WHITE PAPER

Mellanox Seamlessly Integrates with OpenStack, Increasing Efficiency and Reducing Operational Costs

High-speed Ethernet with hardware offloads delivers total infrastructure efficiency for NFV and Cloud Data Centers

Executive Summary

OpenStack, commonly referred to as "the Linux of the Cloud", allows companies to utilize open-source initiatives and a transparent and collaborative approach to implement public and private cloud solutions to achieve business agility, infrastructure elasticity and operational simplicity. While OpenStack provides a flexible framework, it is particularly important that the cloud infrastructure, composed of compute, network and storage resources, runs at maximum performance and efficiency to guarantee overall application performance. This paper discusses the requirements for cloud network infrastructure to properly support web-scale IT with Red Hat OpenStack as the cloud management platform, and tightly integrated with Mellanox end to end networking solutions.

Challenges to Efficient Cloud Deployment

To achieve multi-tenancy and automation goals, cloud deployments often leverage disaggregation and virtualization technologies. However, this comes at a cost of significant performance penalties which manifest themselves as low data communication and storage access performance, and heightened CPU utilization. As a result, organizations compensate by over-provisioning CPU cores to improve the performance resulting in a larger hardware footprint and capital expenditures, thus reducing total infrastructure efficiency.

Understanding Compute Virtualization Penalties

In virtualized environments, multiple virtual machine (VM) instances run simultaneously over physical server hardware. This has necessitated virtual switch software that often reside alongside the hypervisor in the OS kernel to handle network I/O traffic to and from VMs. While virtualization does

Key Benefits

- **Achieve 10X Better Performance** than vanilla OVS with ASAP^2 OVS Offloads
- **Free up 100% CPU Cores** with ASAP^2
- **The Best Industry Throughput and 50% CapEx Savings** for DPDK
- **Improve Latency by 20X** with SR-IOV
- **Improve CPU Utilization by 80%** with Overlay Networks Offloads
- **Deliver 6X Throughput** with RDMA
- **Improve TCO by Marrying Bare-Metal and Virtualized OpenStack Clouds** with Mellanox Spectrum Switches and ConnectX Intelligent Network Adapters

bring flexibility, it also results in significantly degraded I/O performance degradation due to increased layers of processing in software and the CPU is burdened with most of the virtual network I/O processing.

Avoid Network Virtualization Penalty

Overlay tunneling protocols such as VXLAN, NVGRE or GENEV are not all recognized by every server NIC and processing these new packet formats without NIC hardware offload needs to be done by the OS kernel in CPU, resulting in lower and nondeterministic I/O performance and increased CPU load.

Solving Storage Virtualization Penalties

The TCP/IP protocol stack is not the most efficient to power storage networks. The protocols were designed to incorporate a lot of handshakes between endpoints, and it is almost impossible to offload all protocol handling operations into the NIC hardware. As a result, complex protocol software needs to run in the CPU. These can result in low storage access bandwidth, low IOPS, and high CPU overhead.

To overcome these penalties and achieve ultimate infrastructure efficiency and application performance, cloud operators are looking to implement efficient virtual network solutions that provide excellent virtualization and bypass the TCP/IP stack in processing storage IO to achieve acceleration and efficiency in cloud networks.

Mellanox OpenStack Cloud Network Solution

Through an end-to-end suite of interconnect products of adapters, switches, cable/optics, and associated network driver and management software, Mellanox enables cloud data centers to achieve the highest efficiency through a high-performance, low-latency cloud network with rich network offload with acceleration and automation features. Mellanox can mitigate the above-mentioned penalties, delivering cloud networks that can handle line-rate processing at 10, 25, 40, 50, and 100Gb/s speeds, supporting high-throughput, high-IOPS storage operations, with minimal CPU overhead so that infrastructure resources can be dedicated to actual application workload.

Overcoming Penalties

Mellanox provides the foundation for efficient cloud infrastructure through disaggregation and virtualization solutions that mitigate performance penalties associated with compute, network, virtualization and storage. This enables cloud applications to run at the highest performance and efficiency. Mellanox achieves higher cloud efficiency through the following solutions; Open vSwitch Offloads (OVS), OVS over DPDK, Network Overlay Virtualization, SR-IOV, and RDMA.

Increase OVS Efficiency with ASAP^2

Open vSwitch (OVS), vRouter, VPP and Linux Bridge are some of the most popular virtual switch platforms used in OpenStack cloud deployments today. OVS hardware offloads accelerate the traditional slow virtual switch packet performance by an order of magnitude. Essentially, offering the best of both worlds: Hardware acceleration of the data path (fast-path) for high-throughput flows along with unmodified standard OVS control path for flexibility and programming of match-action rules.

Mellanox offers an open source and high performance OVS offload solution called Accelerated Switching and Packet Processing (ASAP^2). ASAP^2 fully and transparently offloads networking functions such as overlays, routing, security and load balancing to the adapter's embedded switch (e-switch). ASAP^2 provides throughput of 66Mpps for small packets and ~line rate performance for large packets while completely freeing up the CPU cores.

ASAP^2 is a significant improvement over traditional approaches because it ensures that SDN and network programmability capabilities are maintained, and at the same time, network I/O achieves highest performance on compute nodes. With virtual switch/router being utilized in almost all production OpenStack deployments, it makes sense to use ASAP^2 with any virtual switch/router implementation to give a tremendous performance boost in terms of higher packet throughput and lower latencies and improve cloud efficiency. Mellanox ASAP^2 is fully integrated with RHEL 7.5 and Red Hat OSP 13.

The Best Performance for DPDK

Data Plane Development Kit (DPDK) reduces overhead caused by interrupts that are sent each time a new packet arrives for processing. DPDK implements a polling process for new packets to achieve the key benefits of significantly improving processing performance while eliminating PCI overhead and maintaining hardware independence. Although DPDK technology consumes CPU cycles, Mellanox ConnectX-5 adapters offer the industry's highest bare-metal packet rate of 139 million packet per second for running OVS or VNF cloud applications over DPDK. Mellanox DPDK is fully Red Hat supported for RHEL 7.5 and OSP 13.

Network Virtualization with VXLAN Offload, and VTEP Gateway

Mellanox started offloading VXLAN protocol processing to the NIC since the ConnectX-3 generation of NICs. The VXLAN Offload feature enables the NIC to handle stateless processing of VXLAN packets such as checksum calculation,

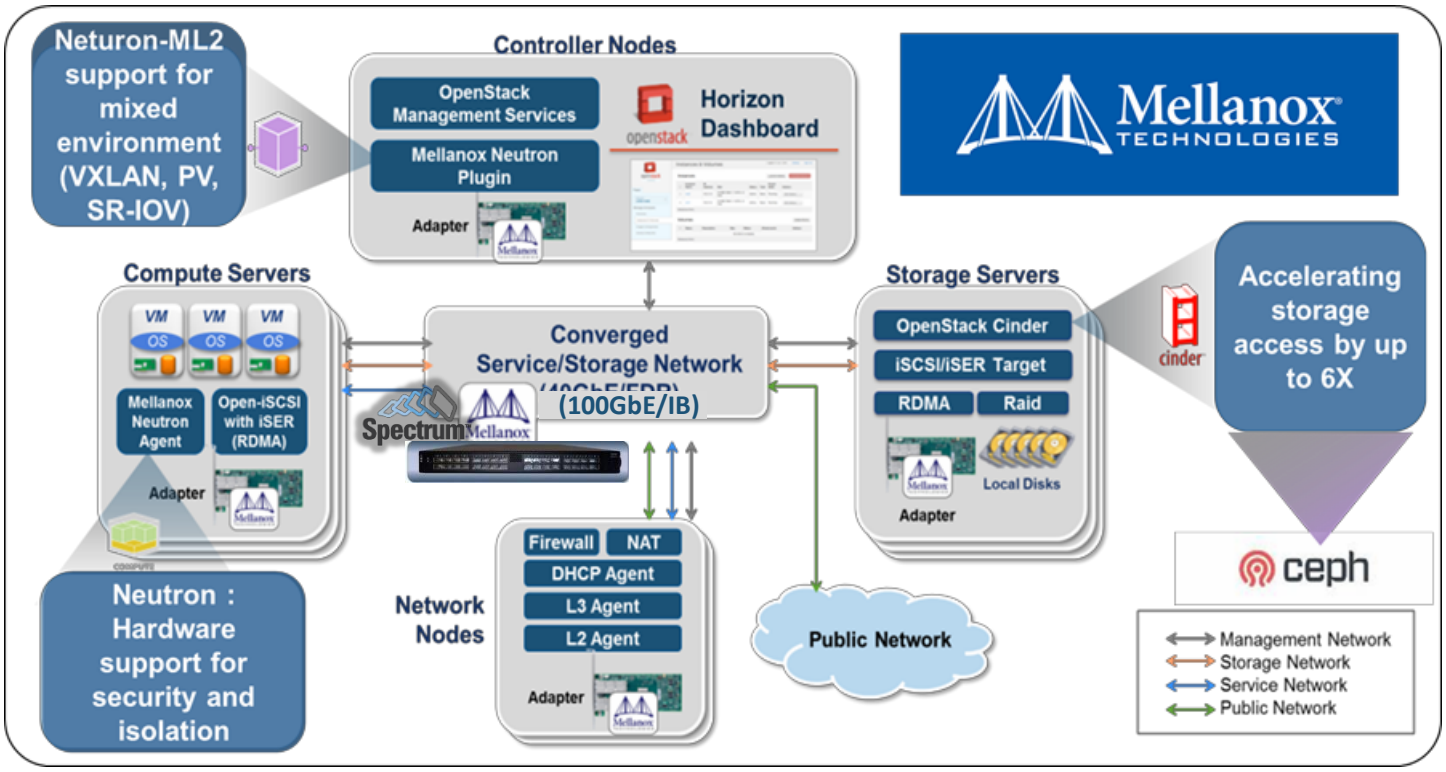

Figure 1. Comprehensive OS Integration for Mellanox Switch and Adapter

Receive Side Scaling (RSS), Large Segmentation Offload (LSO), etc, significantly improving throughput and latency performance, and reducing CPU overhead associated with overlay packet processing. In addition to VXLAN, Mellanox NICs also support offload of other overlay encapsulation protocols such as NVGRE and GENEVE.

Oftentimes VXLAN networks needs to communicate with other networks such as VLAN networks that supports bare metal servers, or wide area networks for data center interconnect, and North-South user traffic. This necessitate VXLAN Tunnel Endpoint (VTEP) gateway which allow to connect VXLAN networks to other type (i.e., VLAN) of networks. Mellanox Spectrum switches support VTEP gateway functionality in hardware, ensuring highest performance when heterogeneous networks in the cloud communicate with each other.

Overcome Compute Virtualization Penalty with SR-IOV

Single Root I/O Virtualization (SR-IOV) allows a device, such as a network adapter, to separate access to its resources among various PCIe hardware functions. This allows traffic streams to be delivered directly between the virtual machines and their associated PCIe partitions, giving applications direct access to the I/O hardware. As a result, the I/O overhead in the software emulation layer is eliminated. SR-IOV enables VMs to achieve network performance that is nearly the same as in non-virtualized environments.

Mellanox NICs support basic SR-IOV as well as advanced features such as SR-IOV High Availability (HA) and Quality of Service (QoS). SR-IOV HA provides a redundancy mechanism for VFs by using Link Aggregation Group (LAG) to bind two VFs from two different port on the same NIC together and exposes the bundle as one VF to the VM. When one VF in the bundle fails, the other VF continues forwarding traffic without affecting VM I/O operations.

Overcome Storage Virtualization Penalty with RDMA/RoCE

The large overhead associated with stateful protocols such as TCP dictates that it is not an ideal transport protocol for software defined scale-out storage applications, especially when storage media gets faster as it transitions from hard disks to solid-state drives (SSD) to Non-Volatile Memory (NVM). Remote Direct Memory Access (RDMA), on the other hand, is a protocol designed for high-speed links within data center environment that can overcome the inefficiencies of TCP. RDMA can run over InfiniBand (IB) or over Converged Ethernet (RoCE). RDMA kernel bypass, read and write network semantics and full transaction offload to RDMA capable NIC devices guarantee the highest possible throughput, lowest latency, and minimal CPU overhead, making it ideal for storage access. Typically, RDMA over Converged Ethernet (RoCE) requires the network to be configured for lossless operation, however, Mellanox has recently enhanced RoCE with built-in error recovery mechanisms. While a lossless network has never been a strict requirement, customers typically configure

Mellanox provides Remote Direct Memory Access (RDMA) capabilities to improve storage performance by up to 6X as compared to conventional adapters without RDMA support

their networks to prevent packet loss and ensure the best performance. With this new version, RoCE can be deployed on ordinary, Ethernet networks. By utilizing RDMA or RoCE, virtual servers can achieve much higher I/O performance because most of the packet processing is offloaded to the NIC. This further enables increased performance, improved latencies and significantly reduced CPU overhead. The net effect is an improvement of overall server and application efficiencies.

Open Composable Networks

The goal of Mellanox is to enable OpenStack deployments with a fully integrated suite of high-performance, highly programmable networking components including switches, network adapters, optical modules and cables. The key to being open and composable is to support open APIs and standard interfaces, as well as disaggregating hardware from software and allowing choice of network operating systems. Mellanox embraces this open philosophy, which completely and truly frees organization from vendor lock-in, all the way down to the switch silicon level. Mellanox Spectrum switching silicon offers a choice of popular open network operating systems like Cumulus, Sonic, and Mellanox Onyx while providing best in class hardware that delivers zero packet loss, fair traffic distribution and more predictable application performance compared to merchant silicon in other switch offerings. Mellanox NEO™, networking orchestration and management software, is a powerful platform for management, monitoring, and visualization of scale-out networks.

OpenFlow On Spectrum Switches

Mellanox Spectrum™ switches are fully open, without locking of cables or features. All features are available including IP unnumbered BGP for the underlay and VTEP for the overlay, with controller integrations or EVPN. Spectrum is Mellanox's 10/25/40/50 and 100Gb/s Ethernet switch that is optimized for SDN to enable flexible and efficient data center fabrics with leading port density, low latency, zero packet loss, and non-blocking traffic flow.

From the ground up, starting at the switch silicon level, Spectrum is designed with a flexible processing capacity so that it can accommodate a programmable OpenFlow pipeline that enables packets to be sent to subsequent tables for further processing, and allows metadata to be communicated between OpenFlow tables. In addition, Spectrum is an OpenFlow-hybrid switch that supports both OpenFlow operation and normal Ethernet switching operation simultaneously. Users can configure OpenFlow at port level, assigning some Spectrum ports to perform OpenFlow based packet processing operations and others to perform normal Ethernet switching operations. Spectrum can even mix and match algorithms on the same switch port by using a classification mechanism to direct traffic to either the OpenFlow pipeline or traditional Ethernet processing.

Figure 2. Mellanox's Comprehensive Cloud Partner Ecosystem

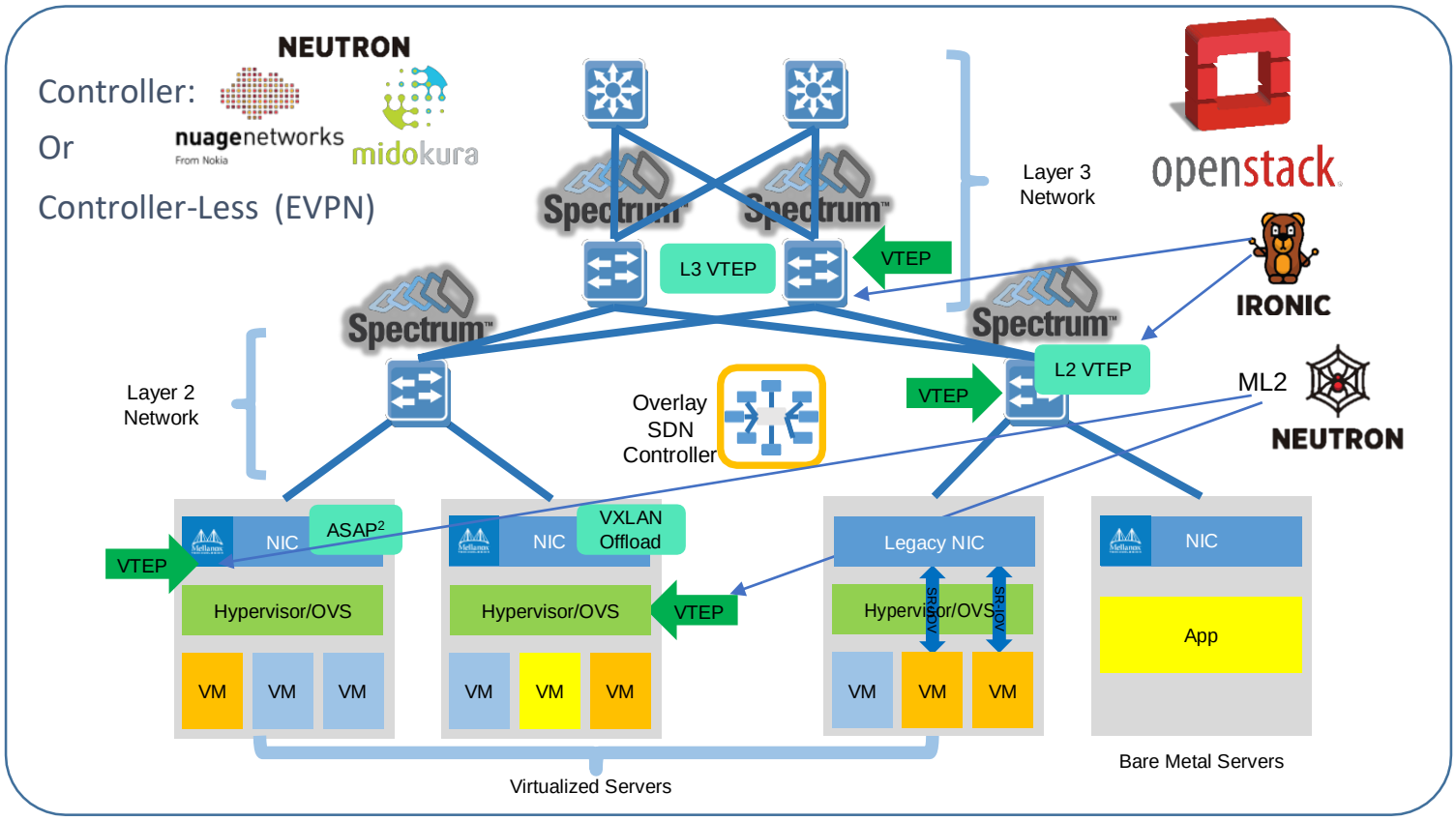

Figure 3. Mellanox's OpenStack VTEP Integration for Virtualized and Bare-Metal Clouds

VTEP Support In Spectrum Switches

There is a need for VXLAN Termination End Point (VTEP) to connect bare-metal servers to the virtual networks in OpenStack. This is best handled by the switch hardware as opposed to on the Hypervisor which can create a bottleneck in the OVS. VTEP can be implemented on Spectrum to offload overlay network technologies such as VXLAN, NVGRE or Geneve. Hardware VTEPs, like the ones deployed by Spectrum can achieve higher performance but pose added complexity on a ToR switch since there is a need for the switch to be VM-aware, which requires a large forwarding table be maintained for VM Mac address or VLAN to VXLAN translations. For this reason, a high-performance switch like Spectrum is required. Mellanox Spectrum supports VTEP gateway functionalities which make it ideal to be deployed as:

- Layer 2 VTEP gateway between virtualized networks using VXLAN and non-virtualized networks using VLAN in the same data center or between data centers. Layer 2 VTEP gateway to provide high-performance connection to virtualized servers across Layer 3 networks and enable Layer 2 features such as VM live migration (VMotion). On virtualized server hosts, where the NIC does not have VTEP capability or software VTEP can't meet the network I/O performance requirement, the VTEP can be implemented on Mellanox Spectrum ToR.
- In some cases, the application running in the VM may desire to use advanced networking features such as Remote Direct Memory Access (RDMA) for inter-VM communication or access to storage. RDMA needs to run in SR-IOV mode on virtualized servers and in cases when Mellanox NIC is not present, the VTEP is best implemented in the ToR. Mellanox Spectrum is the ideal switch to build an ethernet storage fabric. It leverages the speed, flexibility, and cost efficiencies of Ethernet with the best switching hardware and software packaged in ideal form factors to provide performance, scalability, intelligence, high availability, and simplified management for storage.
- Layer 3 VTEP gateway that provides VXLAN routing capability for traffic between different VXLAN virtual networks, or for north-south traffic between an VXLAN network and a VPN network or the Internet. This feature is supported in Spectrum hardware with cumulus network operating system.

Overlay SDN can be deployed to achieve network virtualization and automation without requiring upgrades of physical networking equipment, more specifically, network devices that are NOT the VTEPs. Beyond the virtualized environment with VXLAN/NVGRE/GENEVE, there are often Bare Metal Servers (BMS) or legacy networks that can only use VLAN, or North-South traffic that goes out to a VPN

network or the Internet. In those cases, using a software VTEP gateway adds an extra hop or and performance bottleneck can occur. Best practice is to use the ToR that the BMS is connected to as hardware VTEP. This achieves line rate performance while saving costs. Mellanox NEO is a ethernet fabric monitoring and provisioning tool for Mellanox NICs and switches. NEO is pre-integrated with OpenStack Horizon and supports VTEP provisioning of ConnectX NICs using Neutron ML2 mechanism drivers and Spectrum switch using Ironic conductors.

Marrying Bare-Metal and Virtualized OpenStack Clouds

Servers equipped with Mellanox ConnectX-4/ConnectX-5 provides stateless VXLAN offloads with a simple driver configuration. Further, ConnectX-5 is the best network adapter to provision a high-performance software VTEP per server for inter-VM or VM to public network communication. Mellanox NICs are integrated with popular commercial SDN controllers such as Nuage Virtualized Services Platform(VSP), or with open source controllers such as OpenDaylight. Thus, with ConnectX adapters customers can easily build an efficient and non-blocking virtualized OpenStack cloud with ASAP^2 OVS offload technology.

Mellanox Spectrum is an ideal switch to terminate VXLANs for bare-metal servers in a multitenant cloud. With SDN overlay controller solutions such as Nuage VSP, Open Contrail and vmware NSX, VXLANs can be terminated on the ToR to bare metal servers and traditional VLAN segments. In an environment where an SDN controller isn't needed, Spectrum switches can support a controller-less VXLAN overlay network using standard BGP EVPN protocol. Such a solution eliminates cost of controller licenses and is interoperable with other switches that support standard BGP EVPN based VXLANs.

With a common SDN control layer across ConnectX software VTEPs and Spectrum hardware VTEPs, customers can easily unify deployment and operations of bare-metal and virtualized OpenStack clouds. Further, Mellanox's integration with SDN ecosystem partners enables a turnkey OpenStack cloud solution that achieves lower total cost of ownership.

Conclusion

As an industry leader in high-performance networking technologies, Mellanox understands the risks and rewards of transforming a data center. As IT organizations transition to cloud-based and service-centric infrastructures, the need for gaining network and server efficiencies is paramount to transition beyond 10Gb server I/O. By combining key technologies from the adapter and switch, Mellanox is able to accelerate virtual and bare-metal networks and reduce CPU utilization through hardware-based offloads for increased scalability, greater flexibility and highest efficiency in the modern software-defined data centers to bring highest Return on Investment to our customers.

Learn more about Mellanox and OpenStack

Mellanox OpenStack Reference Architecture:

http://www.mellanox.com/openstack/pdf/mellanox-openstack-solution.pdf

Mellanox Red Hat OpenStack Reference Architecture:

http://www.mellanox.com/related-docs/whitepapers/Mellanox-OpenStack-Solution-for-Red-Hat.pd

Ceph White Paper:

http://www.mellanox.com/related-docs/whitepapers/WP_Deploying_Ceph_over_High_Performance_Networks.pdf

Mellanox Scale-Out Open Ethernet Products:

http://www.mellanox.com/page/ethernet_switch_overview

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com