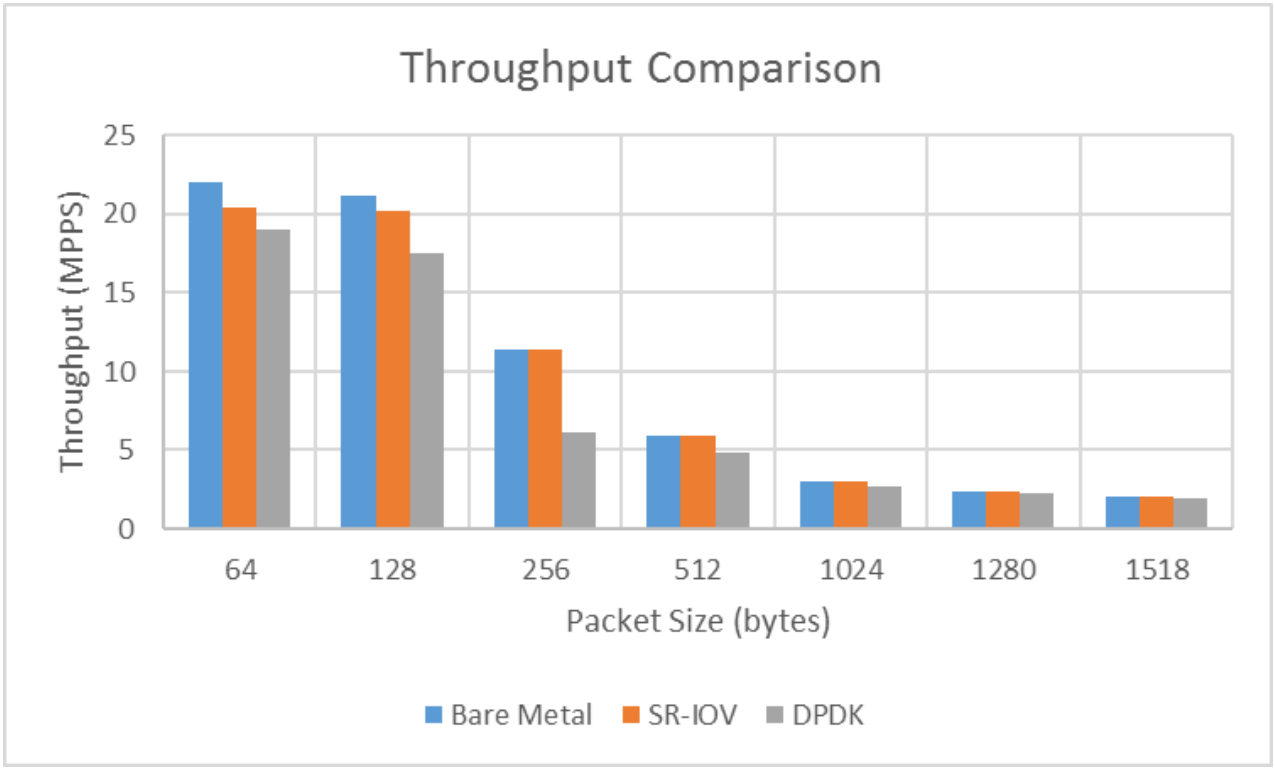


Figure 23. Comparison of Packet Throughput over Bear Metal, SR-IOV and OVS-DPDK

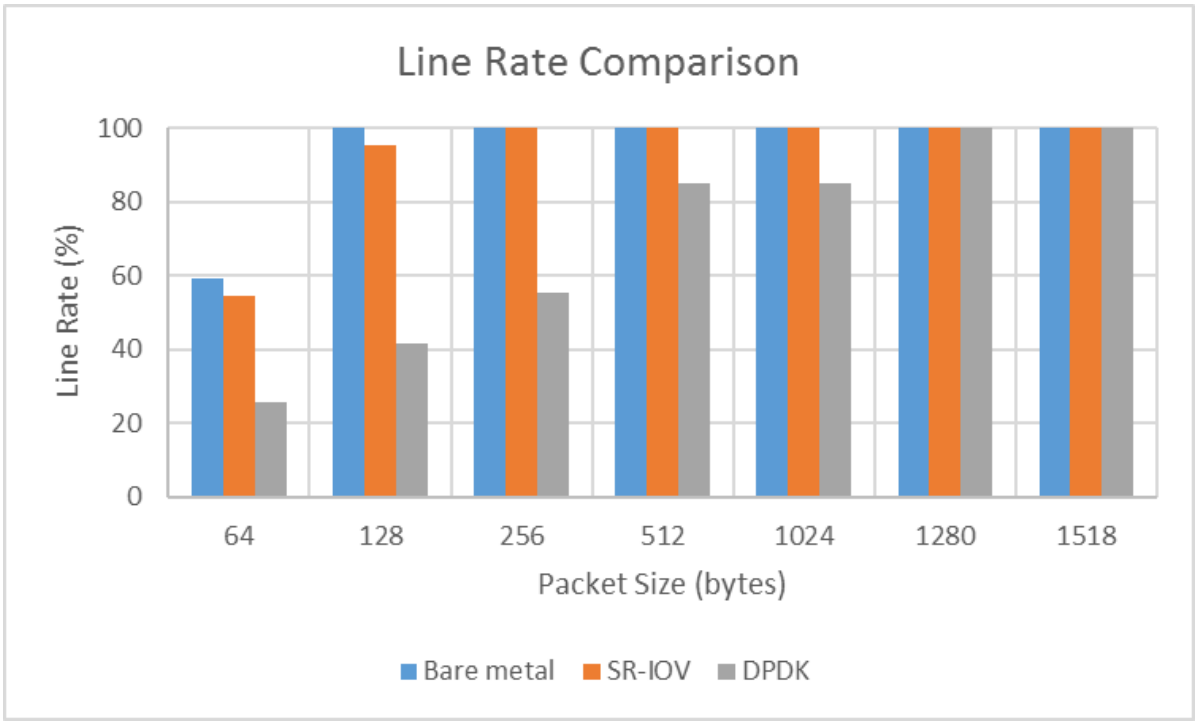


Figure 24. Comparison of Line Rate over Bare Metal, SR-IOV, and OVS-DPDK

8 Deployment example

This section describes an example deployment of Lenovo OpenStack solution for service provider using Red Hat OpenStack Platform 13 with Lenovo hardware.

8.1 Hardware configuration

The Figure 25 shows an example of rack configuration with mix of controllers, compute nodes, storage, and networking equipment.

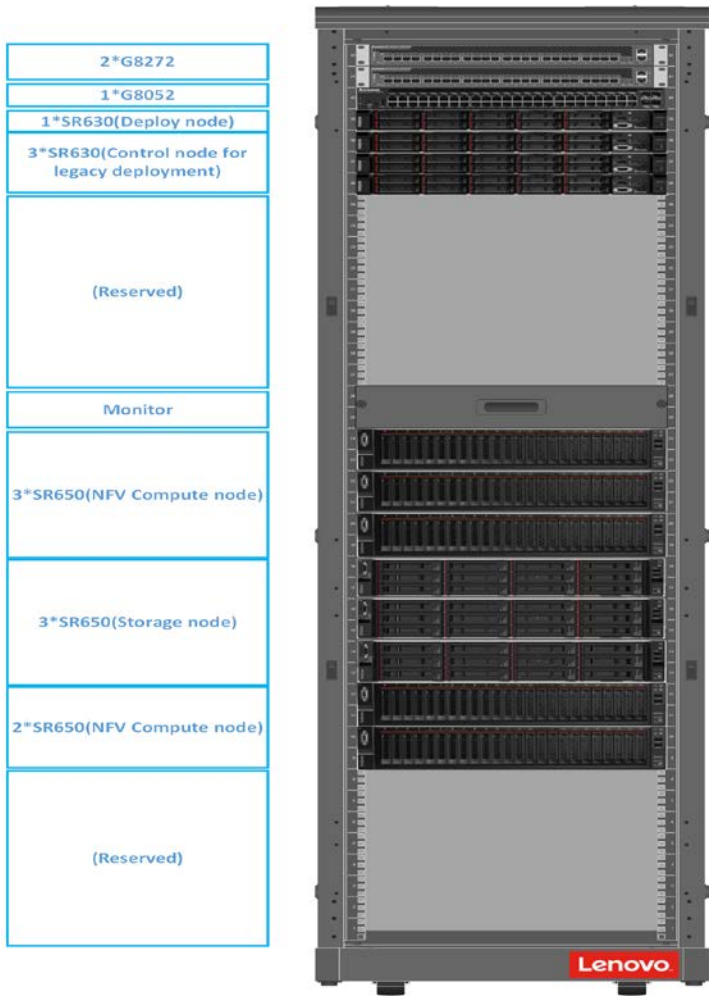
Components	Capacity	Rack layout
Rack	1 (42U)	
Ethernet switches	2 (NE2572), 1 (G8052)	
Utility node	1 (ThinkSystem SR630)	
Controller nodes	3 (ThinkSystem SR630)	
Monitor	1	
Compute nodes (SR-IOV)	3 (ThinkSystem SR650)	
Storage Nodes	3 (ThinkSystem SR650)	
Compute nodes (DPDK)	2 (ThinkSystem SR650)	

Figure 25. Deployment Example 1: Full Rack System Deployment

The rack shown in Figure 25 provides a reference configuration consisting of two NE2572 and one G8052 switches, one utility node, three controller nodes, three storage nodes and five NFV compute nodes. Additional compute nodes and storage nodes can be easily added to expand the available capacity. Two flavors of NFV compute configurations, e.g. SR-IOV and OVS-DPDK are verified in this setup.

8.2 Networking isolation

Table 11 lists the seven logical networks used in this deployment example with a mapping to the physical NICs on servers.

Table 11. OpenStack Logical Networks of Lenovo NFVI solution

Network	Speed	VLAN	Interfaces On Compute	Interfaces On Controller	Interfaces On Storage
Provisioning	1GbE	2	NIC 1	NIC 1	N/A
Tenant	25GbE	3	NIC 2/3 Bond	NIC 2/3 Bond	N/A
External	25GbE	4	N/A	NIC 2/3 Bond	N/A
Storage	25GbE	5	NIC 2/3 Bond	NIC 2/3 Bond	NIC 2/3 Bond
Storage Management	25GbE	6	N/A	NIC 2/3 Bond	NIC 4/5 Bond
Internal API	25GbE	7	NIC 2/3 Bond	NIC 2/3 Bond	N/A
NFV data	25GbE	8-10	NIC 4/5 bond on OVS-DPDK Compute Nodes; NIC 4/5 no bond on SR-IOV Compute Nodes	N/A	N/A

Figure 26 shows the recommended network topology diagram with VLANs for the Lenovo NFVI solution.

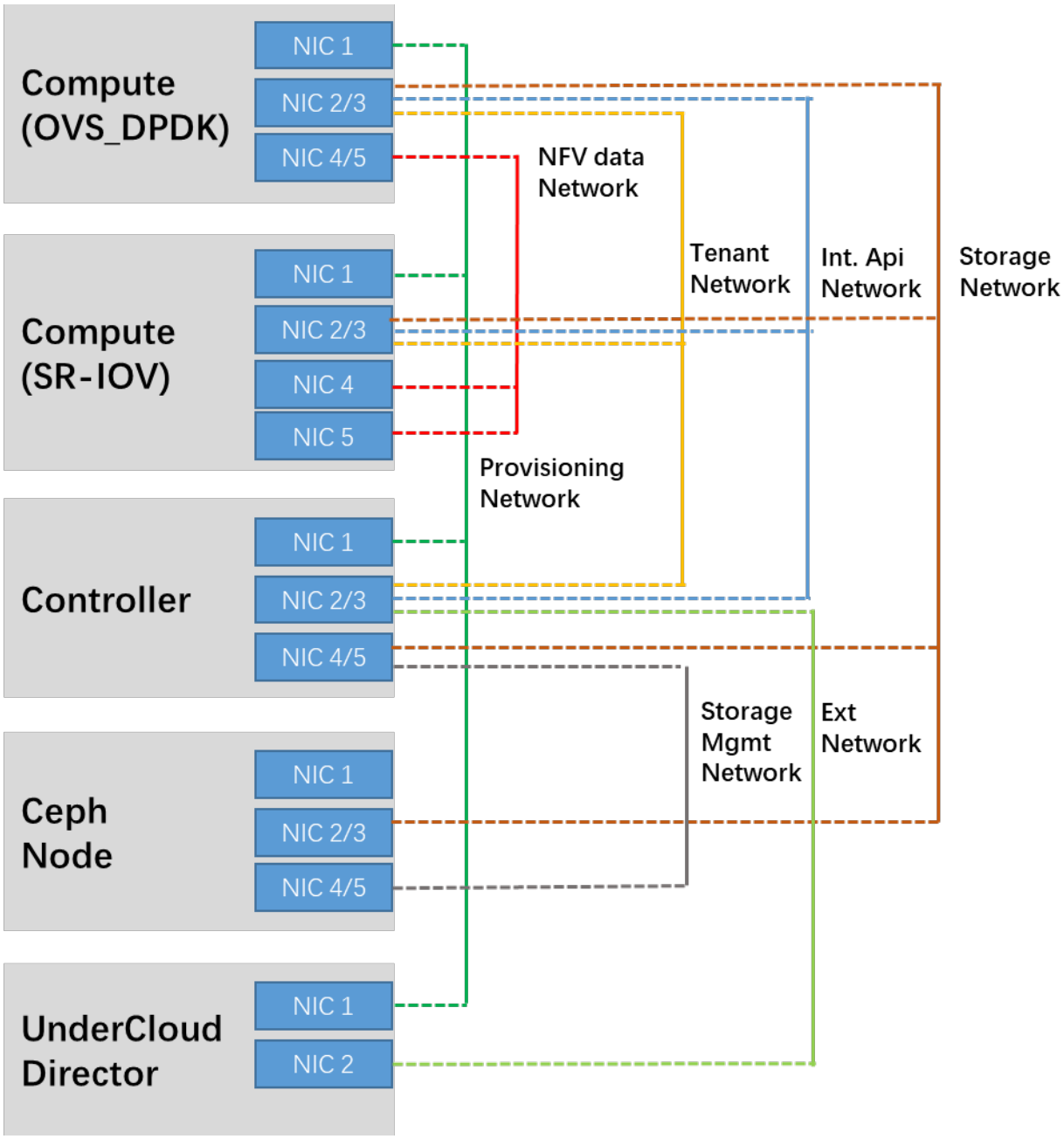


Figure 26. Network Connectivity Overview in the Lenovo NFVI Solution

As shown in the above network connectivity diagram, the storage traffic on Compute Node shares the bonded interface with the Tenant, External, and Internal API networks, and separate provider network is allocated for NFV data. This assumes that typical VNFs have less I/O requirements for storage, but much higher demands on high-throughput and low latency for the data traffic. Two bonded 25GbE interfaces are dedicated to a provider DPDK network on OVS-DPDK compute nodes. For SR-IOV compute nodes, as the physical NIC resources are passed through to VMs via virtual functions, NIC bond is not applied. The number of VFs provided by each of the 25GbE interfaces can be specified during cloud deployment.

8.3 Cloud deployment for accelerated data networking

OVS-DPDK and/or SR-IOV can be enabled on compute nodes to provide VNFs with high performance networking. This section provides guidance to enable OVS-DPDK and/or SR-IOV on the chosen Lenovo ThinkSystem SR650 server with Intel XXV710 NICs. The Intel XXV710 Series Network Adapters support Network Virtualization offloads including VXLAN, NVGRE, MPLS and VXLAN-GPE with Network Service Headers on compute nodes. It also supports DPDK for optimized packet processing. Users can refer to full list of Red Hat supported NICs for DPDK. The Intel XXV710 network adapters also supports SR-IOV of up to 128 VFs for direct assignment in virtualized environment.

8.3.1 Pre-deployment configurations

Before starting the deployment of compute nodes with enhanced IO performance, the following prerequisites must be met by setting options in the server BIOS:

1. Enable “Hyper-Threading” in BIOS
2. Disable all power saving options in BIOS such as: “CPU P-state Control”, “CPU C-States”
3. Select Performance as the CPU Power and Performance policy. It is recommended to use the “Maximum Performance” Operating Mode and set “Platform Controlled Type” to “Maximum Performance” in BIOS Power setting.
4. Enable “DCA” (Direct Cache Access) in BIOS setting;

8.3.2 High-available data path

NIC bonding is a network technology that enables high throughput by aggregating multiple physical links into a single high-speed aggregated network interface. In addition to increasing bandwidth, network bonding provides fault tolerance. A NIC-bond is created over two NICs to ensure that network is up even when one NIC or a leaf switch goes down. Depending on the selected bonding mode, the switch configuration might need to be changed accordingly. For example, if “balance-tcp” is selected as OVS-DPDK bonding mode, switch ports need to be configured properly with vLAG enabled.

The implementation of NIC bonding on SR-IOV compute nodes is slightly different than on DPDK compute node. For the SR-IOV enabled compute nodes NIC ports are connected to switch port without bonding. While NIC bonding is enabled on VNFs which have multiple VF connections. On DPDK compute nodes, NIC bonding is done at compute hosts.

Figure 27 shows the different HA configurations for VNF data paths.

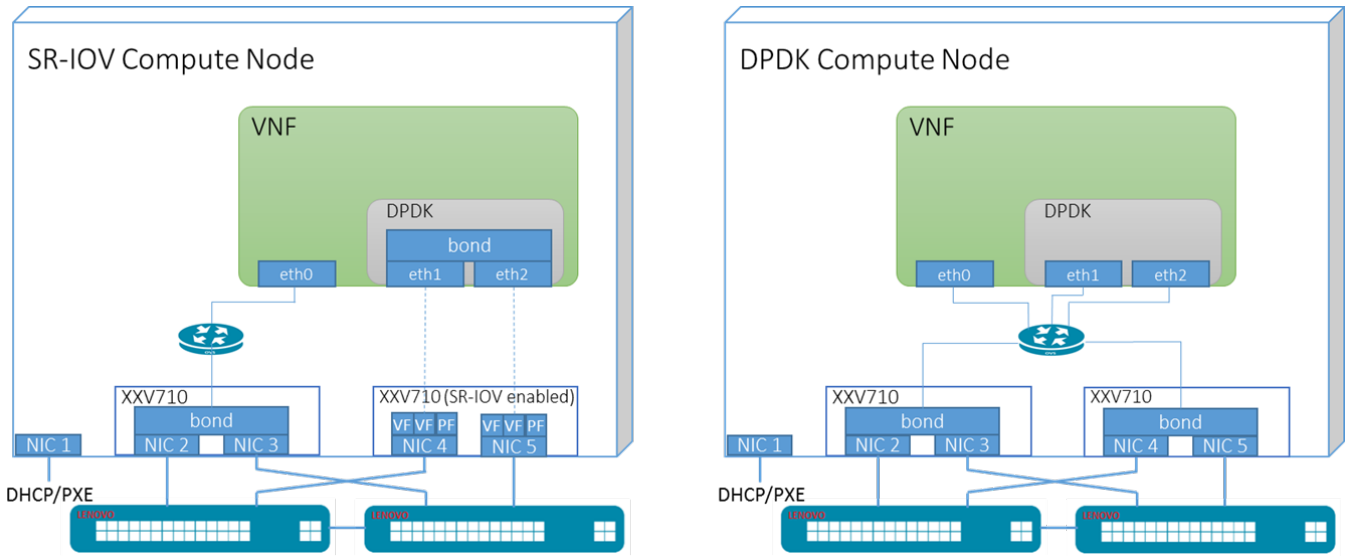


Figure 27. Implementation of NFV Data Path High Availability

8.3.3 Cloud configurations

To enable NFV deployment in Red Hat OpenStack Platform, Cloud Admin needs to make the following changes in Red Hat OpenStack Platform overcloud deployment templates:

1. Update roles_data.yaml to include ComputeOvsDpdk and ComputeSriov roles;
2. Update network-environment.yaml file with network, DPDK and SR-IOV parameters;
3. Update compute-ovsdpdk.yaml or compute-sriov file to match the DPDK and SR-IOV NICs configuration accordingly;
4. Update puppet-ceph-external.yaml to match the external ceph configuration

Updating roles_data.yaml

The basic method of adding services involves creating a copy of the default service list for a node role and then adding services. For example, Cloud Admin can add ComputeOvsDpdk and ComputeSriov role to the default roles_data.yaml file.

```
# Role: ComputeOvsDpdk                                                    #
###############################################################################
- name: ComputeOvsDpdk
description: |
Compute OvS DPDK Role
CountDefault: 1
networks:
- InternalApi
- Tenant
- Storage
HostnameFormatDefault: '%stackname%-computeovsdpdk-%index%'
disable_upgrade_deployment: True
deprecated_nic_config_name: 'compute-dpdk.yaml'
```

```
ServicesDefault:
- OS::TripleO::Services::Aide
- OS::TripleO::Services::AuditD
- OS::TripleO::Services::CACerts
- OS::TripleO::Services::CephClient
- OS::TripleO::Services::CephExternal
- OS::TripleO::Services::CertmongerUser
- OS::TripleO::Services::Collectd
- OS::TripleO::Services::ComputeCeilometerAgent
- OS::TripleO::Services::ComputeNeutronCorePlugin
- OS::TripleO::Services::ComputeNeutronL3Agent
- OS::TripleO::Services::ComputeNeutronMetadataAgent
- OS::TripleO::Services::ComputeNeutronOvsDpdk
- OS::TripleO::Services::Docker
- OS::TripleO::Services::Fluentd
- OS::TripleO::Services::Ipsec
- OS::TripleO::Services::Iscsid
- OS::TripleO::Services::Kernel
- OS::TripleO::Services::LoginDefs
- OS::TripleO::Services::MySQLClient
- OS::TripleO::Services::NeutronBgpVpnBagpipe
- OS::TripleO::Services::NovaCompute
- OS::TripleO::Services::NovaLibvirt
- OS::TripleO::Services::NovaMigrationTarget
- OS::TripleO::Services::Ntp
- OS::TripleO::Services::ContainersLogrotateCrond
- OS::TripleO::Services::OpenDaylightOvs
- OS::TripleO::Services::OVNMetadataAgent
- OS::TripleO::Services::Rhsm
- OS::TripleO::Services::RsyslogSidecar
- OS::TripleO::Services::Securetty
- OS::TripleO::Services::SensuClient
- OS::TripleO::Services::SkydiveAgent
- OS::TripleO::Services::Snmp
- OS::TripleO::Services::Sshd
- OS::TripleO::Services::Timezone
- OS::TripleO::Services::TripleoFirewall
- OS::TripleO::Services::TripleoPackages
- OS::TripleO::Services::Ptp
```

```
# Role: ComputeSriov                                                    #
################################################################################
- name: ComputeSriov
  description: |
    Compute SR-IOV Role
  CountDefault: 1
  networks:
    - InternalApi
    - Tenant
    - Storage
  HostnameFormatDefault: '%stackname%-computesriov-%index%'
  disable_upgrade_deployment: True
  ServicesDefault:
    - OS::TripleO::Services::Aide
```

```
    - OS::TripleO::Services::AuditD
    - OS::TripleO::Services::CACerts
    - OS::TripleO::Services::CephClient
    - OS::TripleO::Services::CephExternal
    - OS::TripleO::Services::CertmongerUser
    - OS::TripleO::Services::Collectd
    - OS::TripleO::Services::ComputeCeilometerAgent
    - OS::TripleO::Services::ComputeNeutronCorePlugin
    - OS::TripleO::Services::ComputeNeutronL3Agent
    - OS::TripleO::Services::ComputeNeutronMetadataAgent
    - OS::TripleO::Services::ComputeNeutronOvsAgent
    - OS::TripleO::Services::Docker
    - OS::TripleO::Services::Fluentd
    - OS::TripleO::Services::Ipsec
    - OS::TripleO::Services::Iscsid
    - OS::TripleO::Services::Kernel
    - OS::TripleO::Services::LoginDefs
    - OS::TripleO::Services::MySQLClient
    - OS::TripleO::Services::NeutronBgpVpnBagpipe
    - OS::TripleO::Services::NeutronSriovAgent
    - OS::TripleO::Services::NeutronSriovHostConfig
    - OS::TripleO::Services::NeutronVppAgent
    - OS::TripleO::Services::NovaCompute
    - OS::TripleO::Services::NovaLibvirt
    - OS::TripleO::Services::NovaMigrationTarget
    - OS::TripleO::Services::Ntp
    - OS::TripleO::Services::ContainersLogrotateCrond
    - OS::TripleO::Services::OpenDaylightOvs
    - OS::TripleO::Services::Rhsm
    - OS::TripleO::Services::RsyslogSidecar
    - OS::TripleO::Services::Securetty
    - OS::TripleO::Services::SensuClient
    - OS::TripleO::Services::SkydiveAgent
    - OS::TripleO::Services::Snmp
    - OS::TripleO::Services::Sshd
    - OS::TripleO::Services::Timezone
    - OS::TripleO::Services::TripleoFirewall
    - OS::TripleO::Services::TripleoPackages
    - OS::TripleO::Services::Vpp
    - OS::TripleO::Services::OVNController
    - OS::TripleO::Services::OVNMetadataAgent
    - OS::TripleO::Services::Ptp
```

Modification of network-environment.yaml

The network-environment.yaml file defines isolated network and related parameters. DPDK related parameters need to be set under parameter_defaults. The parameters need to be properly tuned on Lenovo hardware to achieve optimal performance. The following example of parameters apply only to the Intel Xeon Gold 6138T CPU @ 2.00GHz which is used in this reference architecture that has two NUMA nodes each with 40 logical CPU cores when hyperthreading is enabled.

```
Resource_registry:

  OS::TripleO::Compute::Net::SoftwareConfig:
        /home/stack/templates/compute.yaml
```

```
  OS::TripleO::Controller::Net::SoftwareConfig:
        /home/stack/templates/controller.yaml
  OS::TripleO::ComputeOvsDpdk::Net::SoftwareConfig:
        /home/stack/templates/compute-dpdk.yaml
  OS::TripleO::ComputeSriov::Net::SoftwareConfig:
        /home/stack/templates/compute-sriov.yaml
  OS::TripleO::NodeExtraConfigPost:
        /home/stack/templates/post-install.yaml
  OS::TripleO::Services::NeutronSriovAgent: /usr/share/openstack-tripleo-heat-
templates/docker/services/neutron-sriov-agent.yaml
  OS::TripleO::Services::NeutronSriovHostConfig: /usr/share/openstack-tripleo-heat-
templates/puppet/services/neutron-sriov-host-config.yaml
```

```
ComputeSriovParameters:

    NovaSchedulerDefaultFilters:
“RamFilter,ComputeFilter,ServerGroupAffinityFilter,ServerGroupAntiAffinityFilter,AvailabilityZoneFilt
er,ComputeCapabilitiesFilter,ImagePropertiesFilter,PciPassthroughFilter,NUMATopologyFilter”
    KernelArgs: “isolcpus=4-19,24-39,44-59,64-79 nohz_full=4-19,24-39,44-59,64-79
default_hugepagesz=1GB hugepagesz=1G hugepages=192 iommu=pt intel_iommu=on”
    SriovNeutronNetworkType: ‘flat’
    NovaVcpuPinSet: “4-15,24-35,44-55,64-75”
    IsolCpusList: “4-19,24-39,44-59,64-79”
    NeutronSriovNumVFs:
      - enp47s0f1:16:switchdev
      - enp47s0f0:16:switchdev
    NeutronPhysicalDevMappings: “prov704:enp47s0f1,prov703:enp47s0f0”
    NovaPCIPassthrough:
      - devname: “enp47s0f1”
        physical_network: “prov704”
      - devname: “enp47s0f0”
        physical_network: “prov703”
    TunedProfileName: “cpu-partitioning”
    NeutronSupportedPCIVendorDevs: [‘8086:158b’]
```

```
ComputeOvsDpdkParameters:

    NovaVcpuPinSet: “4-15,24-35,44-55,64-75”
    NovaSchedulerDefaultFilters:
“RamFilter,ComputeFilter,ServerGroupAffinityFilter,ServerGroupAntiAffinityFilter,AvailabilityZoneFilt
er,ComputeCapabilitiesFilter,ImagePropertiesFilter,PciPassthroughFilter,NUMATopologyFilter”
    KernelArgs: “isolcpus=4-19,24-39,44-59,64-79 nohz_full=4-19,24-39,44-59,64-79
default_hugepagesz=1GB hugepagesz=1G hugepages=192 iommu=pt intel_iommu=on”
    IsolCpusList: “4-19,24-39,44-59,64-79”
    OvsEnableDpdk: True
    TunedProfileName: “cpu-partitioning”
    NovaReservedHostMemory: 4096
    OvsDpdkSocketMemory: “2048,2048”
    OvsDpdkMemoryChannels: “4”
    OvsPmdCoreList: “16,17,18,19,36,37,38,39,56,57,58,59,76,77,78,79”
    DpdkBondInterfaceOvsOptions: “bond_mode=balance-tcp lacp=active”
    BondInterfaceOvsOptions: “mode=4 lacp_rate=1 updelay=1500 miimon=200”
    VhostuserSocketGroup: “hugetlbfs”
```

Modification of compute-dpdk.yaml and compute-sriov.yaml

The mappings of networks to NICs should be updated according to the network configuration. The following is an example of configuration of ovs-user-bridge with NIC-bonding on OVS-DPDK compute nodes:

```
-
  type: ovs_user_bridge
  name: br-link
  use_dhcp: false
  members:
    -
      type: ovs_dpdk_bond
      name: dpdkbond0
      ovs_options: {get_param: DpdkBondInterfaceOvsOptions}
      members:
        -
          type: ovs_dpdk_port
          name: dpdk0
          members:
            -
              type: interface
              name: enp47s0f0
        -
          type: ovs_dpdk_port
          name: dpdk1
          members:
            -
              type: interface
              name: enp47s0f1
```

On SR-IOV compute nodes, the NIC ports that have SR-IOV enabled should not be associated to any OVS or Linux bridge.

```
-
  type: interface
  name: enp47s0f0
  use_dhcp: false
  defroute: false
  nm_controlled: true
  hotplug: true
-
  type: interface
  name: enp47s0f1
  use_dhcp: false
  defroute: false
  nm_controlled: true
  hotplug: true
```

Modification of puppet-ceph-external.yaml

```
resource_registry:
  OS::TripleO::Services::CephExternal: /usr/share/openstack-tripleo-heat-
templates/puppet/services/ceph-external.yaml
  OS::TripleO::Services::CephMon: OS::Heat::None
```

```
    OS::TripleO::Services::CephClient: OS::Heat::None
    OS::TripleO::Services::CephOSD: OS::Heat::None
    OS::TripleO::Services::SwiftProxy: OS::Heat::None
    OS::TripleO::Services::SwiftStorage: OS::Heat::None
    OS::TripleO::Services::SwiftRingBuilder: OS::Heat::None

  parameter_defaults:
    CephClusterFSID: 'ba41970b-7d3b-4101-a96e-1c4ba58108ac'
    CephClientKey: 'AQBKuHFbxj3ROBAAfeBUESciizuB/62cZL9KFA=='
    CephExternalMonHost: '192.168.80.182,192.168.80.183,192.168.80.184'

    # the following parameters enable Ceph backends for Cinder, Glance, Gnocchi and Nova
    NovaEnableRbdBackend: true
    CinderEnableRbdBackend: true
    CinderBackupBackend: ceph
    GlanceBackend: rbd
    GnocchiBackend: rbd

    NovaRbdPoolName: cloud5_nova
    CinderRbdPoolName: cloud5_volumes
    CinderBackupRbdPoolName: cloud5_backups
    GlanceRbdPoolName: cloud5_images
    GnocchiRbdPoolName: cloud5_metrics
    CephClientUserName: cloud5_openstack

    CinderEnableIscsiBackend: false
    CephAdminKey: ''
```

8.3.4 Cloud deployment

The following is an example of the deployment script. For a full template example of Red Hat OpenStack Platform 13 with DPDK and SR-IOV enabled on Lenovo ThinkSystem SR650, please visit https://github.com/lenovo/ServiceProviderRA.

```
#!/bin/bash
source ~/stackrc
cd /usr/share/openstack-tripleo-heat-templates
sudo ./tools/process-templates.py    -r /home/stack/templates/roles_data.yaml –n
/home/stack/templates/network_data.yaml

cd /home/stack
openstack overcloud deploy \
--templates \
-e /usr/share/openstack-tripleo-heat-templates/environments/network-isolation.yaml \
-e /usr/share/openstack-tripleo-heat-templates/environments/host-config-and-reboot.yaml \
-e /usr/share/openstack-tripleo-heat-templates/environments/neutron-ovs-dpdk.yaml \
-e /usr/share/openstack-tripleo-heat-templates/environments/ovs-dpdk-permissions.yaml \
-r /home/stack/templates/roles_data.yaml \
-e /home/stack/templates/network-environment.yaml \
-e /home/stack/templates/puppet-ceph-external.yaml \
-e /home/stack/templates/overcloud_images.yaml \
-e /home/stack/templates/node-info.yaml \
--ntp-server pool.ntp.org
```

8.3.5 Post-deployment configurations

After the NFV compute nodes are successfully deployed, flavors created on the NFV compute nodes are recommended to have the following metadata tags:

5. hw:cpu_policy=dedicated
6. hw:cpu_thread_policy=require
7. hw:mem_page_size=large
8. hw:numa_nodes=1
9. hw:numa_mempolicy=strict

The vHost multi-queue can be enabled in VNFs by turning on the hw_vif_multiqueue_enabled option for the Glance images.

8.4 Storage implementation

For NFV deployment, Lenovo recommends using Red Hat OpenStack Director installation to create the Ceph cluster. The Ceph cluster consists of two main components:

- **Ceph OSD** (Object Storage Daemon) which is deployed on three dedicated Ceph nodes. The OSD nodes perform the replication, rebalancing, recovery, and reporting. Lenovo recommends three ThinkSystem SR650 to host Ceph OSDs. If a deployment requires higher storage capacity, more OSDs can be added to the Ceph cluster.
- **Ceph Monitor** maintains a master copy of Ceph storage map with the current state of the storage cluster. In this example, the OpenStack controller nodes are used to host Ceph monitor function.

8.5 Best practices

This section describes best practices for Lenovo NFVI solution using Lenovo ThinkSystem servers.

For detailed deployment steps for OpenStack Platform 13 deployment, please see the Red Hat documentation “Director Installation and Usage”.

The best practices are:

- All nodes must be time synchronized by using NTP at all times.
- Lenovo XClarity Controller (XCC) is a unique management module of Lenovo x86 servers. Users can use web browsers or SSH to connect the management interface to configure and manage servers. Firstly, configure an IP address for XCC (Boot the server → Enter “F1” → Select “RAID Setup ” → Select “Manage Disk Drives” → Drop down to “Change all disk drives state from JBOD to Ugood” → Select “Next” → Select “Advanced configuration” → Select “Next” → Create Virtual disks as your designed configuration.)
- RAID can be configured through XCC web console.(Login XCC web console → “Server Management” dropdown list → “Local Storage” → Select the RAID controller and click “Create Volume” → Complete the wizard)

- The NIC MAC addresses need to be written into “instackenv.json” file. It is used by bare metal service (Ironic). This useful information can be obtained through XCC web console
- Get MAC address (Login XCC web console → Select “Inventory” tab → Drop down to “PCI Adapters” → Launch the network adapter’s properties and view “Physical Ports”)
- The timeout set should be 60 sec in the ironic.conf file, in order to avoid XCC connection failure.
- Both UEFI and Legacy boot mode are supported for overcloud deployment. It is recommended to use UEFI boot mode for better support of peripheral devices:
- For UEFI mode configuration: Lenovo Servers’ default boot mode is UEFI. If you want to deploy through UEFI mode, you need to change the properties of the nodes registered to ironic, and modify the configure file “undercloud.conf”. Please see the Red Hat documentation Director Installation and Usage.
- For Legacy mode configuration: By default, Lenovo Servers’ “PXE Boot” feature is “DISABLED”. If you want to use the Legacy mode, you need to enable the PXE boot(Boot the server → Enter “F1” → Select “Network” → Select Network Boot Settings” -> Select the port you need for PXE Boot -> Set “Legacy PXE Mode” to “Enable”)
- For data plane services, Lenovo recommends using port bonding, e.g. LACP to avoid single point of failure on network equipment. For OpenStack network isolation traffic (InternalAPI, Storage and Tenant), Linux bonds are used on the compute node. On control node ovs_bridge is used for network isolation traffic. When using OVS-DPDK, all bridges on the same compute node should be of type ovs_user_bridge. It is not supported to use ovs_bridge (type: system) and ovs_user_bridge (type: netdev) on the same node. Since some compute nodes need ovs_user_bridge for the higher performance OVS-DPDK, OVS DPDK bond is recommended to use in such scenario.

9 Appendix: Lenovo Bill of Materials

This appendix contains the Bill of Materials (BOMs) for different hardware configurations for the Lenovo NFVI solution. There are sections for compute nodes, utility node, controller nodes, storage nodes, networking, rack options, and software.

9.1 Server BOM

The following section contains the BOM for the Lenovo NFVI solution for service provider implementation using Lenovo ThinkSystem Servers.

9.1.1 Utility node

Code	Description	Quantity
7X02CTO1WW	ThinkSystem SR630 – 3yr Warranty	1
AUW0	ThinkSystem SR630 2.5” Chassis with 8 bays	1
AWEP	Intel Xeon Gold 5118 12C 105W 2.3GHz Processor	2
AUNC	ThinkSystem 16GB TruDDR4 2666 MHz (2Rx8 1.2V) RDIMM	12
BOWY	ThinkSystem Intel XXV710-DA2 PCIe 25Gb 2-port	1
AUKH	ThinkSystem 1Gb 4-port RJ45 LOM	1
AV1X	Lenovo 3m Passive 25G SFP28 DAC Cable	2
AVWB	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
AUMV	ThinkSystem M.2 with Mirroring Enablement Kit	1
B11V	ThinkSystem M.2 5100 480GB SATA 6Gbps Non-Hot-Swap SSD	2
Optional local storage		
AUNJ	ThinkSystem RAID 930-8i 2GB Flash PCIe 12Gb Adapter	1
AUM2	ThinkSystem 2.5” 1.8TB 10K SAS 12Gb Hot Swap 512e HDD	8

9.1.2 Compute node

Code	Description	Quantity
7X06CTO1WW	ThinkSystem SR650 - 3yr Warranty	1
AUVV	ThinkSystem SR650 2.5" Chassis with 8, 16 or 24 bays	1
AWEM	Intel(R) Xeon(R) Gold 6152 CPU @ 2.10GHz	2
AUND	ThinkSystem 32GB TruDDR4 2666 MHz (2Rx4 1.2V) RDIMM	12
B0WY	ThinkSystem Intel XXV710-DA2 PCIe 25Gb 2-port	2
AUKH	ThinkSystem 1Gb 4-port RJ45 LOM	1
AV1X	Lenovo 3m Passive 25G SFP28 DAC Cable	4
AVWF	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
6311	2.8m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
AUNJ	ThinkSystem RAID 930-8i 2GB Flash PCIe 12Gb Adapter	1
B49M	ThinkSystem 2.5" Intel S4610 480GB Mainstream SATA 6Gb HS SSD	2
Optional local storage		
B58G	ThinkSystem U.2 Intel P4510 2.0TB Entry NVMe PCIe 3.0 x4 Hot Swap SSD	2

9.1.3 Controller node

Code	Description	Quantity
7X02CTO1WW	ThinkSystem SR630 - 3yr Warranty	1
AUW1	ThinkSystem SR630 2.5" Chassis with 10 Bays	1
AWEL	Intel Xeon Gold 6126 12C 125W 2.6GHz Processor	2
AUNC	ThinkSystem 16GB TruDDR4 2666 MHz (2Rx8 1.2V) RDIMM	12
BOWY	ThinkSystem Intel XXV710-DA2 PCIe 25Gb 2-port	2
AUKH	ThinkSystem 1Gb 4-port RJ45 LOM	1
AV1X	Lenovo 3m Passive 25G SFP28 DAC Cable	4
AVWB	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
6311	2.8m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2
AUNK	ThinkSystem RAID 930-16i 4GB Flash PCIe 12Gb Adapter	1
B49M	ThinkSystem 2.5" Intel S4610 480GB Mainstream SATA 6Gb HS SSD	2
Optional local storage		
AUM2	ThinkSystem 2.5" 1.8TB 10K SAS 12Gb Hot Swap 512e HDD	8

9.1.4 Storage node

Code	Description	Quantity
7X06CTO1WW	ThinkSystem SR650 - 3yr Warranty	1
AUVV	ThinkSystem SR650 2.5" Chassis with 8, 16 or 24 bays	1
AWER	Intel Xeon Silver 4116 12C 85W 2.1GHz Processor	2
AUNC	ThinkSystem 16GB TruDDR4 2666 MHz (2Rx8 1.2V) RDIMM	12
AUNK	ThinkSystem RAID 930-16i 4GB Flash PCIe 12Gb Adapter	1
B49M	ThinkSystem 2.5" Intel S4610 480GB Mainstream SATA 6Gb HS SSD	2
AUM2	ThinkSystem 2.5" 1.8TB 10K SAS 12Gb Hot Swap 512e HDD	12
B4Y4	ThinkSystem 2.5" SS530 400GB Performance SAS 12Gb Hot Swap SSD	2
BOWY	ThinkSystem Intel XXV710-DA2 PCIe 25Gb 2-port	2
AUKG	ThinkSystem 1Gb 2-port RJ45 LOM	1
AV1X	Lenovo 3m Passive 25G SFP28 DAC Cable	4
AVWF	ThinkSystem 1100W (230V/115V) Platinum Hot-Swap Power Supply	2
6311	2.8m, 10A/100-250V, C13 to IEC 320-C14 Rack Power Cable	2

9.2 Networking BOM

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

9.2.1 G8052 1GbE Switch

Code	Description	Quantity
7159G52	Lenovo RackSwitch G8052 (Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

9.2.2 NE0152T 1GbE Switch

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

9.2.3 G8272 10GbE Switch

Code	Description	Quantity
7159CRW	Lenovo RackSwitch G8272 (Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

9.2.4 NE1032 10GbE Switch

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

9.2.5 NE1032T 10GbE Switch

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

9.2.6 NE1072T 10GbE Switch

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

9.2.1 NE2572 25GbE Switch

Code	Description	Quantity
AV19	Lenovo ThinkSystem NE1072T RackSwitch (Rear to Front)	1
39Y7938	2.8m, 10A/100-250V, C13 to IEC 320-C20 Rack Power Cable	2

9.3 Rack BOM

This section contains the BOM for the 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

9.4 Red Hat subscription options

This section contains the BOM for the Red Hat Subscriptions. See Lenovo Rep for final configuration.

Code	Description	Quantity
00YH835	Red Hat OpenStack Platform, 2 socket, Premium RH Support, 3 yrs	Variable
00YH839	Red Hat OpenStack Platform Controller Node, 2 skt, Prem RH Support, 3 yrs	1
00YH849	Red Hat Ceph Storage, 12 Physical Nodes, to 256TB, Prem RH Support, 3 yrs	1

Resources

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

- OpenStack Project:
  openstack.org
- OpenStack Operations Guide:
  docs.openstack.org/ops/
- Red Hat OpenStack Platform:
  access.redhat.com/documentation/en/red-hat-openstack-platform/13
- Red Hat Ceph Storage:
  redhat.com/en/technologies/storage/ceph
- Red Hat CloudForms:
  access.redhat.com/documentation/en-us/red_hat_cloudforms
- Example of Red Hat OpenStack Platform deployment templates:
  github.com/lenovo/ServiceProviderRA

Document History

Version 1.0	18 September 2018	• Initial version
Version 2.0	21 February 2019	• Updated for Red Hat OpenStack 13 • Updated to support SR-IOV configuration • Upgraded for Intel 25Gbe NICs and 25G switches for networking
Version 2.1	2 May 2019	• Added performance benchmarking

Trademarks and special notices

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo, the Lenovo logo, AnyBay, AnyRAID, BladeCenter, NeXtScale, RackSwitch, Rescue and Recovery, ThinkSystem, System x, ThinkCentre, ThinkVision, ThinkVantage, ThinkPlus and XClarity are trademarks of Lenovo.

Red Hat, Red Hat Enterprise Linux and the Shadowman logo are trademarks of Red Hat, Inc., registered in the U.S. and other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries. The OpenStack mark is either a registered trademark/service mark or trademark/service mark of the OpenStack Foundation, in the United States and other countries, and is used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel Inside (logos), and Xeon are trademarks of Intel Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.