TECHNICAL BRIEF

QCT Validated Red Hat OpenShift Platform with Enhanced Platform Awareness and Optimized Base Frequency

An agile, flexible, and high-throughput performance platform to accelerate network transformation

Introduction

Communications Services Providers (CoSPs) are facing challenges with the growing demands of high-speed networks and increasing Capital Expenditure (CapEx). Novel innovative technologies are indispensable for CoSPs to drastically improve end-user experiences and trigger capital reduction. Containers are penetrating the virtualization environment since their lightweight and standalone software packages implement the cloud-native benefits on Network Functions Virtualization (NFV). Containers enable cloud-native technology into NFV infrastructure with enhanced performance, portability, high interoperability, and agility in workload orchestration. As the world moves towards 5G deployments, cloud-native technology plays a vital role, by accelerating network transformation and supercharging the success of 5G networks with cloud-native NFV enablement.

Key Highlights of QCT Validated OpenShift Platform with EPA:

- *High interoperability and agility in workload orchestration.*
- *Resource isolation in computing.*
- *High throughput performance in data-plane network.*

To strengthen 5G implementation, a network infrastructure with well-defined architecture is required to support containers, address cloud-native principles, and to allow application services to run independently with portable distinctive features. Quanta Cloud Technology (QCT) teamed up with Intel to deliver a QCT validated Red Hat OpenShift Platform applying Enhanced Platform Awareness (EPA) features as well as Intel® Speed Select Technology - Base Frequency (Intel® SST-BF) technology (hereinafter called "the platform"). The platform architecture features enhanced data-plane networking, resource isolation, and hardware capability discovery, all of which are needed for Network Function Virtualization (NFV) workloads to achieve promising deterministic performance.

Overview of QCT Validated Cloud Native Platform for Telco

QCT constructed the cloud native platform with outstanding network performance by selecting QCT hyperscale hardware with high density and extensibility based on Intel® Xeon® Gold processors, and by integrating with Red Hat OpenShift Container Platform, to deliver a fully containerized environment for the telecom operator using cloud native software to build out the next generation network. This QCT cloud native platform provides an auto-scaling feature, self-managing platform, and high-speed network for modern datacenters, as well as support for EPA features and Intel® SST-BF technology. The platform can be adopted to accelerate CNF applications with fast service deployment. It also provides the flexibility for the Telco operator to fulfill innovative services for different verticals and supports the Telco to create unprecedented benefits to the users in the 5G era.

Container technology is essential for NFV in modern networks; Kubernetes is becoming the de-facto standard for managing container base networks. Red Hat® OpenShift®, a container platform built on Kubernetes with full-stack automated operations, unites developers and operations on a single platform to consistently build, deploy, and manage applications across on premise, hybrid-cloud and multi-cloud infrastructures. The platform supports data plane and control plane acceleration to ensure capacity of each isolated network, which helps provide a secure and stable environment to fulfill the needs of Telcos in supplying massive services. Customers using the QCT optimized and validated Red Hat OpenShift platform with acceleration functions can deliver modern and legacy applications with efficient development cycles and reduced operating cost.

QCT Validated Red Hat OpenShift Platform Architecture

QCT QuantaGrid D52B-1U is a server model recommended for OpenShift master nodes and QCT QuantaGrid D52BQ-2U for worker nodes, as shown in Figure 1. The master nodes are equipped with Intel 10GbE OCP network mezzanine cards (NMCs) while the worker nodes are equipped with 25GbE network interface cards (NICs) in which the OpenShift worker nodes are configured to support both

Data Plane Development Kit (DPDK)[1] and Single-Root I/O Virtualization (SR-IOV) technologies. QCT QuantaMesh T4048-IX8D and QuantaMesh T1048-LY4R switches are recommended for data and management switching, respectively. For the detailed hardware configuration, please refer to Appendix A.

QCT partnered with Intel to deliver an OpenShift platform that supports several EPA features, including memory huge pages, CPU Manager for Kubernetes (CMK), Node Feature Discovery (NFD), and DPDK and SR-IOV network adapters. With EPA features enabled, this platform provides resource isolation in computing and achieves high throughput data-plane network performance. The platform supports Rook Ceph Storage to provide object, block, and file storage as the backend storage, which enables the S3-based image registry, persistent volume, and shared file system for containers. For the detailed software configuration, please refer to Appendix B.

Hardware infrastructure includes:

- *1x Deployer node.*
- *3x OpenShift Master nodes.*
- *3x OpenShift Worker nodes.*

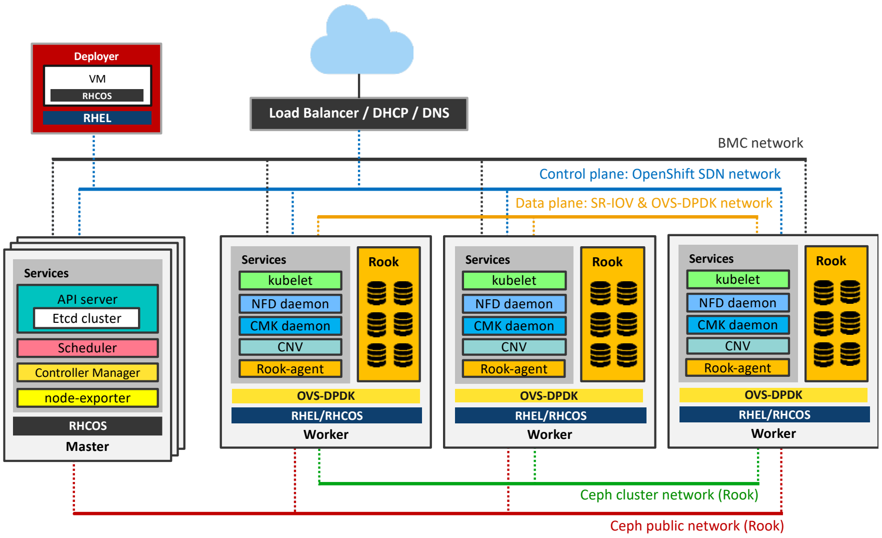

Figure 1. QCT validated OpenShift platform architecture.

Intel EPA in Red Hat OpenShift Platform

To address explosive data growth and complex NFV workloads, the platform features a resilient, scalable architecture with high availability and supports Enhanced Platform Awareness (EPA) to improve the performance and determinism for containerized workloads. Compute overhead can be minimized and network performance can be maximized by utilizing the following key EPA capabilities:

CPU Manager for Kubernetes (CMK) is a command-line program that enables CPU core pinning and isolation for containerized workloads. When a CMK cluster is launched, CMK pods can be spun up on each Kubernetes node to continuously collect CPU information, including topology, resource availability, and resource allocation status. With the collected information, CMK provides capabilities to affinitize and allocate CPU resources to workloads according to the scenario’s demands.

Node Feature Discovery (NFD) is a project that enables hardware capability discovery on OpenShift nodes. When an NFD DaemonSet is launched, NFD pods will be initialized on each node to detect the hardware capabilities such as CPU, memory, network, and kernel features. With the detected information, NFD pods will publish labels attached to the host and advertise the hosts’ capabilities to end-users.

Multus Container Network Interface (Multus CNI) is a plugin that enables multiple network interfaces to be attached to pods. The OpenShift platform supports a variety of CNI plugins to fulfill different needs, such as local-host, flannel, and calico. Typically, only one network interface can be attached to a container pod on the OpenShift platform. With Multus CNI, users can create multi-homed pods with multiple interfaces, which facilitates the separation of control and data network planes, as shown in Figure 2. QCT uses the OpenShift SDN for control plane traffic, the SR-IOV Network Device Plugin (SRIOV-DP)

---

[1] Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

for north-south data plane traffic, and the Userspace Container Network Interface (Userspace CNI) for east-west data plane traffic.

> *"To accelerate data-plane packet processing, the QCT validated OpenShift Platform supports both SR-IOV and OVS-DPDK in worker nodes with memory huge pages."*

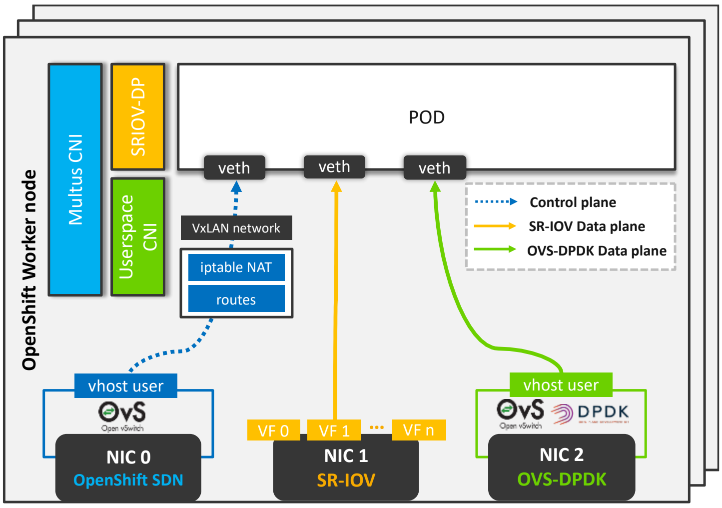

Figure 2. QCT validated OpenShift platform network architecture.

Signal Root I/O Virtualization (SR-IOV) is a specification that allows physical PCI devices to be shared between multiple virtual hosts for enhancing network performance. SR-IOV virtualizes PCI hardware devices to create multiple virtual functions (VFs) on top of physical functions (PFs), in which VFs are the lightweight functions that can be assigned to specific virtual hosts and PFs are the full-featured physical hardware ports. Network traffic flows through a direct-attached VF, which bypasses the software switching layer to achieve near line-rate performance.

SR-IOV Network Device Plugin (SRIOV-DP) is a plugin that enables SR-IOV device discovery and exposure on OpenShift nodes. When an SRIOV-DP DaemonSet is launched, a list of resource configurations will be defined for designated devices, including vendor information, driver information, and PCI device addresses. An SRIOV-DP pod is subsequently spun up on each OpenShift node to continuously collect the availability and allocation status of those devices and to assign PF or VF resources to workloads based on the demands.

Data Plane Development Kit (DPDK)[2] is a set of data plane libraries and user-space network drivers for packet processing acceleration. It significantly eliminates the typical interrupt-driven processing overhead and enables applications to process packets directly to and from the NIC. OVS-DPDK replaces the standard OVS kernel data path with a DPDK-based data path, creating a user-space Open vSwitch (OVS) for packet forwarding. OVS-DPDK efficiently improves network throughput and latency performance while remaining transparent in the overall architecture and exposing the same interfaces—including OpenFlow, Open vSwitch Database (OVSDB), and command lines—as the standard OVS implementation.

Userspace Container Network Interface (Userspace CNI) is a plugin that implements user-space networking on the Kubernetes platform. It enables DPDK-based container applications to run with either the OVS-DPDK or VPP network and significantly accelerates data plane performance for those applications. When a user-space network is attached to a pod, a virtual device is created and attached to the pod, and a network interface is created and configured on a virtual switch on host. In this case, DPDK-based applications can send traffic through the virtual device to the accelerated network stack, to achieve near-native performance.

Hugepages is a kernel feature that provides huge memory segments to reduce the number of pages for the requested memory size. Both SR-IOV and OVS-DPDK take the advantage of memory huge pages to reduce the potential for translation lookaside buffer (TLB) misses and to improve computational performance.

---

[2] Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

Intel® SST-BF in Red Hat OpenShift Platform

DPDK Poll Mode Driver (PMD) is time-consuming in terms of CPU cycles, owing to the constant I/O polling and immediate processing. The network performance impact varies depending on the CPU processing; thus, core isolation and core frequency boosting are extremely critical in accelerating packet forwarding. Intel® SST-BF allows CPUs to be deployed with a guaranteed asymmetric core frequency configuration. When Intel® SST-BF is disabled, all applications on a processor operate at the same core frequency. When Intel® SST-BF is enabled, the targeted applications run at a higher frequency and the remaining applications run at a lower frequency, as shown in Figure 3.

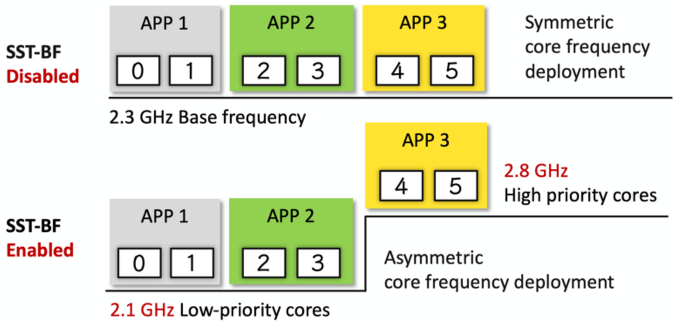

Figure 3. Intel SST-BF deployment.

To demonstrate the benefits of Intel® SST-BF enablement, QCT simulated an asymmetric deployment where the OpenShift worker node hosted high priority workloads and low priority workloads. The high priority workload hosted in OVS-DPDK PMD relies on deterministic compute cycles and benefits from frequency scaling. For the detailed BIOS settings for Intel® SST-BF enablement, refer to Appendix C.

RED HAT OPENSHIFT PLATFORM PERFORMANCE

This section depicts how the aforementioned technologies enhance network performance. To simulate the real-world use case, industry-standard software tools listed below are leveraged to simulate traffic generation and packet forwarding:

- TRex, an open-sourced traffic generator fueled by DPDK, is a stateful and stateless traffic generator, which is able to generate L4-7 multi-stream traffic.
- Testpmd is one of the DPDK applications that can be configured for packet forwarding.

QCT illustrated the necessity of enabling SRIOV-DP and OVS-DPDK data-plane acceleration technologies for north-south traffic and east-west traffic to achieve line-rate throughput performance. The two scenarios, single-node service chaining and cross-node service chaining, were conducted to demonstrate throughput performance improvement with EPA features and SST-BF enabled on the data plane network. The six test cases and the platform configurations for each case are summarized in Table 1.[3]

Table 1. Performance test scenarios.

Test Scenario		**Service Chaining**					
Deployment		**Single Node**			**Cross Node**		
Test Cases (Number of Pod)		**2 Pods**	**4 Pods**	**6 Pods**	**2 Pods (1/1 design)**	**4 Pods (2/2 design)**	**6 Pods (3/3 design)**
Traffic Generator		TRex			TRex		
User Mode Application		DPDK-Testpmd			DPDK-Testpmd		
Data-Plane Acceleration	**North-South Traffic**	SRIOV-DP			SRIOV-DP		
	East-West Traffic	OVS-DPDK			OVS-DPDK		
CPU Core Isolation and CMK		Applied			Applied		
Huge Pages		1GB per page size			1GB per page size		
Intel® SST-BF		Enabled/Disabled			Enabled/Disabled		
Packet Size		64, 128, 256, 512, 1024, 1280			64, 128, 256, 512, 1024, 1280		

---

[3] Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Single-Node Service Chaining

Service chaining is a key approach of NFV, which enables flexibility, dynamic provisioning, and high availability for VNF deployments. A large number of VNFs can be connected across networks in NFV infrastructure as a virtual chain to handle traffic flows. To fulfill the network requirement of throughput-sensitive workloads, the platform adopts SRIOV-DP passthrough technology for north-south traffic which flows to and from the end devices, and OVS-DPDK data plane acceleration for east-west traffic which flows across the chained services. With this design, the benefits from both SRIOV-DP and OVS-DPDK can be leveraged to achieve line-rate performance and to balance the traffic loads between SR-IOV NICs and CPU resources used by the OVS-DPDK PMD. The traffic generated by TRex flows through the SR-IOV VF of NIC 0 and directly passes through to pod 1. The traffic forwarding in each pod is hosted by the DPDK Testpmd application and the traffic redirection across pods on the host is defined by the Open vSwitch flow table. The traffic flow from the last pod directly passes through to the SR-IOV VF of NIC 1 and is sent back to the TRex server, as shown in Figure 4.

> *“Higher throughput performance shows improved platform capability. The platform is able to provide high throughput performance in service chaining scenario across 2 pod, 4 pods, and 6 pods.”*

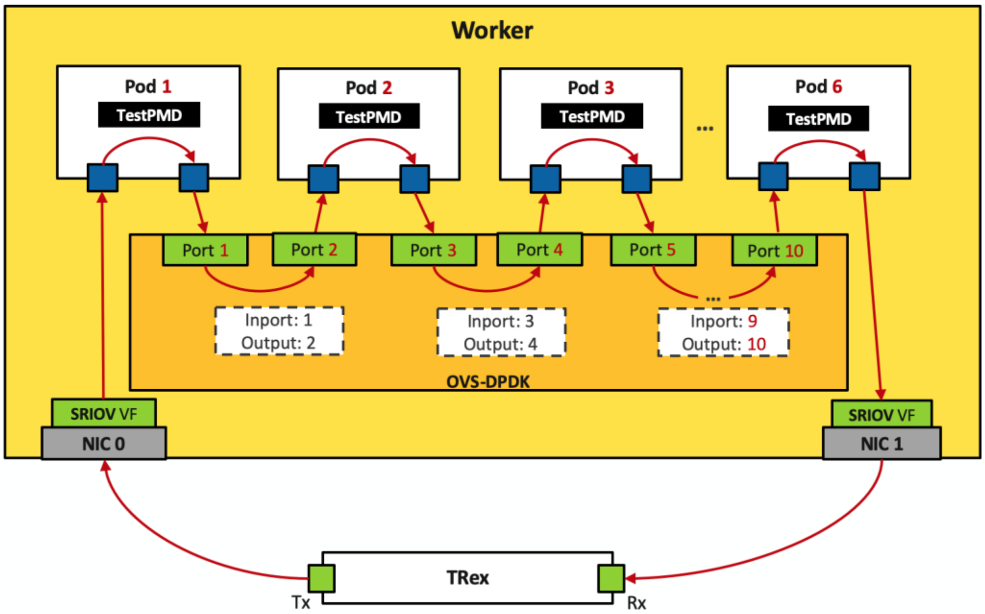

Figure 4. Network topology for single-node service chaining.

In the single-node scenario, the test cases include 2-pod, 4-pod, and 6-pod service chaining with both Intel® SST-BF enabled and disabled, respectively. The throughput performances in single-node service chaining are compared in million packets per second (Mpps) and gigabits per second (Gbps), as shown in Figure 5.[4]

The test results reveal that when the packet size is larger than 128 bytes, the line-rate performances of the platform either with Intel® SST-BF enabled or disabled are achieved in all number of pods. However, compared with the platform with Intel® SST-BF disabled, in 64 bytes, the throughput performance with Intel® SST-BF enabled is improved ranging from 18.5% to 21.3%; in 128 bytes, the throughput performance is improved ranging from 17.1% to 19.8%. In each test case, the maximum throughput is measured and the PMD cores on the host is 100%.

---

[4] Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

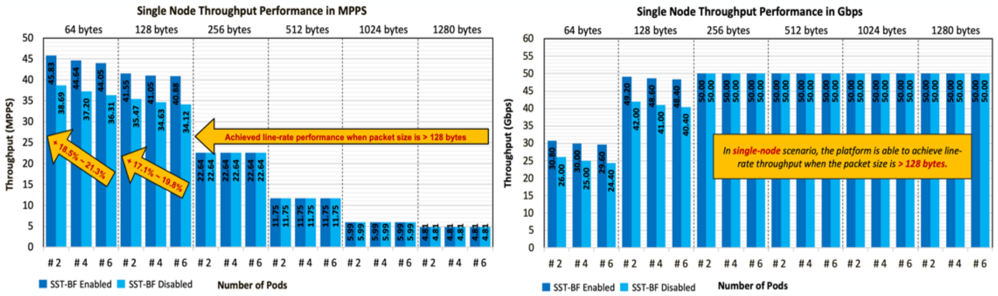

Figure 5. Throughput performance for single-node service chaining.

Cross-Node Service Chaining

Service stability is another challenge for running NFV workloads. The OpenShift platform can be adopted to manage distributed services or containerized applications across a distributed cluster of nodes in order to fully control Kubernetes. By adopting virtual extensible LAN (VXLAN), the platform can provide a mechanism to transfer L2 network packets across L3 networks for node-to-node communication. Hence, QCT simulated cross-node service chaining to demonstrate how the platform with Intel EPA and Intel® SST-BF enabled can achieve optimal performance.

The network topology in the cross-node scenario is similar to that of in the single-node scenario, except that in cross-node scenario, the additional bridge br-phy is built on the OpenShift worker node to connect the DPDK physical NIC with an assigned IP address and the added port dpdk0 on the bridge br-phy is bound to a specific PCI device when Open vSwitch is run in user space. The remote IP is assigned on the bridge br0 for the added port vx1 as destination host's IP address, which allows each Testpmd application on the two hosts to communicate over the VXLAN tunnel, as shown in Figure 6.

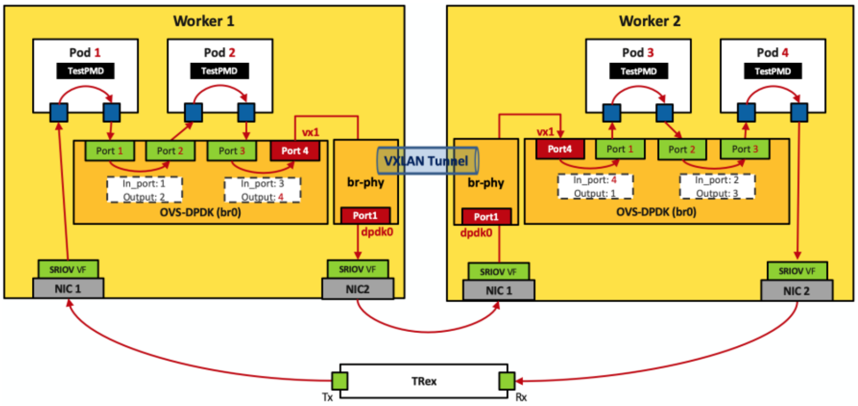

Figure 6. Network topology for cross-node service chaining.

In the cross-node scenario, Testpmd pods are evenly deployed on the two hosts. That is, in the 2-pod, 4-pod, and 6-pod test cases, 1 pod, 2 pods, and 3 pods are respectively allocated in each OpenShift worker node. The throughput performance data of the platform with either Intel® SST-BF enabled or disabled in cross-node service chaining are respectively compared in both million packets per second (Mpps) and gigabits per second (Gbps), as shown in Figure 7.[5]

---

[5] For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

The results[6] reveal that when the packet size is larger than 128 bytes, the platform with either Intel® SST-BF enabled or disabled can achieve near line-rate performance for all number of pods of service chaining. Compared with the platform with Intel® SST-BF disabled, when packet size is larger than 128 bytes, the throughput performance of the platform with Intel® SST-BF enabled in the 6-pod test case is slightly improved, ranging from 0.6% to 0.73%. In 64 bytes, the throughput performance of the platform with Intel® SST-BF enabled is improved, ranging from 18.31% to 19.83%. In 128 bytes, the throughput performance of the platform with Intel® SST-BF enabled is improved, ranging from 16% to 16.84%. In each test case, the maximum throughput is measured and the PMD cores on the host is 100%.

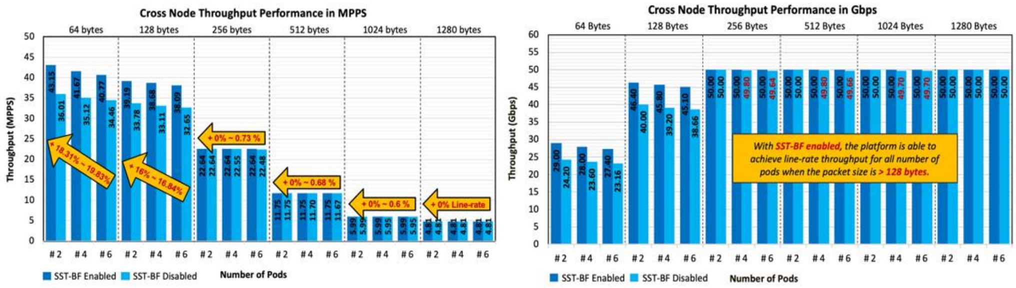

Figure 7. Throughput performance for cross-node service chaining.

SUMMARY

With a view to rapidly delivering microservices and accelerating time to 5G market, implementing cloud-native frameworks with integrated NFV is one of the approaches for CoSPs to optimize the operation process and achieve high efficiency in deploying and scaling applications. Container-based platforms can provide required services at millisecond level which can help CoSPs quicken the path from the development environment to production scenarios. The OpenShift platform with integrated network acceleration technologies improves automatic application delivery and provides guaranteed network performance.

> *"Compared to the general OpenShift platform with Intel® SST-BF disabled, the platform with Intel® SST-BF enabled is able to achieve line-rate throughput for all number of pods of service chaining when packet size is larger than 128 bytes."*

QCT collaborated with Intel to design the test scenarios, simulating the real-world production environment as a reference, and proving that the platform with enabled EPA and Intel® SST-BF features optimizes hardware usage and sustains the throughput-sensitive workload. According to the test results, the platform in the single-node and cross-node service chaining are generally verified to achieve line-rate performances in different packet sizes. Specifically, VXLAN topology designed as a tunneling protocol in cross-node service chaining simulates the practical use case for node-to-node communication and ensures network stability. With robust, stable, and reliable orchestration for container-based applications, QCT validated the platform can efficiently provide accelerated digital transformation and highly optimized resource utilization.

APPENDIX A

Hardware Configuration	OpenShift Master Node	OpenShift Worker Node
Server Model	QuantaGrid D52B-1U	QuantaGrid D52BQ-2U
Platform	Intel® Xeon® Processor Scalable Family	Intel® Xeon® Processor Scalable Family
Processor	2x Intel® Xeon® Gold 5218N Processor (16 cores, 32 threads, 2.30 GHz)	2x Intel® Xeon® Gold 6252N Processor (24 cores, 48 threads, 2.30 GHz)
Memory	192GB to 384GB	192GB to 384GB
Storage	2x 800GB SATA SSD with RAID 1	2x 800GB SATA SSD with RAID 1
Network Interface	1x 10GbE Dual Port OCP Mezzanine Card Intel® Ethernet Network Controller 82599ES SFI/SFP+	2x 25GbE Dual Port Intel® Ethernet Network Adapter XXV710 SFP28+
Switch	Management Switch: QuantaMesh T1048-LY4R Data Switch: QuantaMesh T4048-IX8D	

---

[6] For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

APPENDIX B

Software Configuration	Software Version	Location
Host OS	RHEL 7.6 (Linux version 5.4.1-1.el7.elrepo.x86_64) RHCOS 4.1 (Linux version 4.18.0-147.0.3.el8_1.x86_64)	https://www.redhat.com/en/technologies/linux-platforms/enterprise-linux
Container Orchestration	OpenShift Container Platform v4.1.24	https://www.openshift.com/products/container-platform
OpenShift Deployment Software	OpenShift Installer (RHCOS) v4.1	https://github.com/openshift/installer
	Ansible (RHEL) v2.7.11	https://github.com/openshift/openshift-ansible
Container Runtime Engine	CRI-O v1.13.11	https://github.com/cri-o/cri-o
CPU Manager for Kubernetes	CMK v1.3.1	https://github.com/intel/CPU-Manager-for-Kubernetes
Node Feature Discovery	NFD v0.2.0	https://github.com/kubernetes-sigs/node-feature-discovery
Data Plane Development Kit	DPDK v18.11.0	https://www.dpdk.org/
Open vSwitch	OpenvSwitch v2.11.0	http://www.openvswitch.org/
Userspace CNI	Userspace-CNI v1.2	https://github.com/intel/userspace-cni-network-plugin
SRIOV-CNI	dev/sriov-network-device-plugin-alpha	https://github.com/intel/sriov-cni
SR-IOV Network Device Plugin	SRIOV-DP v2.1.0	https://github.com/intel/sriov-network-device-plugin
Intel Ethernet Drivers	i40e-2.1.14-k i40evf-3.0.1-k	https://downloadcenter.intel.com/product/95260/Intel-Ethernet-Network-Adapter-XXV710-DA2

APPENDIX C

Path to BIOS Setting		BIOS Function	Required Setting for Deterministic Performance
Processor Configuration		Hyper-Threading	Enabled
PCI Subsystem Setting		SR-IOV Support	Enabled
IIO Configuration		Intel® VT for Directed I/O (VT-d)	Enabled
Power Management Configuration	**Energy Perf BIAS**	ENERGY_PERF_BIAS_CFG mode	Performance
		Power Performance Tuning	BIOS Controls EPB
	CPU P State Control	SleepStep (Pstates)	Enabled
		Activate PBF	Enabled
		Configure PBF	Disabled
		Turbo Mode	Enabled
		Energy Efficient Turbo	Disabled
	Hardware PM State Control	Hardware P-States	Native mode with no legacy support
		EPP Enable	Enabled
		RAPL Prioritization	Enabled
	CPU C State Control	CPU C6 report	Disabled
		Enhanced Halt State (C1E)	Disabled
	Package C State Control	Package C State	C0/C1 state
Memory Configuration		IMC interleaving	2-way Interleave

ABOUT Intel

Intel (NASDAQ: INTC), a leader in the semiconductor industry, is shaping the data-centric future with computing and communications technology that is the foundation of the world's innovations. The company's engineering expertise is helping address the world's greatest challenges as well as helping secure, power and connect billions of devices and the infrastructure of the smart, connected world – from the cloud to the network to the edge and everything in between. Find more information about Intel at newsroom.intel.com and intel.com.

Intel Disclaimers and Notices

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Performance results are based on testing conducted by QCT and Intel as of Jul. 1st, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Intel technologies may require enabled hardware, software or service activation.

ABOUT QCT

Quanta Cloud Technology (QCT) is a global data center solution provider. We combine the efficiency of hyperscale hardware with infrastructure software from a diversity of industry leaders to solve next-generation data center design and operation challenges. QCT serves cloud service providers, telecoms, and enterprises running public, hybrid and private clouds.

Product lines include hyperconverged and software-defined data center solutions as well as servers, storage, switches and integrated racks with a diverse ecosystem of hardware components and software partners. QCT designs, manufactures, integrates and services cutting-edge offerings via its own global network. The parent of QCT is Quanta Computer, Inc., a Fortune Global 500 corporation.

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

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 株式會社
日本国東京都港区芝大門二丁目五番八号牧田ビル３階
TEL: +81-3-5777-0818
FAX: +81-3-5777-0819

Germany
Quanta Cloud Technology Germany GmbH
Hamborner Str. 55, 40472 Düsseldorf
TEL: +49-2405-4083-1300

Korea
QCT Korea, Inc. (주식회사 큐씨티코리아)
서울특별시 영등포구 의사당대로 97
교보증권빌딩 10층, 07327
TEL: +82-10-5397-1412
FAX: +82-2-6336-6710

Other regions
Quanta Cloud Technology
No. 211 Wenhua 2nd Rd., Guishan Dist.,
Taoyuan City 33377, Taiwan
TEL: +886-3-327-2345
FAX: +886-3-397-4770