Utilize Efficient Hardware to Scale-Out Open Source Cloud Environments

High-speed Ethernet with stateless offloads deliver network and server efficiencies in hyper-scale cloud services

KEY BENEFITS

Mellanox integrates with OpenStack and allows customers to deploy with confidence

- **Deliver 6X** Better Throughput with RDMA
- **Improve** Connectivity by 20x with SR-IOV
- **Reduce** CPU Utilization by 80% with Offloads

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 operation 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 requirements for cloud network infrastructure to properly support web-scale IT with OpenStack as the cloud management platform, and corresponding Mellanox solutions.

Deployment Efficiency Challenges

In order to achieve multi-tenancy and automation goals, cloud deployments often leverage virtualization technologies, which started with compute virtualization and extended to network and storage virtualization. Virtualization significantly improves hardware utilization as well as ease of resource orchestration, but when deployed improperly, can cause performance and efficiency degradation. The degradation manifests itself as low data communication and storage access performance, and heightened CPU utilization or eventually larger hardware footprint and capital expenditure.

Compute Virtualization Penalty

In a hypervisor-managed virtualized environment (as opposed to bare metal), 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. The virtual switch does bring enhanced flexibility but also results in I/O performance degradation due to increased layers of processing in software. As server I/O moves to 10Gb/s and beyond, it is challenging for vanilla virtual switches to reach line rate. In some cases where applications demand small packet performance, fewer than 1Mpps (million packets per second) are achieved on a 10 Gbps link which can in theory support 15 Mpps.

Network Virtualization Penalty

Software Defined Networking (SDN) technology is becoming a key component of OpenStack cloud deployment to improve network programmability, elasticity and automation through network virtualization. Mainstream SDN vendors have adopted overlay-style SDN to support rapid virtual network provisioning and ultimately cloud agility without heavy dependence on the underlay physical switch fabric. Overlay SDN introduces new tunneling protocols such as VxLAN, NVGRE or GENEVE. Not all server NICs can recognize and process these new packet formats and without NIC hardware offload of overlay tunnel protocol processing, the majority of virtual network I/O processing needs to be done by virtual switches in CPU, resulting in low and nondeterministic I/O performance and increased CPU load.

Storage Virtualization Penalty

The new generation of cloud and Big Data technologies drive the need for distributed software-defined, scale-out and object-based storage and a preference for Ethernet speeds and converged infrastructure as opposed to legacy Fiber Channel based storage networks. But the TCP/IP protocol stack commonly riding over Ethernet is not the most efficient to power storage networks. TCP/IP protocols originated from wide area networks where bandwidth used to be scarce and congestions were a common phenomenon. As a result, the protocols were designed to incorporate a lot of handshakes between the endpoints, and it is almost impossible to offload all protocol handling operations into the NIC hardware as 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 challenges and achieve ultimate cloud efficiency and application performance, cloud operators are looking to implement efficient virtual network solutions that provide excellent virtualization, acceleration and automation support.

Mellanox OpenStack Cloud Network Solution

Through an end-to-end suite of interconnect products such as NICs, switches, cable/optics, and associated network driver and management software, Mellanox Efficient Virtual Networks (EVN) solution enables cloud data centers to achieve the highest efficiency through a high-performance, low-latency cloud network with rich network offload, acceleration and automation features. EVN can mitigate the above-mentioned virtualization 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 dedicated to actual application workload can be maximized.

Mellanox EVN achieves higher cloud efficiency goal through the virtualization efficiencies, acceleration techniques and automation technologies which will be discuss in the following sections.

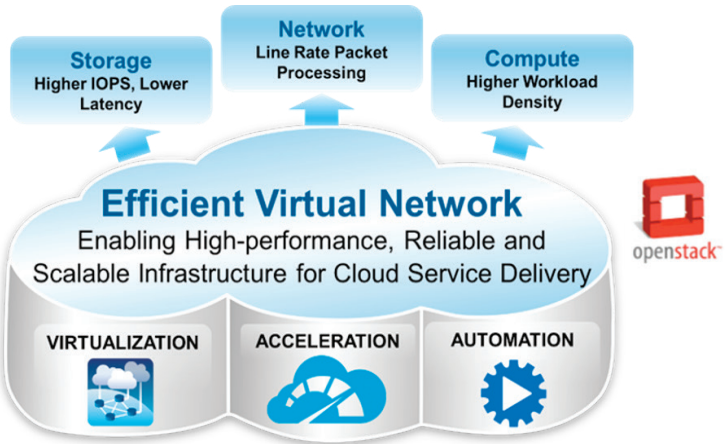

Figure 1. *Mellanox EVN – Foundation for Efficient Cloud Infrastructure*

Virtualization

Mellanox virtualization solution mitigates performance penalties associated with compute, network and storage virtualization and enable cloud applications to run at the highest performance and efficiency.

Overcome Compute Virtualization Penalties 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 partition, giving applications direct access to the I/O hardware. As a result, the I/O overhead in the software emulation layer is diminished. SR-IOV can enable 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 like 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 together two VFs from two different port on the same NIC, and expose 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. SR-IOV QoS feature allows much more granular control of rate limiting and bandwidth guarantee on a per VF basis.

Overcome Network Virtualization Penalty with VxLAN Offload, VTEP Gateway and ASAP2

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

Built on top of the success of VxLAN Offload, Mellanox introduced Accelerated Switching and Packet Processing (ASAP2) starting from ConenctX-4 generation of NICs. This feature enables the majority of data plane switching and packet processing that a virtual switch performs to be offloaded to the Mellanox NIC, while maintaining control plane operations of virtual switches. ASAP2 is a significant improvement from traditional SR-IOV 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. ASAP2 is also a significant improvement over VxLAN Offload, as it can support additional packet processing operation offload (such as VxLAN encapsulation, decapsulation and packet classification) beyond the stateless offload supported by VxLAN Offload. With OVS being utilized in almost half of all production OpenStack deployments, it makes sense to use ASAP2 with any OVS implementation to give a tremendous performance boost in term of higher packet throughput and lower latencies, and improve cloud efficiency.

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 the VTEP gateway. Mellanox Spectrum switches support VTEP gateway functionality in hardware, ensuring highest performance when heterogeneous networks in the cloud communicate with each other.

Figure 2. *Mellanox Partners with and Integrates with a Comprehensive set of OpenStack vendors*

Overcome Storage Virtualization Penalty with RDMA/RoCE

The large overhead associated with stateful protocols such as TCP dictates that it is not suitable as an interconnect protocol for software defined scale-out storage applications, especially when storage media gets faster from harddisks 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 inefficienies of TCP. RDMA can run over InfiniBand (IB) or over Converged Ethernet (RoCE). With RDMA, all data transfer operations can be offloaded to RDMA-capable NICs, guaranteeing the highest possible throughput, lowest latency at minimal CPU overhead, making it ideal for storage access. Typically this is implemented over enhanced Ethernet which is 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 their networks to prevent packet loss and ensure the best performance. With this new version, RoCE can be deployed on ordinary, standard, Ethernet networks. By utilizing RDMA or RoCE, virtual servers can achieve much higher I/O performance because the majority of packet forwarding can be offloaded to the NIC. This further enables increased performance, decreased latencies and significantly reducing the CPU burden. The net effect is an improvement of overall server and network efficiencies.

Mellanox provides RDMA capabilities to improve storage performance by up to **6 times** when compared to conventional performance

Acceleration

Mellanox adapters and switches provide the network-acceleration needed to run an efficient and scalable public and private cloud solutions based on OpenStack. The flexible framework of OpenStack is complemented by the Mellanox high-performance Ethernet solutions. Together they form a cloud infrastructure, composed of compute, network and storage resources which run at maximum performance and efficiency. This guarantees overall application performance while supporting web-scale IT now and into the future with up to 100GbE end-to-end products.

Automation

The provisioning and management of cloud networking resources is a crucial component to achieving deployment goals of business agility, and infrastructure elasticity. This means the ability to act or react quickly and requires operational simplicity. Mellanox NEO™, a powerful platform for data-center network orchestration, was designed to simplify network provisioning, monitoring and operations. This includes easily integration with OpenStack components and core services through REST APIs. NEO's network REST API technology allows OpenStack services access to fabric-related data and provisioning activities. This allows for simplified, yet robust automation capabilities from network staging and bring-up, to day-to-day operations.

Open Composable Networks

The goal of OpenStack deployments are to provide a set of high-performance, highly programmable networking components including switches, server 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 complements this approach which completely and truly frees organization from vendor lock-in strategies, all the way down to the switch silicon level. Mellanox Spectrum switching silicon offers choice of popular open source operating systems like Cumulus while providing best in class hardware that deliver zero avoidable packet loss, fair traffic distribution and offer more predictable application performance compared to merchant silicon in other competitive switches.

Mellanox NEO™, networking orchestration and management software, is a powerful platform for managing scale-out networks. Mellanox NEO enables data center operators to efficiently provision, monitor and operate a modern software-defined data center fabric. NEO serves as interface to the fabric, thus extending existing tools capabilities and enhancing the ability to configure and provision switches and routers. With integration to common advanced cloud networking software solutions like OpenStack Neutron and VMware vSphere, NEO enhances provisioning and removes the need for human intervention.

Conclusion

As an industry leader in high-performance networking technologies, Mellanox understand 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 when transition beyond 10Gb server I/O. By combining key technologies from the adapter and switch, Mellanox is able to accelerate virtual networks and reduce CPU utilization through hardware-based stateless offloads for increased scalability and greater flexibility in the modern software-defined data centers.

About Mellanox

Mellanox Technologies (NASDAQ: MLNX) is a leading supplier of end-to-end Ethernet and InfiniBand intelligent interconnect solutions and services for servers, storage, and hyper-converged infrastructure. Mellanox intelligent interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications and unlocking system performance.

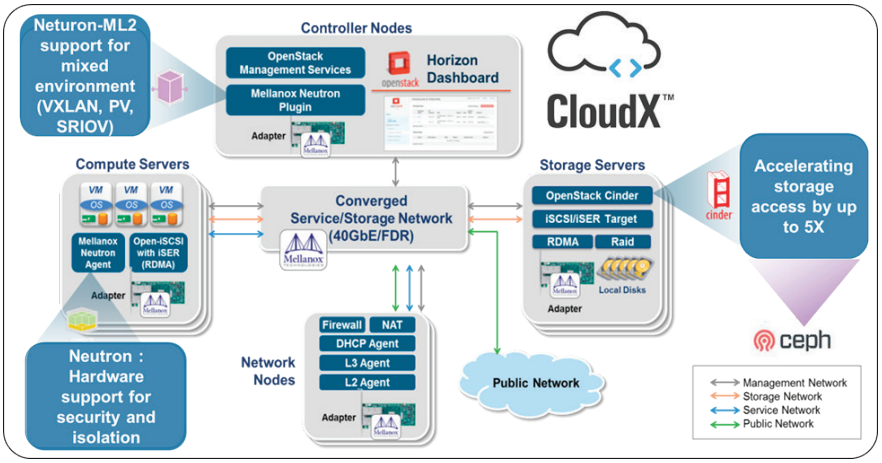

Figure 3. *Mellanox offers seamless integrate between its products and OpenStack services*

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.pdf

Mellanox and OpenStack Ceph White Paper:
http://www.mellanox.com/related-docs/whitepapers/WP_Deploying_Ceph_over_High_Performance_Networks.pdf

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

MLNX-31-426