TECHNICAL BRIEF

QCT Validated Kubernetes Platform with Enhanced Platform Awareness

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

EXECUTIVE SUMMARY

Currently, Communications Services Providers (CSPs) are facing challenges with the growing demands of high-speed networks and the corresponding increase in Capital Expenditure (CapEx). Novel innovative technologies are indispensable for CSPs to improve the end user experience and transform from LTE to 5G. More and more CSPs are deploying Virtualized Network Functions (VNFs) on standard x86 servers to avoid vendor lock-in and gain flexibility. Cloud native has become an alternative approach to conventional virtual machines for accelerating network transformation. Kubernetes, originally proposed by Google, is one of the most well-known and widely-used open source container orchestration platforms. According to a RedMonk[1] article, more than 50% of the Fortune 100 companies expect Kubernetes to rapidly deploy new features. Consequently, Kubernetes is regarded as the de facto standard for container orchestration to build a flexible, scalable, and highly-reliable platform.

Quanta Computer Technology (QCT) constantly develops and coordinates with both ecosystem partners and service providers to contribute value in the ecosystem. This technical brief describes how QCT collaborated with Intel to deliver a QCT validated Kubernetes platform with Enhanced Platform Awareness (EPA) (hereinafter called "the platform"), bringing high interoperability and agility in workload orchestration. The platform architecture features enhanced data-plane networking, resource isolation, and hardware capability discovery, which are tailored for Network Function Virtualization (NFV) workloads to achieve cloud-native NFV with promising deterministic performance.

Key Highlights of QCT Validated Kubernetes Platform with EPA:

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

KUBERNETES PLATFORM OVERVIEW

Container technology provides process-level namespace isolation in which all containers on the same host share the same kernel space. This means that in comparison to VMs, containers do not simulate a separate kernel and hardware. Container technology is relatively lightweight and is able to provide workload portability to be transferred from one host to another.

Kubernetes is a production-grade container-orchestration system hosted by the Cloud Native Computing Foundation (CNCF). It provides a framework with portability, extensibility, and high-availability architecture that hosts containerized applications. Kubernetes is widely used to automate deployment, scale, maintain and validate operation of applications. It also provides features such as service discovery, storage orchestration, secret management, and horizontal scaling.

QCT combines hyperscale hardware with infrastructure software from industry leaders to solve modern datacenter challenges. During platform development, QCT applied Enhanced Platform Awareness (EPA) features in Kubernetes to maximize the platform flexibility, resource utilization, and data-plane network performance. The platform accelerates NFV adoption and deployment through software and hardware pre-integration, simplified installation tooling, and performance optimization.

QCT Validated Kubernetes Platform

The QCT QuantaGrid D52B-1U server is recommended for all Kubernetes nodes, including deployer node, master nodes, and minion nodes, and the QCT QuantaGrid D52T-1ULH servers are recommended for storage nodes, as shown in Figure 1. All nodes are equipped with Intel® 25G network interface cards (NICs), in which the Kubernetes minion nodes are configured to support both Data Plane Development Kit (DPDK) and Single-Root I/O Virtualization (SR-IOV) technologies. QCT QuantaMesh T4048-IX8D and

[1] RedMonk Website — Cloud Native Technologies in the Fortune 100
https://redmonk.com/fryan/2017/09/10/cloud-native-technologies-in-the-fortune-100/

QuantaMesh T1048-LY4R switches are recommended for data and management switching, respectively. For detailed hardware configuration and BIOS configuration, please refer to Appendices A and B.

QCT partnered with Intel to deliver a Kubernetes 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 also achieves high throughput, low-latency data-plane network performance. The platform supports Red Hat Ceph Storage to provide object, block, and file storage as backend storage, which enables S3-based image registry, persistent volume, and shared file systems for container pods. For detailed software configuration, please refer to Appendix C.

Hardware infrastructure includes:

- *1x Deployer node.*
- *3x Kubernetes Master nodes.*
- *3x Kubernetes Minion nodes.*
- *3x Red Hat Ceph Storage nodes.*

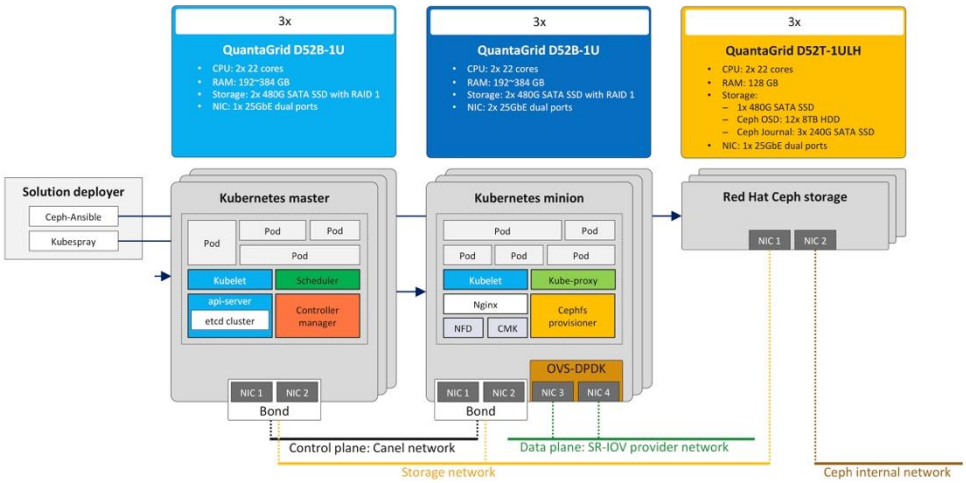

Figure 1. QCT validated Kubernetes Platform architecture.

Enhanced Platform Awareness in Kubernetes

To address explosive data growth and complex NFV workloads, the platform features a resilient, scalable architecture with high availability, and also supports Enhanced Platform Awareness (EPA) to improve the performance and determinism for container-based 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 container-based 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 Kubernetes hosts. When an NFD DaemonSet is launched, NFD pods will be initialized on each Kubernetes 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 Kubernetes hosts 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 Kubernetes 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 Kubernetes 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 Canal network for control plane traffic, the SR-IOV Network Device Plugin (SRIOV-DP) for north-south data plane traffic, and the Userspace Container Network Interface (Userspace CNI) for east-west data plane traffic.

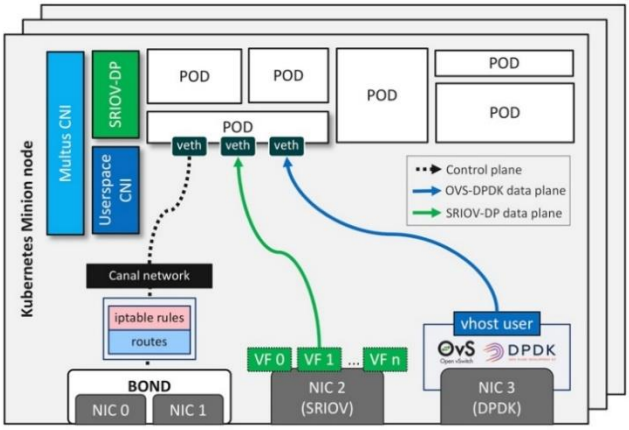

Figure 2. QCT validated Kubernetes Platform network architecture.

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

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 Kubernetes hosts. 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. A SRIOV-DP pod is subsequently spun up on each Kubernetes 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) 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 are able to 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 advantage of memory huge pages to reduce the potential for translation lookaside buffer (TLB) misses and to improve computational performance.

KUBERNETES SOLUTION PERFORMANCE

This section demonstrates how the aforementioned technologies enhance the performance of the platform. To simulate a real application, the industry-standard software tools listed below are leveraged to simulate traffic generation, packet forwarding, and CPU load generation:

- 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.

- Stress-ng is a Linux-based tool that can stress a Linux system in various situations, including CPU compute, cache trashing, I/O syncs, and context switching.

To conduct the evaluation tests, the following two scenarios are utilized:

- In the Service Chaining scenario, QCT is aimed at demonstrating end-to-end performance of container-based service chaining. This scenario illustrates the necessity of enabling SRIOV-DP and OVS-DPDK data-plane acceleration technologies to host north-south traffic and east-west traffic, achieving near line-rate throughput performance.
- In the CMK-NN scenario, QCT is aimed at demonstrating that when the CMK features are applied to the platform, the noisy-neighbor (NN) effect is effectively eliminated.

The Service Chaining and CMK-NN scenarios were designed to demonstrate that the platform can achieve near-native performance with EPA enabled. The seven test cases and platform configurations for each case are summarized in Table 1.

Test Scenarios		Service Chaining			CMK-NN			
Test Cases		1 Pod	2 Pods	4 Pods	No CMK	No CMK with Noisy Neighbor	CMK	CMK with Noisy Neighbor
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		applied			applied			
Huge Pages		1GB per page size			1GB per page size			
CMK		V	V	V	V	V		
Noisy Neighbor (Stress-ng)						V		V
Number of Pods		1	2	4	2	2	2	2
Packet Size		64, 128, 256, 512, 1024, 1280			64, 128, 256, 512, 1024, 1280			
Number of Streams		10000 bi-directional streams			10000 bi-directional streams			

Table 1. Performance test scenarios.

Service Chaining Performance Results

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 near line-rate performance and to balance the traffic loads between the SR-IOV NICs and the CPU resources used by the OVS-DPDK Poll Mode Driver (PMD).

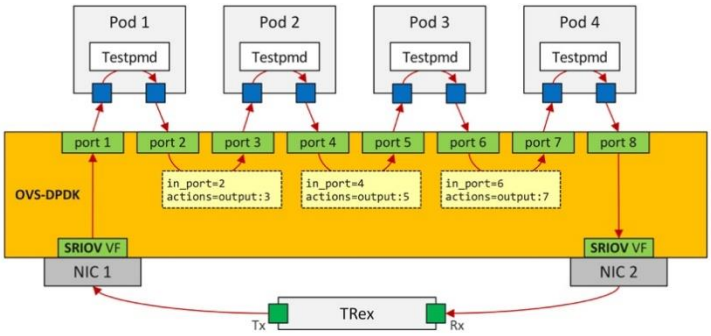

Figure 3. Network topology for service chaining scenario.

In this scenario, traffic generated by TRex flows through the SR-IOV VF of NIC 1 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 from the last pod directly passes through to the SR-IOV VF of NIC 2 and is sent back to the TRex server, as

shown in Figure 3. Based on this topology, the throughput performance with comparisons between 1 pod, 2 pods, and 4 pods are illustrated in both gigabits per second (Gbps) and million packets per second (Mpps), as shown in Figure 4. The results reveal that the platform can be adopted to achieve near line-rate performance for both 1 pod and 2 pods of service chaining for large packet sizes. When the number of pods is 4, 32 Gbps throughput was measured and the PMD cores on the host is 100%.

"Higher throughput performance shows better platform capability. The platform is able to provide high throughput performance in service chaining scenario across 1 pod, 2 pods, and 4 pods."

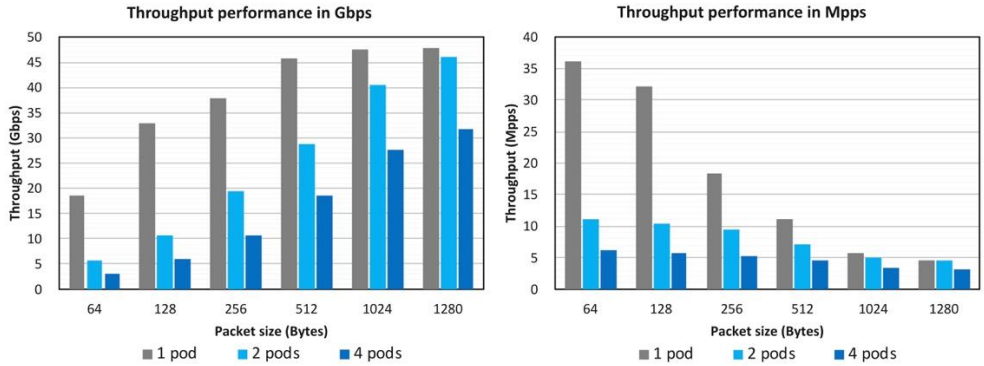

Figure 4. Throughput performance for service chaining scenario.

CMK-NN Performance Results

Service stability is another challenge for running NFV workloads. As all workloads running on the same infrastructure share the same resource pool, resource isolation has become a key feature of NFV to ensure resource availability for performance sensitive workloads. In the CNK-NN scenario, the results demonstrate that the platform is capable of eliminating the noisy-neighbor effect when the CMK features are enabled.

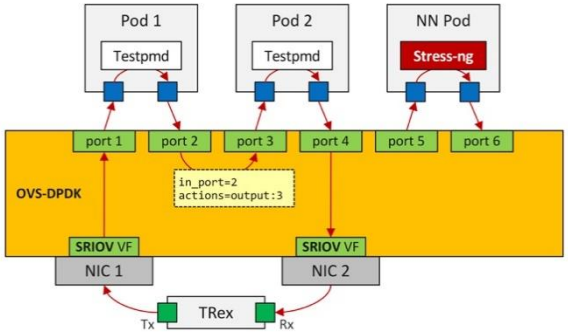

Figure 5. Network topology for CMK-NN scenario.

Similar to the service chaining scenario, SRIOV-DP and OVS-DPDK are enabled to host the north-south traffic and the east-west traffic, respectively. In the CMK NN scenario, the traffic forwarding in each pod is hosted by the DPDK Testpmd application and the traffic redirection across pods on host is defined by the Open vSwitch flow table, as shown in Figure 5. All the pods, including pod 1, pod 2, and the NN pod require 2 CPU cores. If the CMK feature is applied, 2 dedicated CPU cores will be assigned to each pod for resource affinitization; otherwise, CPU resources will be assigned according to the OS scheduler, which means all the pods will share the same CPU resource without resource isolation.

The results of the throughput performance in the CMK-NN scenario is presented in gigabits per second (Gbps) and million packets per second (Mpps), as shown in Figure 6. The results reveal that when the CMK feature is not applied on the platform, the noisy-neighbor effect will degrade the performance approximately 40 to 60 percent. However, when the CMK feature is enabled, the platform shows no performance degradation, even when the CPU is heavily loaded, and the platform effectively eliminates the noisy-neighbor effect.

"Compared to the general Kubernetes platform without CMK enabled, the platform with CMK enabled is able to eliminate the noisy-neighbor effect and provide better resource isolation."

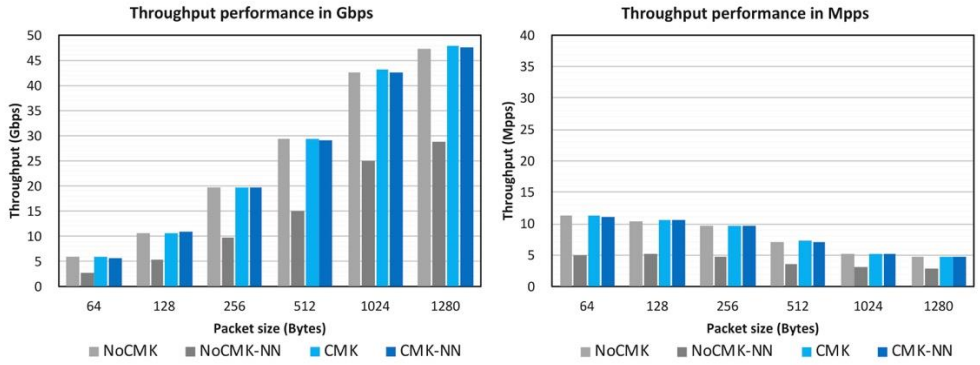

Figure 6. Throughput performance for CMK-NN scenario.

SUMMARY

To take the leading position of network transformation, cloud-native approach is best suited to evolve the NFV infrastructure and change the way the CSPs industry functions. In the next few years, leveraging the lightweight, highly-portable, easy-to-use container technology for NFV is serving the move to Cloud Native. Thus, it is essential to introduce container orchestration platforms such as Kubernetes to improve operation efficiency and provide guaranteed performance in CSPs deployments.

QCT collaborated with Intel to deliver a platform with EPA features that provides stable and reliable orchestration for container-based applications. According to the test results in two scenarios, the platform can be adopted to achieve near line-rate performance among different packet sizes in both single-pod and two-pod service chaining while the noisy-neighbor effect can be eliminated efficiently when CMK is enabled. QCT has developed a robust, resilient, and scalable Kubernetes platform to accelerate cloud-native transformation and minimize operational expenses to realize NFV success.

APPENDIX A

Hardware Configuration	Kubernetes Master Node	Kubernetes Minion Node	Ceph Storage Node
Server Model	**QuantaGrid D52B-1U**	**QuantaGrid D52B-1U**	**QuantaGrid D52T-1ULH**
Platform	Intel® Xeon® Processor Scalable Family	Intel® Xeon® Processor Scalable Family	Intel® Xeon® Processor Scalable Family
Processor	2x Intel® Xeon® Gold 6152 Processor (22 cores, 44 threads, 2.10 GHz) Microcode Revision: 02000029	2x Intel® Xeon® Gold 6152 Processor (22 cores, 44 threads, 2.10 GHz) Microcode Revision: 02000029	2x Intel® Xeon® Gold 6152 Processor (22 cores, 44 threads, 2.10 GHz) Microcode Revision: 02000029
Memory	192GB to 384GB	192GB to 384GB	128GB
OS Disk	2x 480GB SATA SSD with RAID 1	2x 480GB SATA SSD with RAID 1	1x 480GB SATA SSD
Local Storage	N/A	N/A	Ceph OSD: 12x 8TB HDD Ceph Journal: 3x 240G SATA SSD
Network Interface	1x Dual Port 25GbE Intel® Ethernet Network Adapter XXV710 SFP28+	2 x Dual Port 25GbE Intel® Ethernet Network Adapter XXV710 SFP28+	1 x Dual Port 25GbE Intel® Ethernet Network Adapter XXV710 SFP28+
Switch	Management Switch: QuantaMesh T1048-LY4R Data Switch: QuantaMesh T4048-IX8D		

APPENDIX B

BIOS Configuration		BIOS Setting	Settings for MAX Performance	Required/Recommended
PCI Subsystem Setting		SR-IOV Support	Enable	Required
IIO Configuration		Intel® VT for Directed I/O (VT-d)	Enable	Required
Power Management Configuration	CPU P State Control	SleepStep (Pstates)	Enable	Recommended
		Turbo Mode	Enable	Recommended
	Hardware PM State Control	Hardware P-States	Native Mode	Recommended
	CPU C State Control	Autonomous Core C-State	Disable	Recommended
		CPU C6 Report	Auto	Recommended
		Enhanced Halt State (C1E)	Enable	Recommended
	Package C State Control	Package C State	Auto	Recommended
Memory Configuration		Volatile Memory Mode	1 LM mode	Recommended

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 10 층, 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

APPENDIX C

Software Configuration	Software Version	Location
Host OS	RHEL 7.6	https://www.redhat.com/en/technologies/linux-platforms/enterprise-linux
Ansible	Ansible 2.7.1	https://www.ansible.com/
Python	Python 2.7.5	https://www.python.org/
Kubespray	Kubespray v2.8.0	https://github.com/kubernetes-sigs/kubespray
Docker	Docker 18.06.1-ce	https://www.docker.com/
Container Orchestration	Kubernetes v1.12.2	https://github.com/kubernetes/kubernetes
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

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

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.

Disclaimers

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 technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at intel.com.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice Revision #20110804

Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation.