REFERENCE ARCHITECTURE

QxStack NFV Infrastructure with Red Hat OpenStack Platform

A Carrier-Grade Infrastructure Pre-Integrated and Validated for Network Service Providers

Version 1.0

CONTENT

TABLES

FIGURES

1. Overview

Exploding traffic on media, mobile, and high-tech applications is accompanied by the increasing need of high speed network in the era of the Internet. Traditional fixed appliances are inadequate to deal with the growing data, and it is too pricey to expand the legacy network infrastructure to process the complicated workloads. Facing the inevitable digital transformation, Communication Service Providers (CSPs) are looking for emerging software-defined infrastructure to easily manage network resources and keep a dominant position in the telco market.

Network Function Virtualization (NFV) is recognized as a key, which decouples network service from proprietary hardware, and runs workloads on common hardware and open source software. Featuring the modularized software-defined infrastructure, OpenStack is considered to be a matching foundation for NFV for its flexibility on cloud infrastructure and management. According to OpenStack Foundation Report, an agile and scalable OpenStack platform with compelling technical and business benefits is gaining popularity among telecommunications companies and CSPs.

Quanta Cloud Technology (QCT) are dedicated to delivering performance-optimized reference designs for a broad enterprise use cases on OpenStack solutions. We offer an NFV infrastructure solution to meet carrier-grade requirements for datacenter and network innovations. This solution delivers exceptional network performance to satisfy digital demands and accelerates cloud transformation based on integrated hardware and software with non-uniform memory access (NUMA) balanced design and accelerated data plane packet processing technologies.

The QxStack NFV Infrastructure with Red Hat OpenStack Platform solution proposes the industry-standard x86 hardware, solution software, and network planning as the reference architecture for CSPs to efficiently implement the designated NFV infrastructures. Meanwhile, this paper also introduces the integration of network virtualization technologies, resource allocation strategies, and performance evaluation with industry-leading OPNFV test suite.

2. Solution Hardware

A main objective of NFV is to decouple network functions from proprietary appliances and bring a flexible, high-performance platform based on industry-standard servers. The QxStack NFV Infrastructure with Red Hat Open-Stack Platform solution is built up with QCT high expandable x86 hardware, supporting a fully virtualized infra-structure including servers, storage, and networking.

2.1. QuantaGrid D52B-1U: controller and compute node

QuantaGrid D52B-1U[1] is an ultra-dense server with up to 5 PCIe expansion slots per chassis. It features flexible I/O options, including a variety of SAS Mezzanine and OCP NIC/ PHY Mezzanine cards in three different SKUs. With NUMA-balanced design, *QuantaGrid D52B-1U* supports data plane acceleration technologies like DPDK and SR-IOV to accommodate the most demanding Telco workloads. This 1U server with Intel® Xeon® Processor Scalable Family is designed to enhance computing power, provide faster socket interconnect, and strengthen memory bandwidth.

Table 1. Solution hardware - QuantaGrid D52B-1U

	QuantaGrid D52B-1U **Default Configuration**
Processor	Intel® Xeon® Processor Scalable Family
Storage	SFF tiered SKU: • 8x 2.5" hot-plug SATA/SAS drives • 4x 2.5" hot-plug NVMe/SATA/SAS drives
Memory	Total slots: 24 Capacity: Up to 3TB of memory for RDIMM/LRDIMM Memory type: 2666 MHz DDR4 RDIMM
Network	LOM: Dedicated 1x GbE management port
Expansion Slot	SFF tiered SKU, option 1: • 1x PCIe Gen3 x16 SAS mezzanine slot • 1x PCIe Gen3 x16 OCP 2.0 mezzanine slot or PHY card • 1x PCIe Gen3 x16 LP MD-2 • 1x PCIe Gen3 x8 LP MD-2

[1] *QCT QuantaGrid D52B-1U:*
https://www.qct.io/product/index/Server/rackmount-server/1U-Rackmount-Server/QuantaGrid-D52B-1U

2.2. QuantaGrid D51PH-1ULH: storage node

QuantaGrid D51PH-1ULH[2] features hybrid-tiered storage architecture in an ultra-dense 1U platform. It is tailored for hyperscale data centers and software-defined storage solutions based on Intel® Xeon® processor E5-2600 v3, v4 product families and with up to 1TB memory capacity. Equipped with 12 hot-swappable 3.5" disk drives and 4 hot-swappable 2.5" SATA SSDs, *QuantaGrid D51PH-1ULH* is ideal for tier storage planning to accelerate IOPs and throughput without sacrificing large data storage capacity.

Table 2. Solution hardware – QuantaGrid D51PH-1ULH

	QuantaGrid D51PH-1ULH **Default Configuration**
Processor	Intel® Xeon® processor E5-2600 v3, v4 product family
Storage	Options: • 12x 3.5"/2.5" hot-plug 12Gb/s SAS or 6Gb/s SATA HDD/SSD • 4x 2.5" hot-plug 7mm 6Gb/s SATA SSD • 1x Internal SATA DOM
Memory	Total slots: 16 Capacity: Up to 512GB RDIMM Memory type: 2400/2133/1866/1600/1333 MHz DDR4 RDIMM
Network	LOM: Dedicated 10/100/1000 management port
Expansion Slot	Options: • 1x SAS Mezzanine x8 • 1x OCP LAN Mezzanine slot x8

2.3. QuantaMesh T3048-LY9: management switch

QuantaMesh T3048-LY9[3] is a new generation 10GBASE-T solution for data center networking which features 48 triple speed (100/1000/10GBase-T) ports and 6 QSFP+ ports in a 1U form factor. It offers hardware-based VXLAN to support virtual machine mobility. *QuantaMesh T3048-LY9* is designed for high availability from both hardware and software perspectives. It also aims to facilitate Infrastructure-as-a-service (IaaS) networking deployment, boosting more cost-effective performance and performing the best TCO.

[2] *QCT QuantaGrid D51PH-1ULH:*
https://www.qct.io/product/index/Storage/Storage-Server/1U-Storage-Server/QuantaGrid-D51PH-1ULH
[3] *QCT QuantaMesh T3048-LY9:*
https://www.qct.io/product/index/Networking/Ethernet-Switch/T3000-Series/QuantaMesh-T3048-LY9

Table 3. Solution hardware - QuantaMesh T3048-LY9

	QuantaMesh T3048-LY9 **Default Configuration**
Physical ports	Port configuration: • 48x 100/1000/10GBASE-T • 6x QSFP+ ports
Performance	Switching capacity: 1.44Tbps Maximum forwarding rate: 1071Mpps Latency: <3us
High Availability	Features: • 1+1 hot-swappable power supplies • 3+1 hot-swappable fans • Up to 48 paths ECMP routing for load balancing and redundancy • Multi-chassis Link Aggregation (MLAG)

2.4. QuantaMesh T4048-IX2: Top-of-Rack (ToR) switch

By leveraging merchant silicon chipsets, *QuantaMesh T4048-IX2*[4] is a high-performance, high-density Ethernet switch for the deployment of datacenter infrastructure. *QuantaMesh T4048-IX2* is pre-loaded with Open Network Installation Environment (ONIE) to provide high agility. Any operating system that supports ONIE installer can be easily installed when customer's demand changes. *QuantaMesh T4048-IX2* is a BMC built-in Ethernet switch which can provide health monitoring of the temperature, power status, and aids in the deployment and management of software and hardware peripherals.

Table 4. Solution hardware - QuantaMesh T4048-IX2

	QuantaMesh T4048-IX2 **Default Configuration**
Physical ports	Port configuration: • 48x SFP28 (10/25GbE) • 8x QSFP28 ports (10/25/40/50/100GbE)
Performance	Switching capacity: 4Tbps Maximum forwarding rate: Line Rate Performance Latency: < 450ns
High Availability	Features: • Redundant power supply: 1+1 • Hot-swappable fan tray: N+1

[4] *QCT QuantaMesh T4048-IX2:* *https://www.qct.io/product/index/Networking/Bare-Metal-Switch/Leaf-Switch/QuantaMesh-BMS-T4048-IX2*

3. Solution Software

In the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution, the BIOS and firmware are upgraded to the latest version for hardware management, and the switch is installed with QNOS 5 to support VLAN, Multi-Chassis Link Aggregation Group (MC-LAG), and LACP Bypass. The detailed software information is listed below:

Table 5. Solution software version

Hardware management	
QuantaGrid D52B-1U	BIOS Version: 3A08.E1 Firmware Version: 3.46.00
QuantaGrid D51PH-1ULH	BIOS Version: S2P_3A16 Firmware Version: 3.30.00
QuantaMesh T4048-IX2	Operating System: QNOS 5
Operating system and solution software	
Red Hat Enterprise Linux	Version 7.4
Red Hat OpenStack Platform	Version 10
Red Hat OpenStack Platform director	Version 10
Red Hat Ceph Storage	Version 2.0

3.1. Red Hat OpenStack Platform 10

Red Hat OpenStack Platform is the de-facto choice for a virtualized infrastructure manager (VIM) for leading edge Network Function Virtualization (NFV) deployments. The OpenStack Networking Service's pluggable architecture is a key enabling technology for NFV. Many Red Hat-supported open source projects and communities are focused on accelerating cloud networking performance and NFV with the support of all major network equipment providers and telcos. A short list of relevant Red Hat technologies includes libvirt, Data Plane Development Kit (DPDK), Open vSwitch, QEMU/KVM, and Linux.

Red Hat's NFV focus is on the infrastructure (NFVI), especially the VIM layer for enabling Virtual Network Functions (VNFs). Red Hat partners with Independent Software Vendors (ISVs), for NFV management and organization (MANO), and with infrastructure providers like QCT, building the technology that enables advanced network functionality both in Red Hat OpenStack Platform and across the entire network stack. By developing the capabilities into the platform and operating system and leveraging those features and tools, high throughput and low latency are achieved for applications and use cases, while being fully integrated, supported and performance-tuned.

Red Hat OpenStack Platform can be turned into a private, public, or hybrid cloud platform that includes:

- Authentication and authorization mechanisms
- Fully distributed object storage

- Integrated networking
- Persistent block-level storage
- Virtual machine provisioning engine and image storage
- Web browser-based interface accessible to users and administrators

Other noteworthy and important benefits and features of Red Hat OpenStack Platform include:

- **Reliable deployments with live upgrades:** The director tool in Red Hat OpenStack Platform checks systems throughout the installation process to provide consistent, automated cloud deployment. It features live orchestrated system upgrades and updates, ensuring long-term, production-ready stability with little downtime
- **Integrated orchestration:** Red Hat OpenStack Platform director provides system-wide orchestration of OpenStack resources, including bare-metal provisioning.
- **Unlimited Red Hat Enterprise Linux uses:** Red Hat Enterprise Linux can operate on host nodes and unlimited virtualized workloads on OpenStack[5].
- **Included workload and infrastructure management:** Red Hat CloudForms can manage OpenStack workloads and infrastructure. It provides resource management and data collection over OpenStack clouds, such as resource monitoring and reporting, compliance assurance, chargeback and showback, service cataloging, user management, and heat template management
- **Reliable storage:** Every company or organization that purchases Red Hat OpenStack Platform receives one subscription of 64TB of Red Hat Ceph Storage. This helps customers get started with highly scalable and redundant object, block, and file storage[6].
- **Hardened for communication service providers:** An extensive patching, bug-fixing, testing, and certification process ensures broad compatibility and performance with upstream community releases.
- **Highly available infrastructure:** Red Hat OpenStack Platform maintains high availability and policy-driven measures, including infrastructure failure recognition, automated host node evacuation, and downed node fencing, and automatically restarts workloads on remaining available hosts.
- **Enterprise software life cycle:** Red Hat provides stable branch releases of OpenStack and Linux that are supported for an enterprise production life cycle—beyond the six-month release cycle of the OpenStack community. Customers can choose to standardize for up to five years on certain releases or stay on top of the latest features by updating every six months to one year. Red Hat OpenStack Platform's lifecycle can be found here.
- **Expansive ecosystem:** Red Hat has built an expansive certified OpenStack partner ecosystem for Red Hat OpenStack Platform. It includes thousands of certified servers and third-party software, plus an

---

[5] *Red Hat OpenStack Platform is available for purchase with or without unlimited Red Hat Enterprise Linux guests. Both versions include Red Hat Enterprise Linux for OpenStack host controller nodes.*

[6] *One subscription of 64TB of Red Hat Ceph Storage per company or organization regardless of number of Red Hat OpenStack Platform subscriptions purchased. Additional capacity of Red Hat Ceph storage sold separately.*

OpenStack-specific certification program with partners in the compute, storage, networking, independent software vendor (ISV) software, and service fields.

- **Technology leadership:** Red Hat is a top code contributor to many OpenStack projects and a longtime leader in the OpenStack, Ceph storage, and broader Linux communities—making Red Hat an ideal supporter of full-scale OpenStack deployments.
- **Security:** Security-Enhanced Linux (SELinux) military-grade security technologies prevent intrusions and protect data when running in public or private OpenStack clouds.
- **Performance:** The Red Hat Virtualization Hypervisor provides superior performance for OpenStack workloads. Based on Kernel-based Virtual Machine (KVM), the hypervisor holds record-breaking performance scores on the SPECvirt_sc2013 benchmark[7].
- **World-class global support, professional consulting services, and certified training:** Red Hat provides global support services to help customers running critical infrastructure like Red Hat OpenStack Platform and Red Hat Ceph Storage. Customers with an active subscription can contact Red Hat support engineers via telephone, e-mail, and an available web portal. Additionally, Red Hat also has dedicated consulting services and a Technical Account Manager (TAM) team available to work closely with customers. To ensure customers are prepared to operate the system, Red Hat develops end-user training and certification programs.

<u>Red Hat OpenStack Platform director</u>

Red Hat OpenStack Platform director is the installation and lifecycle management tool for Red Hat OpenStack Platform. Director is based primarily on the OpenStack project TripleO. This project takes advantage of OpenStack components to install an OpenStack environment.

The Red Hat OpenStack Platform director uses two main concepts: an Undercloud and an Overcloud, where the Undercloud installs and configures the Overcloud.

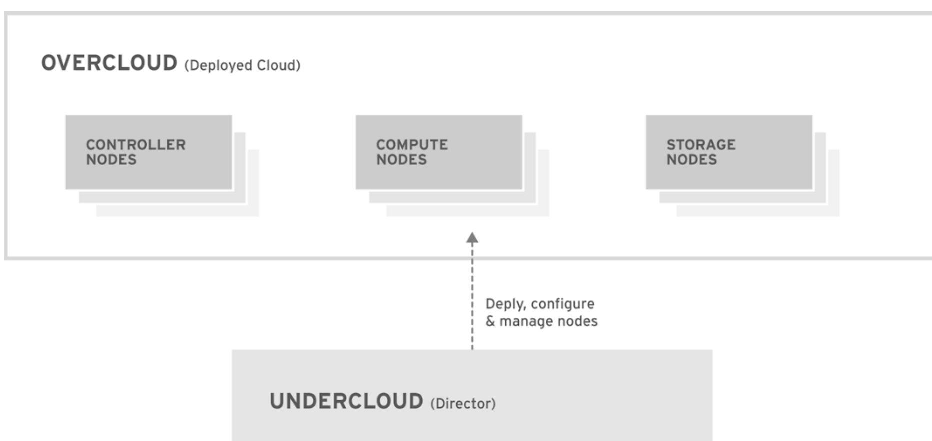

Figure 1. Basic layout of undercloud and overcloud

[7] *All comparisons are based on a benchmark addressing performance evaluation of datacenter servers used in virtualized server consolidation at www.spec.org/virt_sc2013/ as of March 10, 2016. SPEC® and the benchmark name SPECvirt_sc® are registered trademarks of the Standard Performance Evaluation Corporation (SPEC).*

Undercloud

The Undercloud is the main director node. It is a single-system OpenStack installation that includes components for provisioning and managing the OpenStack nodes that comprise the OpenStack environment (the Overcloud). The components that form the Undercloud provide the following functions:

- **Environment planning:** The Undercloud provides planning functions for users to assign Red Hat OpenStack Platform roles, including Compute, Controller, and various storage roles.
- **Bare metal system control:** The Undercloud uses various configurable drivers including Intelligent Platform Management Interface (IPMI) for power management control and a PXE-based service to discover hardware attributes and install OpenStack on each node. This provides a method to provision bare metal systems as OpenStack nodes.
- **Orchestration:** The Undercloud provides and reads a set of YAML templates to create an OpenStack environment.

The Red Hat OpenStack Platform director utilizes these Undercloud functions through both a web-based graphical user interface and a terminal-based command line interface. The Undercloud uses the following components:

- **OpenStack Identity (keystone):** Provides authentication and authorization for the director's components.
- **OpenStack Bare Metal (ironic) and OpenStack Compute (nova):** Manages bare metal nodes.
- **OpenStack Networking (neutron) and Open vSwitch:** Controls networking for bare metal nodes.
- **OpenStack Image Service (glance):** Stores images that are written to bare metal machines.
- **OpenStack Orchestration (heat) and Puppet:** Provides orchestration of nodes and configuration of nodes after the director writes the Overcloud image to disk.
- **OpenStack Telemetry (ceilometer):** Performs monitoring and data collection. This also includes:
    - **OpenStack Telemetry Metrics (gnocchi):** Provides a time series database for metrics.
    - **OpenStack Telemetry Alarming (aodh):** Provides a an alarming component for monitoring.
    - **OpenStack Workflow Service (mistral):** Provides a set of workflows for certain director-specific actions, such as importing and deploying plans.
    - **OpenStack Messaging Service (zaqar):** Provides a messaging service for the OpenStack Workflow Service.
- **MariaDB:** Database for the director.
- **RabbitMQ:** Messaging queue for the director's components

The Undercloud can be deployed both as a baremetal node or virtualized on RHEL KVM, or Red Hat Virtualization (RHV), or other supported Virtualization platforms. In this solution, the Undercloud is deployed atop of RHEL KVM.

Overcloud

The Overcloud is the resulting Red Hat OpenStack Platform environment created using the Undercloud and where the user facing workloads or vNFs will run. This includes one or more of the following node types:

Controller nodes provides administration, networking, and high availability for the OpenStack environment. This solution contains three nodes together in a high availability cluster for the resulting OpenStack environment, which is a recommended best practice. A default Controller node contains the following components: Horizon, Keystone, Nova API, Neutron Server, Open vSwitch, Glance, Cinder Volume, Cinder API, Swift Storage, Swift Proxy, Heat Engine, Heat API, Ceilometer, MariaDB, MongoDB, and RabbitMQ. The Controller also uses Pacemaker and Galera for services high availability. Compute nodes provide computing resources for the OpenStack environment. Add more Compute nodes to scale your environment over time. A default Compute node contains the following components: Nova Compute, Nova KVM, Ceilometer Agent, Open vSwitch, Neutron L2 agent. Storage nodes provide storage for the OpenStack environment. This includes nodes for:

- **Ceph Storage nodes:** Used to form storage clusters. Each node contains a Ceph Object Storage Daemon (OSD). In addition, the director installs Ceph Monitors on to Controller nodes in situations where it deploys Ceph Storage nodes.
- **Block storage (Cinder):** Used as external block storage for HA Controller nodes. This node contains the following components: Cinder Volume, Ceilometer Agent, Open vSwitch.
- **Object storage (Swift):** Used as external object storage for HA Controller nodes. This node contains the following components: Cinder Storage, Ceilometer Agent, Open vSwitch.

3.2. Enhanced platform awareness features for NFVI

3.2.1. PCI passthrough technologies

In generic OpenStack compute nodes, network traffic from VM flows through a complicated network virtualization stack to access the Internet including TAP device, Linux bridge, veth pair and Open vSwitch, as shown in Figure 1. To accelerate data plane packet processing, the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution supports memory huge pages in the compute nodes through either DPDK or SR-IOV.

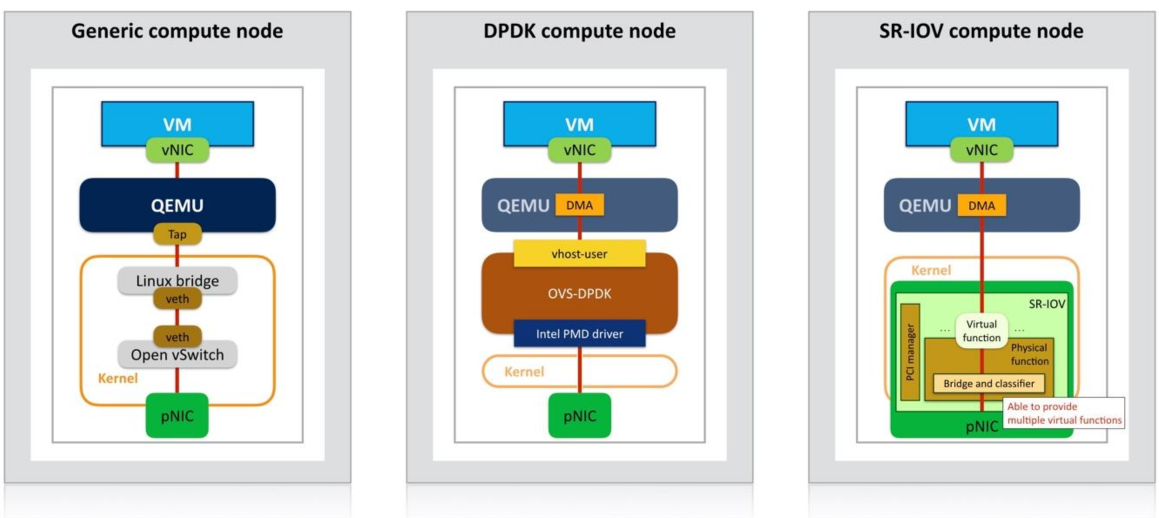

Figure 2. Comparison of network virtualization stacks

DPDK consists of a set of data plane libraries and user-space network drivers for accelerating packet processing. It provides a programmable framework that implements a run-to-completion model, eliminates packet interrupt processing overhead, and enables applications to perform packet processing operations directly from and to the NIC. This significantly improves network throughput and latency performance in Red Hat OpenStack Platform. As shown in Figure 2, the QCT NFV infrastructure solution uses a DPDK-accelerated version of Open vSwitch (OVS-DPDK) to enhance network performance. In this case, OVS-DPDK replaces the standard OVS kernel datapath with a DPDK-based datapath, creating a user-space Open vSwitch (OVS) for packet forwarding. OVS-DPDK efficiently allocates virtual host (vhost) memory across NUMA nodes while remaining transparency in the overall architecture and exposing the same interfaces—including OpenFlow, Open vSwitch Database (OVSDB), and command lines—as the standard OVS implementation.

SR-IOV is a specification that allows physical PCI devices to be shared between multiple virtual machines (VMs) for increasing network performance. SR-IOV virtualizes PCI hardware devices to create multiple virtual functions (VFs)—lightweight functions that can be assigned to specific VMs—on top of a physical functions (PFs)—full-feature physical hardware ports. A VF driver is required to implement SR-IOV. This driver resides in the VM, introduces VFs to the VM as physical NICs, and allows the VM to communicate directly with the physical device. Network traffic from a VM with a direct-attached VF bypasses the software switching layer to achieve near line-rate performance.

Both OVS-DPDK and SR-IOV take advantage of memory huge pages. Physical memory is typically segmented into 4KB pages. Memory huge pages increase the size of these memory blocks to either 2MB or 1GB, reducing the number of pages needed for a given amount of data. This increases the amount of memory that can be mapped by the translation lookaside buffer (TLB) to reduce the potential TLB misses and improve computational performance.

3.2.2. Resource allocation strategy

To optimize resource allocation, the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution supports CPU pinning and features a NUMA-aware design.

In virtualized infrastructures, a pool of physical CPUs (pCPUs) on a host are shared across multiple vCPUs associated with VMs. CPU pinning enables one-to-one mapping between vCPUs and pCPUs to increase VM performance. Because VMs run as user-space tasks within the host operating system, CPU pinning provides similar advantages to task pinning. As shown in Figure 3, CPU pinning dedicates specific compute resources to specific VMs and increases cache efficiency.

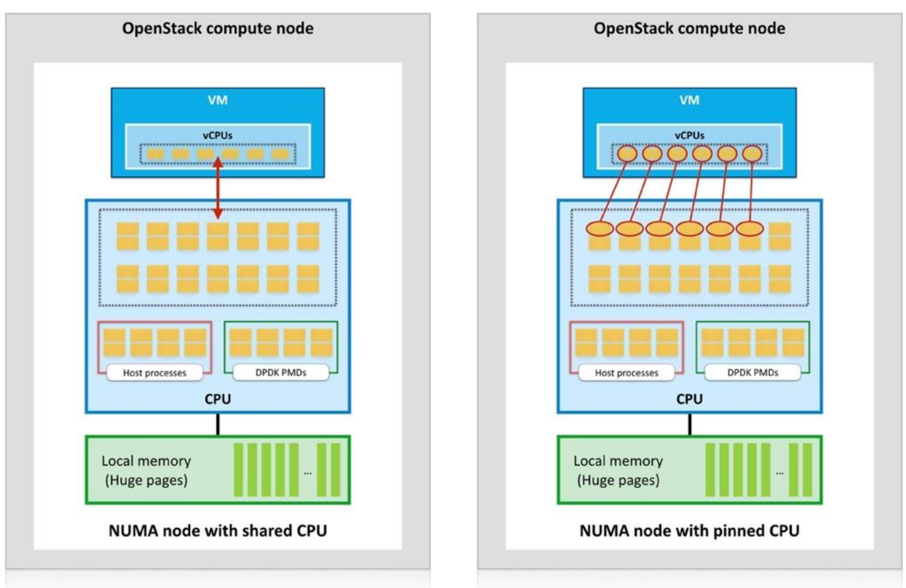

Figure 3. CPU topology and core allocation

Traditional uniform memory access (UMA) architecture models share memory resources evenly across all CPUs and sockets in a multiprocessor system. This often results in long memory access time, regardless of the location of the memory in relation to the CPU or socket. NUMA architecture models geographically-distributed system memory by considering its location in relation to each CPU, speeding the access to memory that is closer to the CPU. Processes can then access local CPU memory—rather than another CPU's local memory or shared memory—to improve computational performance. In Red Hat OpenStack Platform, OpenStack Compute (Nova) intelligently schedules and places memory when launching instances. Administrators can create instance configurations customized for specific performance levels to target specialized workloads like NFV and high-performance computing (HPC).

The QxStack NFV Infrastructure with Red Hat OpenStack Platform solution uses a NUMA-balanced design that supports local memory access, and distributes NICs across CPUs and sockets, as shown in Figure 4.

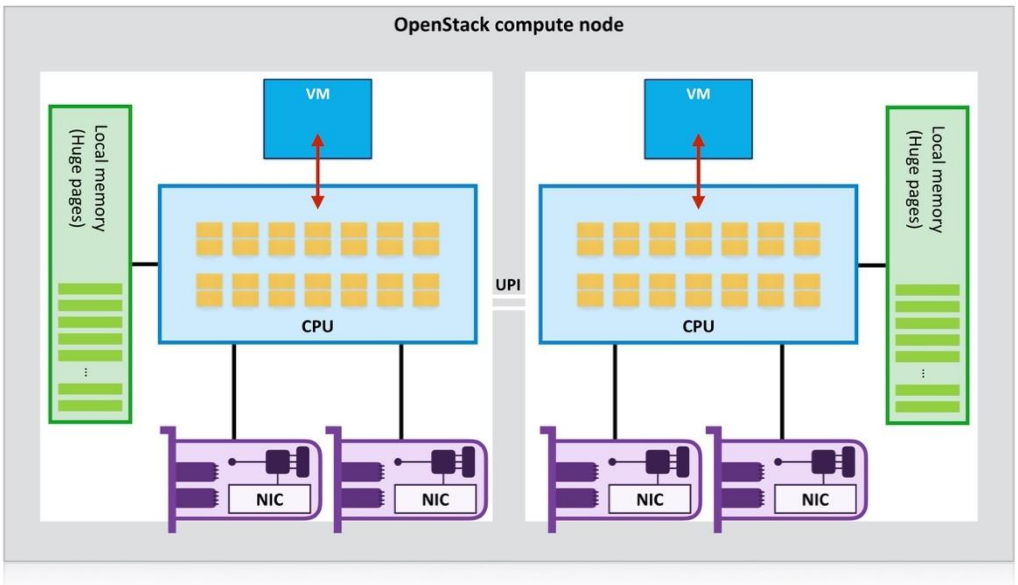

Figure 4. NUMA-balanced hardware design

3.3. Solution functionality

The QxStack NFV Infrastructure with Red Hat OpenStack Platform solution is designed to fulfill the strict requirements of CSPs and features the following benefits.

<u>Flexible design with scalability</u>

Based on standard x86 hardware design and the integration of Red Hat OpenStack Platform and Red Hat Ceph Storage, the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution provides a flexible design with scalable resources. Unlike traditional operating model with proprietary equipments, telecom operators are able to resize computing and storage sources based on their business strategies with little overhead, which successfully improves the capital expenditure (CAPEX) and operating expense (OPEX).

Reliable design with high resource availability

A reliable platform is critical to introduce VNFs into NFV infrastructure. To keep an environment up and running with strong reliability, QCT work on several measures aimed at providing extraordinary resource availability, load balancing, and fault tolerance to avoid business downtime. In this reference architecture, three controllers with Pacemaker and HAProxy are employed to ensure the high availability of cluster resource and control plane network traffics. Three Red Hat Ceph storage nodes are recommended as the backend of OpenStack services which stores at least three copies of data for data safety and periodically checks its own state for data retrieving and OSD rebalancing.

In addition to existing HA features in compute and storage nodes, QCT also provides high availability over networking. The top-of-rack (ToR) networking switches are designed using link aggregation technology to reduce the occurrence of high severity faults on network infrastructures. For controller and compute nodes, four dual-port Network Interface Cards (NICs) are used, and the ports on the same NIC are connected to different switches to prevent system crash resulting from switch failure.

Feasible deployment using the QxStack Auto-Deployment Tool

Considering the whole picture of NFV infrastructure deployment from datacenters to central offices across different countries and areas, a feasible deployment tool is indispensable to facilitate the deployment and to ease the pressure of time-to-market. The QCT QxStack Auto-Deployment Tool is an Ansible-based automation which simplifies the heat template customization and provides streamlined setup process with error tolerance. The tool defines the required parameters in key-value format and scripts the deployment process from platform setting, EPA features enablements, and other customization, which dramatically decreases the installation time from a week to just several hours. For more details, please refer to QxStack with Red Hat OpenStack Platform[8].

[8] *QxStack with Red Hat OpenStack platform:*
http://go.qct.io/solutions/enterprise-private-cloud/qxstack-with-red-hat-openstack-platform/

4. Solution Configuration

4.1. Hardware configuration

The hardware configuration designed in this reference architecture includes one director node, three controller nodes, three compute nodes with DPDK or SR-IOV enabled, and three Red Hat Ceph Storage nodes. For Red Hat OpenStack Platform director, controller, and compute nodes, Intel® Xeon® Gold 6152 processor with 22 CPU cores is chosen to provide outstanding performance. As for the storage nodes, Intel® Xeon® E5-2620 v4 with 10 CPU cores is selected to provide sufficient processing power. The Solid-State Drives (SSDs) with RAID 1 are enabled on the Red Hat OpenStack Platform director, controller, and compute nodes to guarantee high availability for the operating system. The hardware configuration is shown in Table 6.

Table 6. Solution hardware configuration

Role	Model	Specification per node	Node quantity
Red Hat OpenStack Platform director	QuantaGrid D52B-1U	- CPU: 2x Intel® Xeon® Gold 6152 - RAM: 192 GB - Storage: 2x 480G S3520 SSD with Raid 1 - NIC: 1x 25 GbE dual ports	1
Controller	QuantaGrid D52B-1U	- CPU: 2x Intel® Xeon® Gold 6152 - RAM: 192 GB - Storage: 2x 480G S4600 SSD with Raid 1 - NIC: 4x 25 GbE dual ports	3
DPDK/SR-IOV Compute	QuantaGrid D52B-1U	- CPU: 2x Intel® Xeon® Gold 6152 - RAM: 384 GB - Storage: 2x 480G S4600 SSD with Raid 1 - NIC: 4x 25 GbE dual ports	3
Storage	QuantaGrid D51PH-1ULH	- CPU: 2x Intel® Xeon® E5-2620 v4 - RAM: 128 GB - Storage: 12x 6TB Raw Capacity 4x 480G S4600 SSD - NIC: 1x 25 GbE dual ports	3
Management Switch	QuantaMesh T3048-LY9	- 48 100/1000/10GBASE-T - 6 QSFP+ ports	1
ToR Switch	QuantaMesh T4048-IX2	- 48 10/25GbE SFP28 - 8 QSFP28 ports	2

4.2. Network planning

A well-designed network topology ensures the efficiency, correctness, and availability of the communication among running services. Red Hat OpenStack Platform uses Neutron networking service to manage the software-based networks, static and floating IP addresses, and the DHCP service. The OpenStack services are mapped to separate network traffic types assigned with various network subnets. In order to optimize the network performance to fulfill the strict requirements of CSPs, the network topology of the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution is designed according to the Red Hat OpenStack Platform 10 Network Functions Virtualization Planning Guide[9].

In the following section, we will present the default network configurations in the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution, including network types and subnet assignment, an overview of logical network topology, and logical layout respectively for SR-IOV and DPDK compute nodes.

4.2.1. Network types and subnet assignment

The network types, VLAN IDs, and subnets used in the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution are shown in Table 7.

Table 7. Network types and subnet assignment in overcloud

Network Type	VLAN ID	Subnet Details	Description
IPMI management	1250	10.102.50.0/24	IPMI management network provides access to BMC which allows system administrators to perform out-of-band management and monitoring for hardware.
External network	1252	10.102.52.0/24	External network not only hosts the OpenStack Dashboard (Horizon) for graphical system management but also handles the public API for OpenStack services. Moreover, this network performs SNAT service, which provides the external access for the running VMs.
Provisioning network	1251	10.102.51.0/24	Provisioning network handles the deployment of the overcloud nodes over PXE boot and orchestrates the installation of the overcloud environment on the bare metal servers. This network is predefined before the installation of the Red Hat OpenStack Platform director.
Internal API network	201	172.16.0.0/24	Internal API network is used for the communication between the OpenStack services using API communication, RPC messages, and database communication.
Tenant network	202	172.17.0.0/24	Neutron service allows each cloud user (tenant) to manage their own network environment using either VLAN segregation or tunneling.

[9] *Network Functions Virtualization Planning Guide:* https://access.redhat.com/documentation/en-us/red_hat_openstack_platform/10/paged/network_functions_virtualization_planning_guide/index

Storage network	203	172.18.0.0/24	Storage network handles in and out traffics of the storage service, including block storage, NFS, and iSCSI. Ideally, storage network can be isolated to a dedicated network interface for performance optimization.
Storage management network	204	172.19.0.0/24	Storage management network is used to synchronize data between replica nodes as well as to handle the proxy traffic between user requests and the underlying storage.
SR-IOV provider network	1253 1254 1255 1256	10.102.53.0/24 10.102.54.0/24 10.102.55.0/24 10.102.56.0/24	SR-IOV provider network allows virtual network ports to be attached to VFs, and handles in and out traffics of the VM dataplane directly through SR-IOV NICs.
DPDK provider network	206 207 208 209	172.26.0.0/24 172.27.0.0/24 172.28.0.0/24 172.29.0.0/24	DPDK provider network handles in and out traffics of the VM dataplane through DPDK-accelerated Open vSwitch.

4.2.2. Logical network topology

The Red Hat OpenStack Platform director provides multiple node types, including controller, SR-IOV compute, DPDK compute, and Red Hat Ceph Storage, to build the NFVI overcloud environment. Different services are assigned to each node for easy management and optimal system resource utilization. Likewise, network types are attached to the overcloud nodes based on the assigned services.

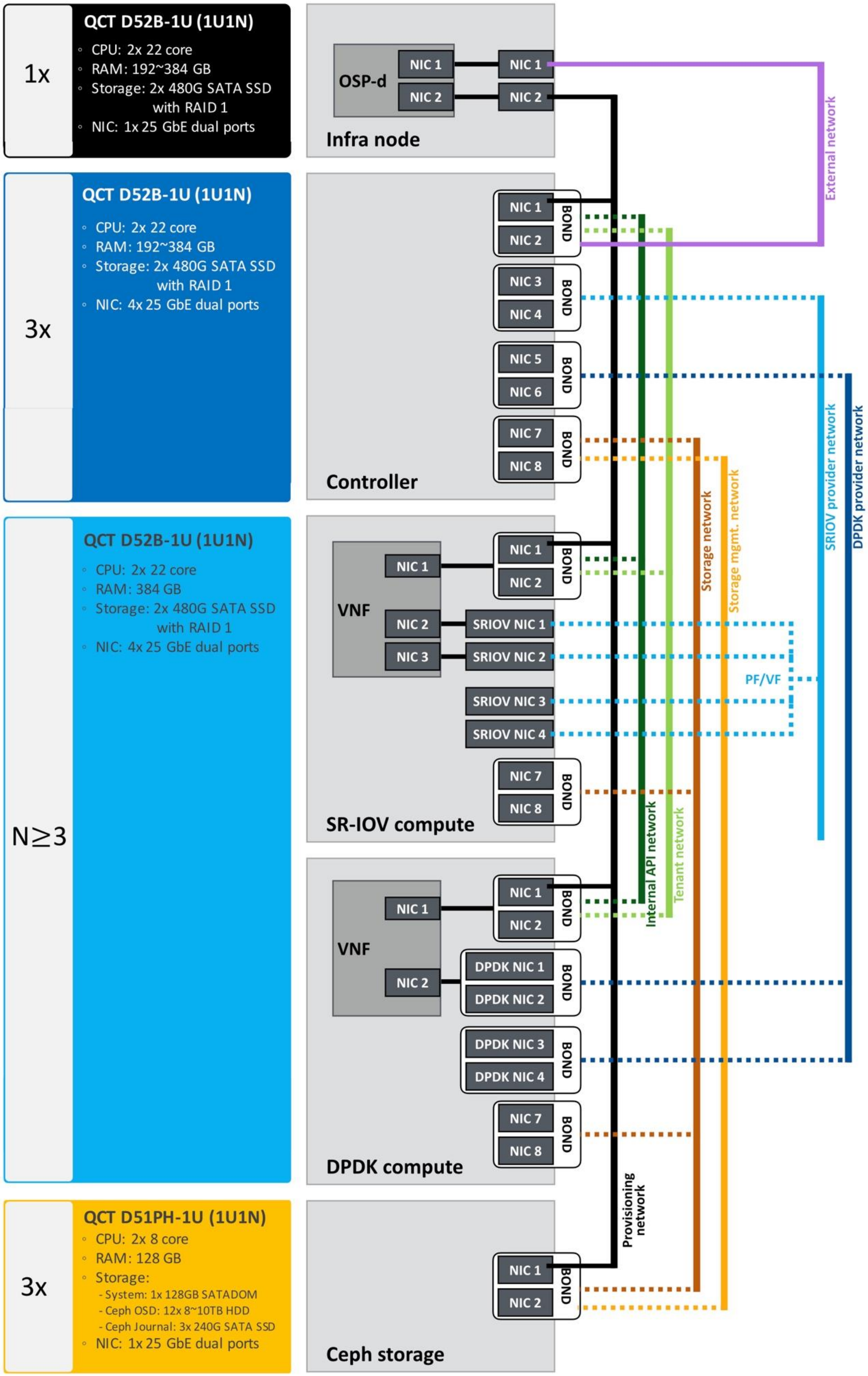

Figure 5. Network layout in overcloud

4.2.3. Logical layout for SR-IOV and DPDK compute nodes

The QxStack NFV Infrastructure with Red Hat OpenStack Platform solution supports two kinds of compute nodes, SR-IOV compute node and DPDK compute node. Figure 6 shows the logical layout of SR-IOV compute node. All the SR-IOV compute nodes are equipped with four dual-port 25G NICs, distributed among two NUMA nodes. One port from each NIC is assigned to provide SR-IOV VFs for the running VMs. With QCT selected NUMA-balanced hardware design, VMs can always access the NIC that is local to the vCPUs to achieve better performance. In order to provide high availability on physical NICs, cross-NIC bonds are designed respectively for management traffics and storage traffics. However, under the limit of hardware device passthrough, VMs attached to SR-IOV provider network should provide vNIC bonding to achieve vNIC high availability if required.

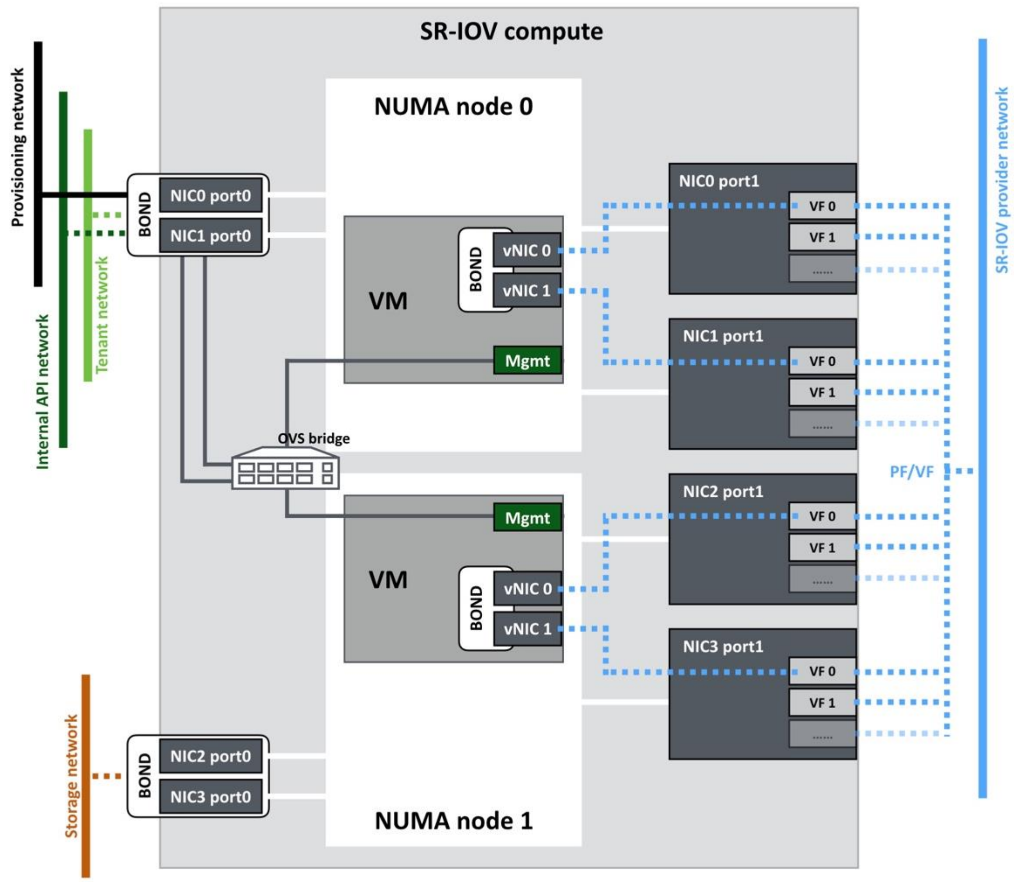

Figure 6. Network layout in SR-IOV compute node

Figure 7 shows the logical layout of DPDK compute node. Like SR-IOV compute nodes, all DPDK compute nodes are equipped with four dual-port 25G NICs, distributed between two NUMA nodes. In order to provide high availability on physical NICs, cross-NIC bonds are designed respectively for management traffics, storage traffics, and DPDK provider networks. With QCT selected NUMA-balanced hardware design, VMs can always access the NIC that is local to the vCPUs to achieve better performance.

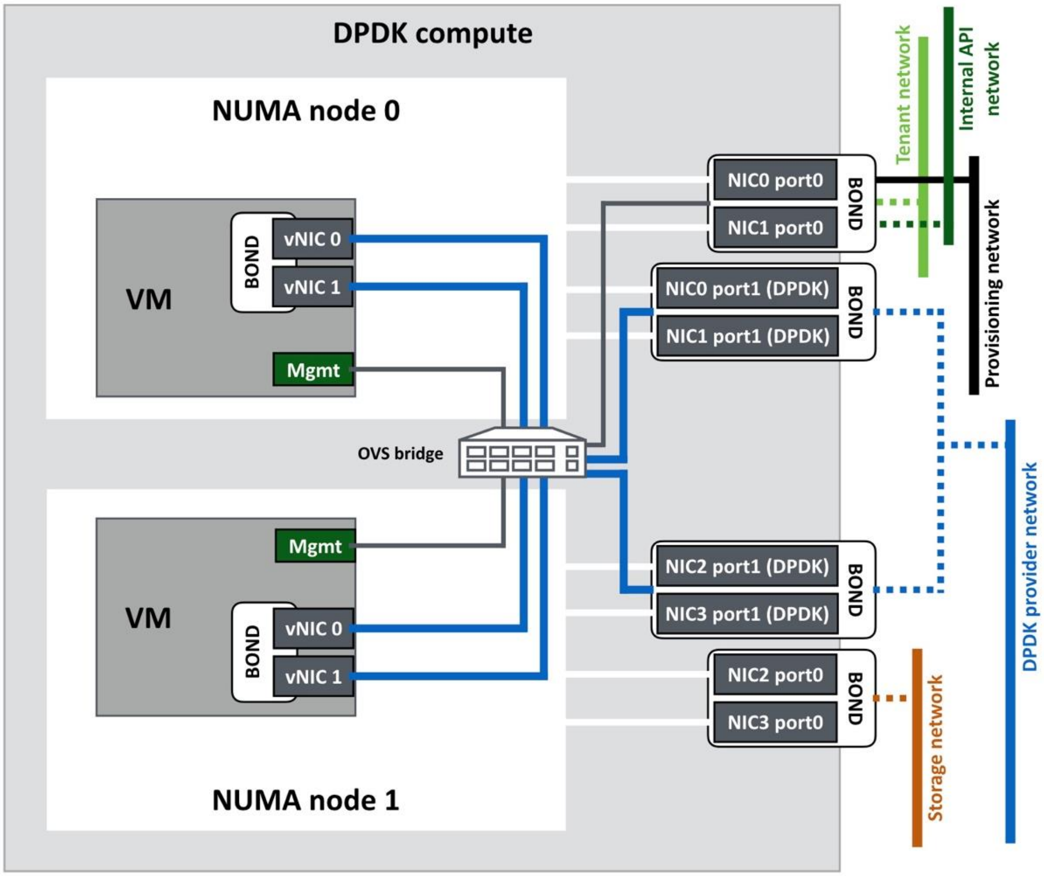

Figure 7. Network layout in DPDK compute node

4.3. Physical cabling and network diagram

In order to guarantee the network availability and maximize the network throughput, the top-of-rack (ToR) switch networking architecture is implemented in the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution, as shown in Figure 8. From the perspective of the server, two cross-NIC network ports are bonded as an interface except SR-IOV ports; on the other hand, link aggregation technology is enabled in the switch operating system to aggregate the ports from separate switches for HA purpose. The diagram below shows the logical wiring of our solution. Ports in a bonded interface are separately connected to different switches where each is tagged with its corresponding VLAN IDs.

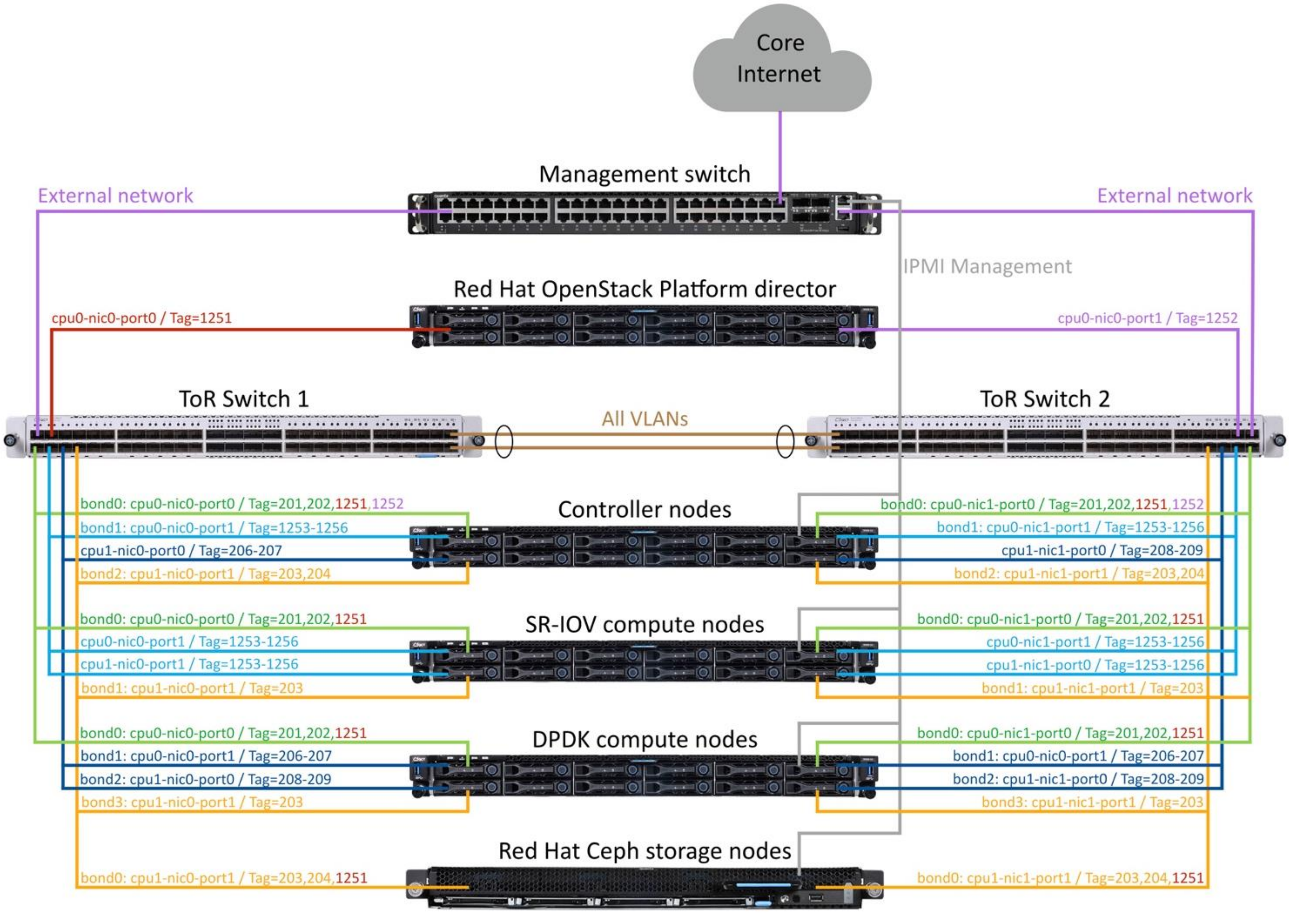

Figure 8. Physical cabling arrangement

5. Solution Performance

5.1. OPNFV test ecosystem overview

Open Platform for NFV (OPNFV)[10] is an open source project under the Linux Foundation, aimed at facilitating the cross-community integration with system-level deployment and testing. In order to demonstrate a comprehensive performance evaluation of the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution, QCT leverages OPNFV test ecosystem for solution validation. OPNFV test ecosystem includes several testing projects which cover functionality tests, performance measurements, stress tests, benchmarking as a service as well as compliance verification. By integrating NFV-related upstream projects, OPNFV delivers a comprehensive test framework with numerous test scenarios.

Figure 9 shows the test topologies of dataplane performance evaluation including east-west (VM-to-VM) traffic and north-south (Phy-VM-Phy) traffic. The east-west traffic tests evaluate the dataplane performance of underlying infrastructure from the perspective of VMs running on the virtual infrastructure manager (VIM) platform while the north-south traffic evaluates the network performance between the infrastructure and the rest of the network. In this paper, QCT leverages OPNFV Yardstick[11] and NFVbench[12] to demonstrate the solution performance respectively for east-west and north-south traffics.

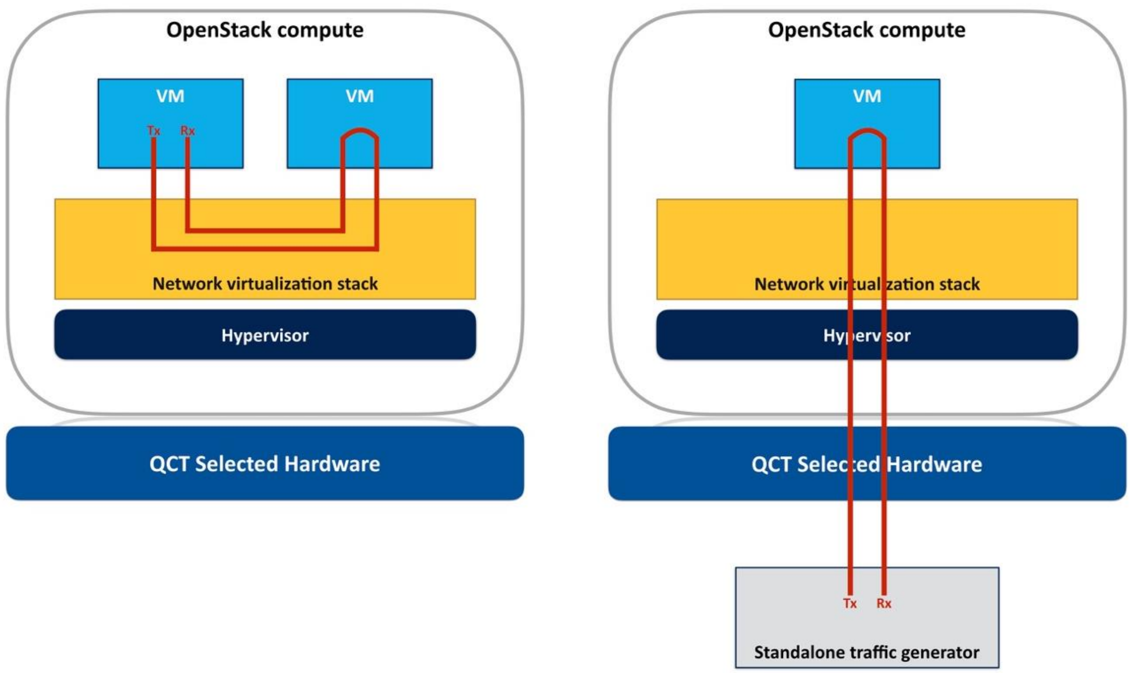

Figure 9. Test topology comparison between east-west and south-north traffic

[10] *Open Platform for NFV (OPNFV): https://www.opnfv.org/*
[11] *OPNFV Yardstick project: https://www.opnfv.org/community/projects/yardstick*
[12] *OPNFV NFVbench project: http://docs.opnfv.org/en/latest/submodules/nfvbench/docs/testing/user/userguide/index.html*

5.2. YardStick test cases and performance results

As a testing project sponsored by the Linux Foundation, Yardstick implements system-level validation aligned with the European Telecommunications Standards Institute (ETSI) TST 001 specification[13] to verify the underlying NFV infrastructure. To accommodate a variety of NFV use cases, Yardstick test cases decompose typical workload performance metrics into several characteristics and performance vectors. From the numerous test cases in Yardstick, QCT selected the test cases including throughput and latency to measure network performance. We not only customize a test suite composed of a set of test metrics and corresponding test cases in Yardstick but also verify the OpenStack compute nodes with EPA features and data plane enhancement in terms of Generic, DPDK, and SR-IOV. The performance testing is evaluated by the customized test suite running on the optimized NFV infrastructure based on Yardstick framework.

5.2.1. East-west traffic test methodology

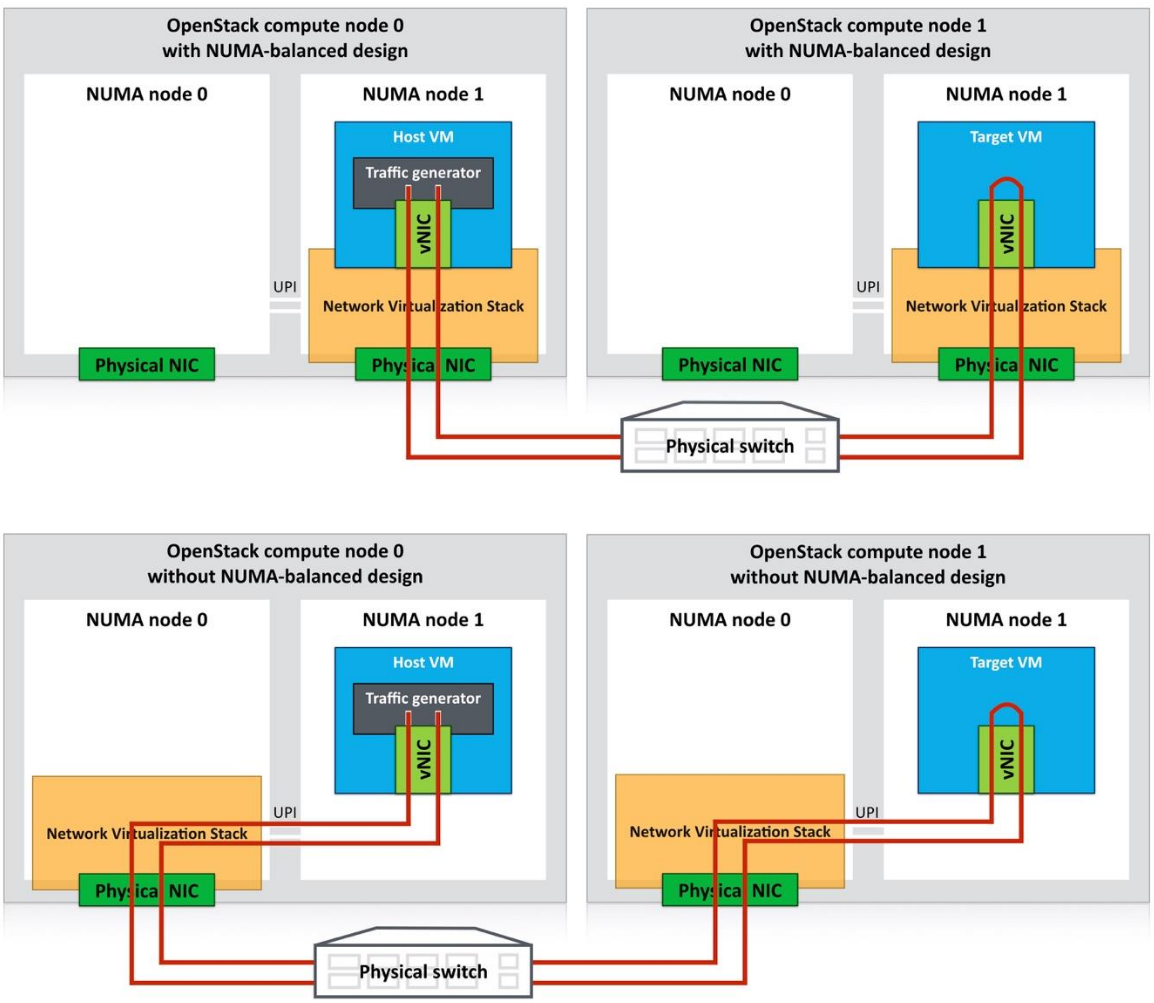

Figure 10. East-west traffic test topology with and without NUMA-balanced hardware design

---

[13] *ETSI TST 001 specification:*
http://www.etsi.org/deliver/etsi_gs/NFV-TST/001_099/001/01.01.01_60/gs_nfv-tst001v010101p.pdf

Yardstick runs a test case or a test suite with access and credentials to the underlying NFV infrastructure, in this case, the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution. Each test case is described by a YAML configuration file and a test suite is composed of a set of test cases. Yardstick converts the input configuration into a Heat template and deploys the template into a stack on the underlying OpenStack. After both host and target VMs are launched, Yardstick activates each measurement tool and runs the the corresponding commands with specific measurement types (e.g., Pktgen, Ping, etc.) in the VM via SSH. The output results are collected and can be dispatched to a JSON file, the influxDB, or the HTTP server for further query. Finally, Yardstick undeploys the Heat stack after the test is finished.

Figure 10 shows the east-west traffic test topology with and without NUMA-balanced hardware design. A NUMA-balanced hardware design supports local memory access and distributes NICs across CPUs and sockets. With this design, running VMs can always use vCPUs, memory, and NICs on the same local socket to provide consistent, high performance. In contrast, VMs running on a non-NUMA-balanced hardware design might not be able to use the system resources from the same local socket such as NICs and memory.

5.2.2. Performance evaluation

The following section shows the detailed results of east-west traffic performance evaluation based on OPNFV Yardstick project.

Comparison between network virtualization stacks

Based on the aforementioned test topology, Figure 11 shows the throughput and latency performance comparisons among generic, DPDK, and SR-IOV network virtualization stacks. The test results demonstrate that both DPDK and SR-IOV can significantly improve network performance regardless of the packet size. For the throughput performance, generic network virtualization stack shows relatively low performance while SR-IOV is able to get near line-rate performance in large packet size. On the other hand, the results of latency performance show that both DPDK and SR-IOV are comparatively stable and are able to reduce up to 80% of end-to-end latency compared to generic stack.

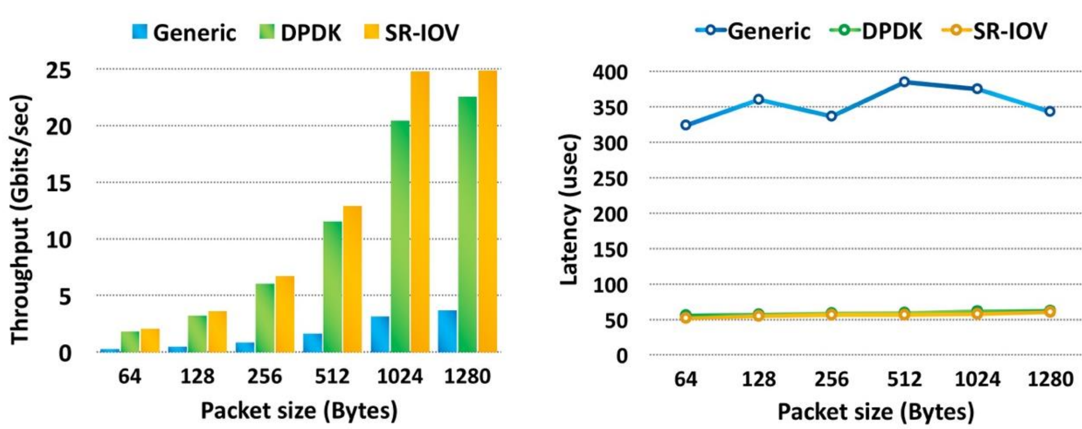

Figure 11. Network performance of generic, DPDK, and SR-IOV

Comparison between VMs running with/without EPA features based on DPDK

Figure 12 shows the throughput and latency performance comparisons between DPDK with and without EPA features enabled. The "DPDK without EPA features" scenario is tested under the aforementioned topology without NUMA-balanced hardware design while the VM is launched with default CPU shared policy. On the other hand, the "DPDK with EPA features" scenario is tested under the topology with NUMA-balanced hardware design while the VM is launched with CPU pinning applied. The test results demonstrate that with EPA features enabled, both throughput and latency performances are improved by 3% to 15% regardless of the packet size.

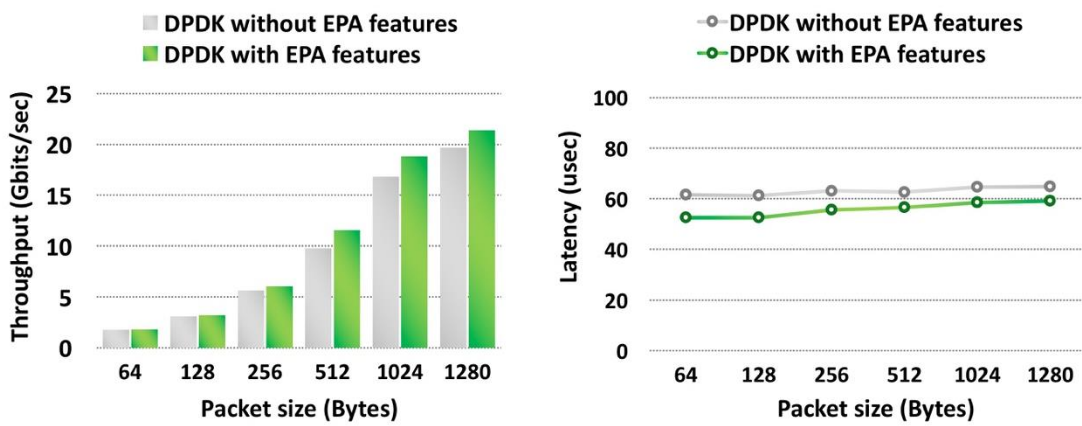

Figure 12. Network performance of DPDK with/without EPA features

Comparison between VMs running with/without EPA features based on SR-IOV

Figure 13 shows the throughput and latency performance comparisons between SR-IOV with and without EPA features enabled. As mentioned in section 3.2.1, SR-IOV is a specification that allows physical PCI hardware devices to create multiple VFs to be shared between VMs. VM should reside in the CPU socket which is local to the VFs for the direct access to hardware devices. Therefore, both "SR-IOV without EPA features" and "SR-IOV with EPA features" scenarios are tested under topology with NUMA-balanced hardware design. The former launches VMs with default CPU shared policy and the latter launches VMs with CPU pinning applied. With low system load and only 1 CPU core used for traffic generation and transmission, we can barely get the performance difference with CPU pinning policy applied. Nevertheless, the results of the throughput and latency performance respectively improved by 1% and 5% to 9% can still be provided as a reference.

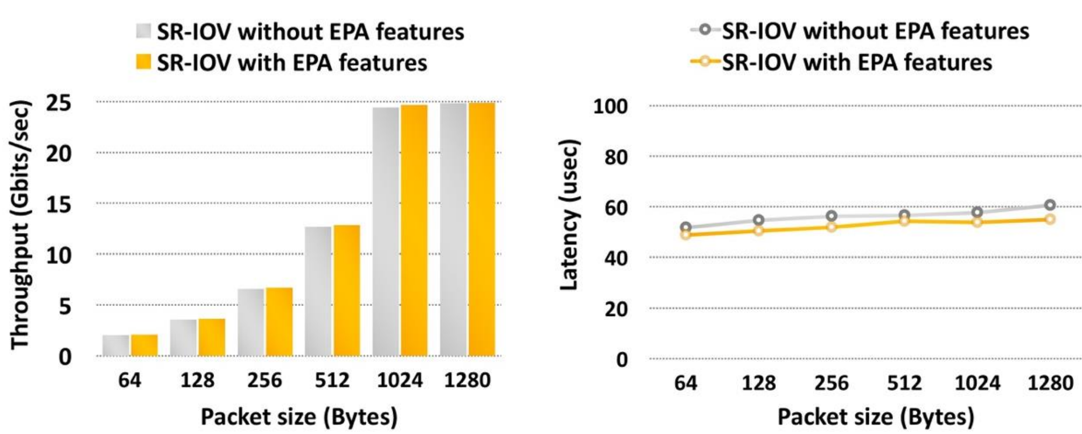

Figure 13. Network performance of SR-IOV with/without EPA features

5.3. NFVbench test cases and performance results

NFVbench is a new OPNFV testing project in the Euphrates release. The main goal of NFVbench is to provide a representative data plane benchmark by developing a measurement toolkit for the assessment of a fully OpenStack-based NFVI production. In the Euphrates release, NFVbench not only supports both DPDK and SR-IOV enhanced techniques on a NFVI solution stack but also provides three types of measurements such as a fixed rate with detailed parameters, the highest throughput with no drop rate, and with a maximum-allowed drop rate. Both L2/L3 forwarder and pre-built Phy-VM-Phy/Phy-VM-VM-Phy packet paths for service chains are well configured in the NFVbench, and the test process is automatically launched via standard OpenStack APIs. To validate network performance of the north-south traffics, we adopt the test scenario of a Phy-VM-Phy service chain with a L2 forwarder on the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution.

5.3.1. North-south traffic test methodology

NFVbench runs a test on a standalone server and interacts with NFVI solution stack via OpenStack credentials and APIs. Each test case is described by a YAML configuration file and a customized traffic profile is configured with various packet sizes. All the tests in NFVbench are measured through the integrated TRex generator which follows a well-defined subset of the RFC-2544 standard and the most representative of NFV traffics. After the target VM is successfully launched on an OpenStack stack, TRex generator can generates the network traffics and the NFVbench is able to assess the performance results. The output results are collected and stored in a JSON file. Finally, NFVbench can automatically clean up the testing resources deployed on the OpenStack environment after the test is finished.

Figure 14 shows the north-south traffic test topology with and without NUMA-balanced hardware design. With NUMA-balanced hardware design, running VMs can always use vCPUs, memory, and NICs on the same local socket to provide consistent, high performance.

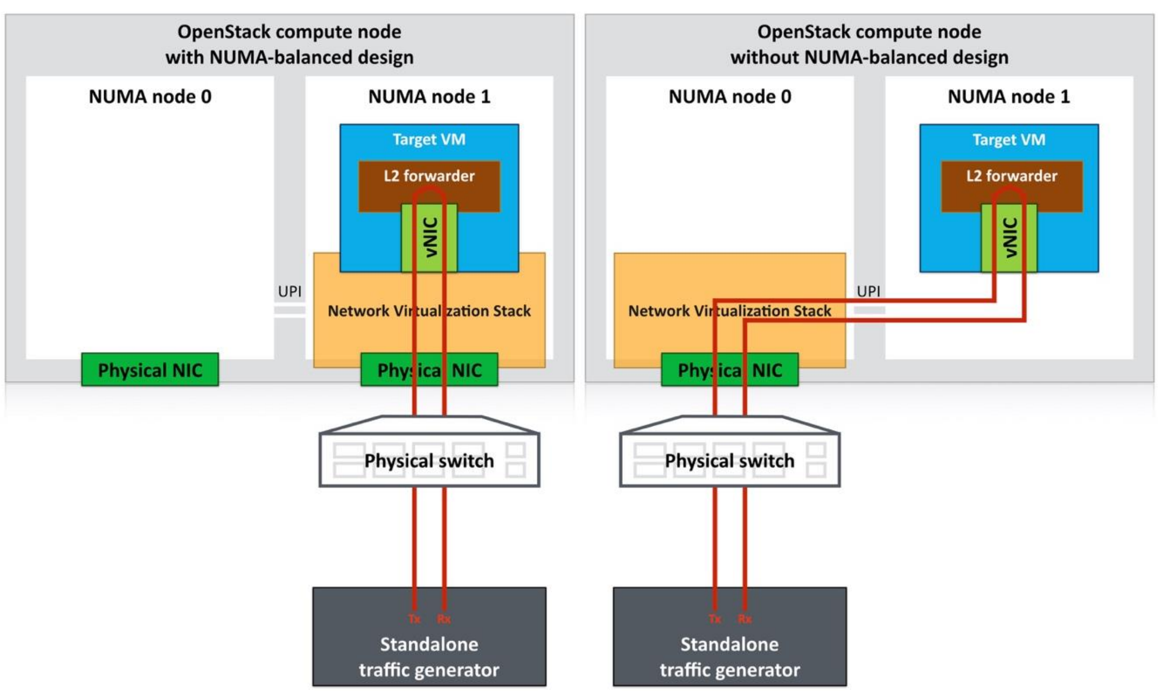

Figure 14. Phy-VM-Phy test topology with and without NUMA-balanced hardware design

5.3.2. Performance evaluation

The following section shows the detailed results of north-south traffic performance evaluation based on OPNFV NFVbench project.

Comparison between network virtualization stacks

Based on the aforementioned north-south test topology, Figure 15 shows the throughput and latency performance comparison between DPDK and SR-IOV network virtualization stacks. For the throughput performance, SR-IOV is able to get near line-rate performance in the packet size 128 bytes or above while DPDK is able to achieve 80% of line-rate performance in the packet size 1280 bytes. Moreover, the results of latency performance reveal that DPDK can significantly reduce 90% of the end-to-end latency compared to SR-IOV.

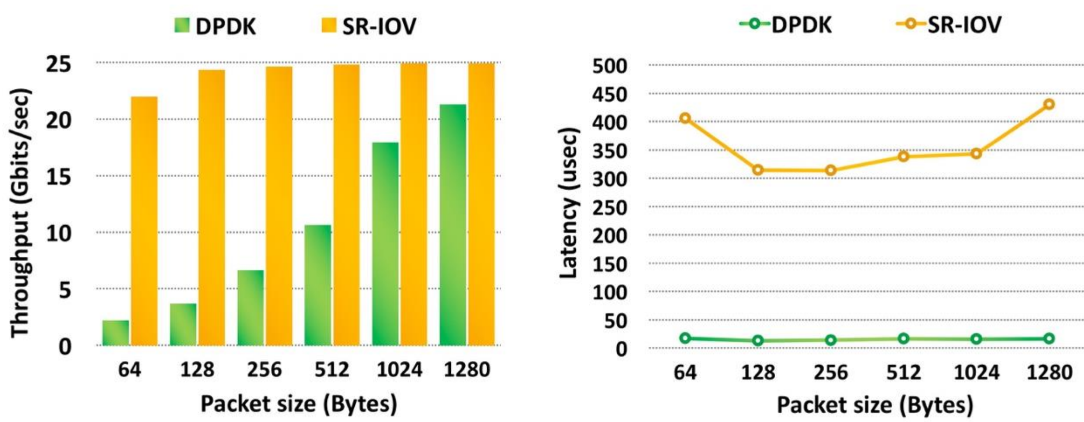

Figure 15. Network performance of generic, DPDK, and SR-IOV

Comparison between VMs running with/without EPA features based on DPDK

Figure 16 shows the throughput and latency performance comparisons between DPDK with and without EPA features enabled. Like the east-west traffic tests, the "DPDK without EPA features" scenario launches VMs with default CPU shared policy and without NUMA-balanced hardware design while the "DPDK with EPA features" scenario launches VMs with both CPU pinning applied and NUMA-balanced hardware design. The test results demonstrate that with EPA features enabled, the end-to-end latency performance is improved up to 90% and the throughput performance is improved 40%.

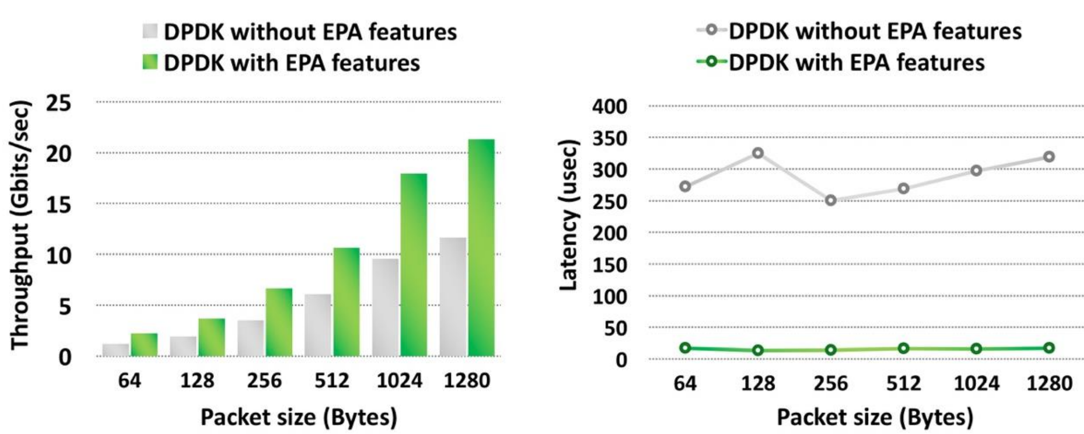

Figure 16. Network performance of DPDK with/without EPA features

Comparison between VMs running with/without EPA features based on SR-IOV

Figure 17 shows the throughput and latency performance comparisons between on SR-IOV with and without EPA features enabled. Like the east-west traffic tests, both “SR-IOV without EPA features” and “SR-IOV with EPA features” scenarios are tested with NUMA-balanced hardware design. The former launches VMs with default CPU shared policy and the latter launches VMs with CPU pinning applied. The test results demonstrate that with EPA features enabled, both throughput and latency performances are significantly improved 30% to 60%, and the end-to-end performance is comparatively stable when CPU pinning policy is applied.

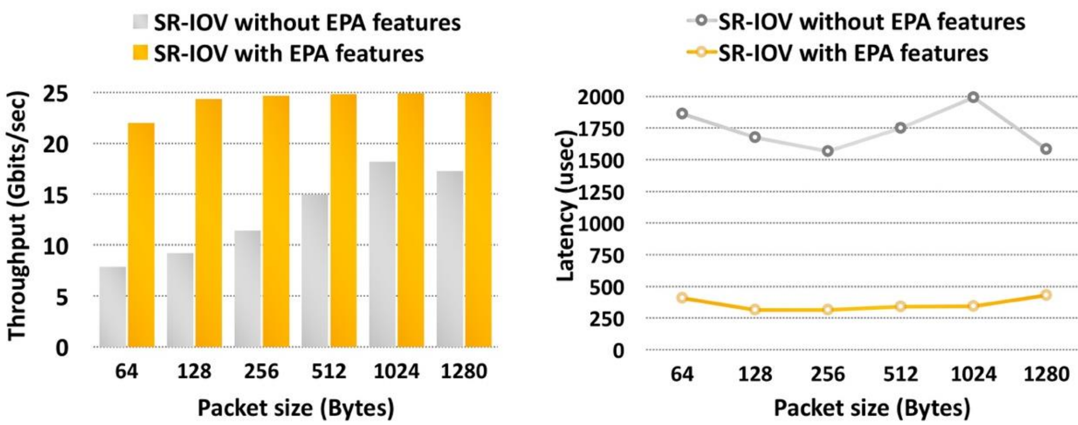

Figure 17. Network performance of SR-IOV with/without EPA features

6. Conclusion

The need to digitally transform the legacy network infrastructure has never been greater than it is today for CSPs. The QxStack NFV Infrastructure with Red Hat OpenStack Platform solution provides a flexible, reliable, high-performance NFV foundation to realize the cloud transformation strategy.

The building block pre-integrated with QCT NUMA-balanced hardware and Red Hat open source software brings the benefits of agility and flexibility. To maximize resource for CSPs, a well-designed QxStack auto-deployment tool is equipped with NFVI foundations for business expansion.

This reference architecture recommends the network planning with EPA features for your environment to ensure your network performance and high availability. With the comprehensive validation based on industry standard OPNFV test framework, CSPs could rely on the solution and build up a NFV infrastructure on the fly.

To learn more about the QxStack NFV Infrastructure with Red Hat OpenStack Platform solution, please visit www.qct.io/q/NFVI.

ABOUT QCT

QCT (Quanta Cloud Technology) is a global datacenter solution provider extending the power of hyperscale datacenter design in standard and open SKUs to all datacenter customers.

Product lines include servers, storage, network switches, integrated rack systems and cloud solutions, all delivering hyperscale efficiency, scalability, reliability, manageability, serviceability and optimized performance for each workload.

QCT offers a full spectrum of datacenter products and services from engineering, integration and optimization to global supply chain support, all under one roof.

The parent of QCT is Quanta Computer Inc., a Fortune Global 500 technology engineering and manufacturing company.

http://www.QCT.io

UNITED STATES

QCT LLC., Silicon Valley office
1010 Rincon Circle, San Jose, CA 95131
TOLL-FREE: 1-855-QCT-MUST
TEL: +1-510-270-6111
FAX: +1-510-270-6161
Support: +1-510-270-6216

QCT LLC., Seattle office
13810 SE Eastgate Way, Suite 190, Building 1,
Bellevue, WA 98005
TEL: +1-425-633-1620
FAX: +1-425-633-1621

CHINA

云达科技, 北京办公室（Quanta Cloud Technology）
北京市朝阳区东大桥路 12 号润诚中心 2 号楼
TEL +86-10-5920-7600
FAX +86-10-5981-7958

云达科技, 杭州办公室（Quanta Cloud Technology）
浙江省杭州市西湖区古墩路浙商财富中心 4 号楼 303 室
TEL +86-571-2819-8650

JAPAN

Quanta Cloud Technology Japan 株式会社
東京都港区芝大門 2-5-8 芝大門牧田ビル 3F, 105-0012
TEL +81-3-5777-0818
FAX +81-3-5777-0819

GERMANY

Quanta Cloud Technology Germany GmbH
Hamborner Str. 55, 40472 Düsseldorf
TEL +492405-4083-1

TAIWAN

雲達科技（Quanta Cloud Technology）
桃園市龜山區文化二路 211 號 1 樓
1F, No. 211 Wenhua 2nd Rd., Guishan Dist., Taoyuan City 33377, Taiwan