WHITE PAPER

Software Defined Networking, Done Right

Contents

Overview 2
SDN Deployment Models 2
Device-Based SDN Deployment Model 2
Overlay SDN Deployment Model 3
Proprietary SDN Solutions 4
Mellanox SDN Technology Highlights 4
VXLAN Offload on Connect-X NICs 4
ASAP^2 (Accelerated Switching and Packet Processing) on ConnectX-4 NICs 4
ASAP^2 Direct 5
ASAP^2 Flex 5
OpenFlow support on Spectrum Switches 6
VTEP support in Spectrum Switches 6
Building the Most Efficient SDN Networks with Mellanox Interconnect 7
Recommended deployment for OpenFlow based SDN networks 7
Recommended Deployment for Overlay SDN Networks 8
Conclusion 8

Overview

Software-Defined Networking (SDN) is a revolutionary approach to designing, building and operating networks that aims to deliver business agility in addition to lowering capital and operational costs through network abstraction, virtualization and orchestration. Conceptually, SDN decouples the control and data planes, and logically centralizes network intelligence and control in software-based controllers that maintain a global view of the network. This enables more streamlined policy-driven external control and automation from applications, which ultimately enhances network programmability and simplifies network orchestration. As such, SDN-based design allows for highly elastic networks that can readily adapt to changing business needs.

The first wave of SDN deployment focuses on functionality, but with many innovations and enhancements in data center interconnect technologies, it is time to take a fresh look at more efficient, higher performance SDN deployment options.

In this white paper, we focus on SDN solutions for data centers, which is often an essential part of building cloud, whether it is private cloud or public cloud. It covers the following topics:

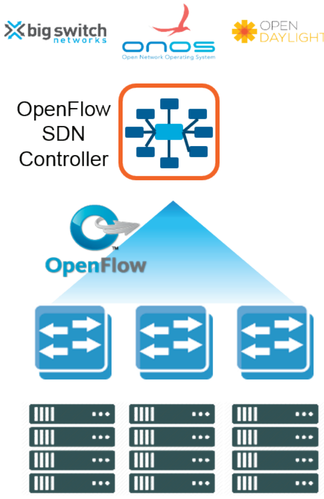
- Overview of the dominant SDN deployment models;
- Mellanox SDN technology highlights, where we discussed the key Mellanox features and products that make SDN deployment more efficient;
- Mellanox SDN solutions for OpenFlow and Overlay SDN deployment models. We show how the Mellanox products and features are put together to deliver total solution in various SDN deployment scenarios and the key benefits the Mellanox solutions deliver.

SDN Deployment Models

Three different deployment models dominate today's SDN landscape:

Device-Based SDN Deployment Model

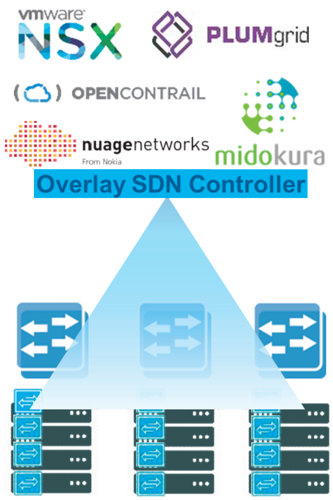
In this model, the SDN Controller uses a south-bound device control protocol to directly communicate policy or forwarding table information to the physical and virtual switching and routing devices. OpenFlow is the most commonly used protocol, and some of the early SDN architectures are based on OpenFlow to decouple control plane from network devices.



Examples of SDN implementations based on this model are BigSwitch's Big Cloud Fabric, Open Networking Lab(ON.LAB)'s ONOS, and Open Daylight (ODL). Beyond OpenFlow, ONOS and ODL also support other southbound protocols such as Netconf and SNMP for device configuration and management.

Essentially for every new flow, all the devices that this flow traverses potentially needs to be programmed to handle the proper flow operations. This model requires the network devices to be OpenFlow-aware, which can sometimes be a challenge when you have legacy networks or a mixture of various generation of network devices.

Overlay SDN Deployment Model

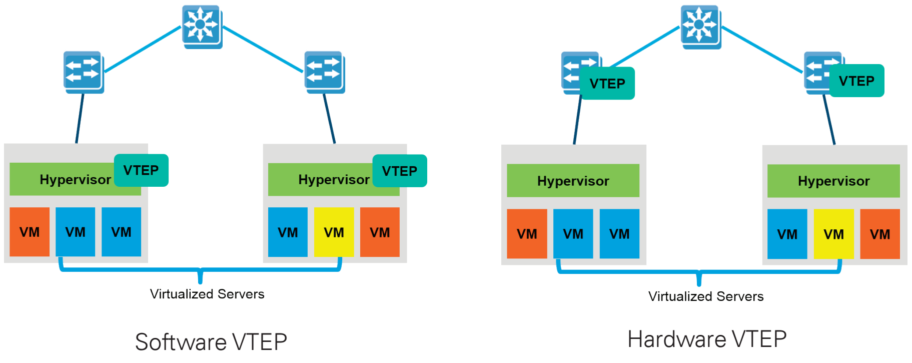


Many customers have an installed base of networking equipment that is not yet OpenFlow-enabled, and doing a network-wide upgrade may not be an option. Overlay approach of SDN deployment model came into being to bring SDN/network virtualization to these customers without requiring forklift network upgrade that can be both expensive and disruptive to business services. Overlay SDN has been the most commonly seen architecture, and mainstream SDN solutions such as VMware NSX, Nuage Networks (Now part of Nokia) VSP, PLUMGrid ONS, OpenContrail and Midokura MidoNet all primarily follow this model.

As the name indicates, in overlay model, logical networks are established through tunnels between endpoints, and these tunnels are overlaid onto an existing physical network. The intelligence about multi-tenancy, network and security policies are pushed to the network edge. Some of the most commonly used tunneling protocols include Virtual eXtensible LAN (VXLAN), Network Virtualization using GRE (NVGRE), and Generic Network Virtualization Encapsulation (GENEVE). In case of VXLAN, the tunnel endpoints are known as VXLAN Tunnel End Points (VTEP). The physical network, or the underlay, becomes the "core" network and its functionalities can potentially be simplified to providing high-performance IP connectivity between these VTEPs. An overlay SDN controller will primarily communicate with the VTEPs, which oftentimes are the virtual switching and routing device residing on servers.

Overlay SDN can be deployed to achieve network virtualization and automation without requiring upgrades of physical networking equipment, more specifically, the network devices that are NOT the VTEPs. Despite its pluses, overlay SDN introduce added complexity when it comes to managing both the overlay and underlay, and correlating information from both layers during troubleshooting.

There are two common ways to deploy VTEP: software VTEP in virtual switches, normally running in server hypervisors; or hardware VTEP in Top of Rack (ToR) switches, and there are tradeoffs between these two approaches. Software VTEP is flexible and conceptually simple, but can impact performance as well as raise CPU overhead on edge devices due to the packet processing associated with the relatively new tunneling protocols that not all server Network Interface Cards (NICs) can offload from CPU. This can be even more pronounced when the applications themselves are virtualized network functions (VNFs) in Network Function Virtualization (NFV) deployment. Hardware VTEPs can often achieve higher performance but pose added complexity on the ToR switch since the ToR switch needs to be VM-aware, maintain a large forwarding table, and performance VM Mac address or VLAN to VXLAN translations.



Software VTEP

Hardware VTEP

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 extra hop or potentially performance bottleneck and the best practice is to use the ToR that the BMS is connected to as hardware VTEP.

Proprietary SDN Solutions

There are other proprietary SDN solutions in the market, such as Cisco Application Centric Infrastructure (ACI), Plexxi and Pluribus. With these solutions, the SDN controller and the SDN switching and routing elements are often tightly coupled. This category of SDN solutions are not as open as the above two, and pose limitations for ecosystem vendors to integrate with them. Mellanox currently works with only open SDN solutions.

Mellanox SDN Technology Highlights

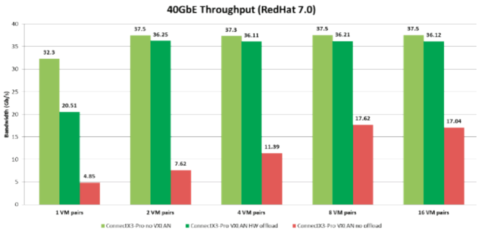
VXLAN Offload on Connect-X NICs

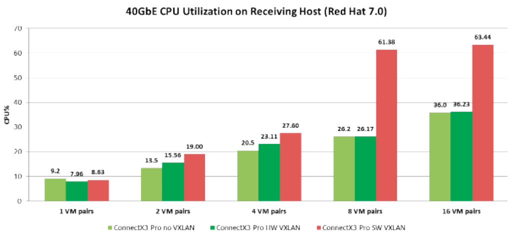
In the good old days when network virtualization was realized through VLAN, achieving line rate performance on the server host is possible because the server can offload some of the CPU-intensive packet processing operations such as checksum, Receive Side Steering (RSS), Large Receive Offload (LRO) etc. into the NIC hardware. This both improves network I/O performance and reduces CPU overhead, ultimately making the infrastructure run more efficiently.

As mentioned in the above section, with overlay SDN, a tunneling protocol such as VXLAN, NVGRE or GENEVE is introduced to encapsulate the original payload. For NICs that don't recognize these new packet header formats, even the most basic offloads stop functioning, resulting in all packet manipulating operations to be done in software in CPU. This can cause significant network I/O performance degradation and excessive CPU overhead, especially when server I/O speed evolves from 10Gb/s to 25, 40, 50, or even 100Gb/s.

Starting from ConnectX-3 Pro series of NIC, Mellanox supports VXLAN hardware offload that includes stateless offloads such as checksum, RSS, and LRO for VXLAN/NVGRE/GENEVE packets. With VXLAN offload, I/O performance and CPU overhead can be restored to similar levels as VLAN.

The following two graphs shows the bandwidth and CPU overhead comparison in three scenarios: VLAN, VXLAN without offload, and VXLAN with offload. VXLAN offload results in greater than 2X throughput improvement with approximately 50% lower CPU overhead.



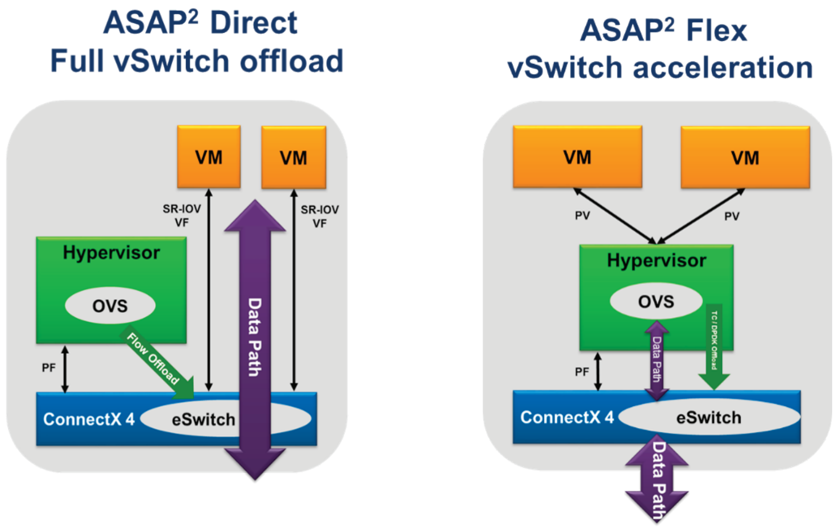


VXLAN Offload is supported at OS/hypervisor kernel level for Linux, Microsoft Hyper-V, and VMWare ESXi, and does not depend on the type of virtual switch or router used.

ASAP^2 (Accelerated Switching and Packet Processing) on ConnectX-4 NICs

Starting from ConnectX-4 series of NICs, Mellanox support VTEP capability in server NIC hardware through the ASAP2 feature. With a pipeline-based programmable eSwitch built into the NIC, ConnectX-4 can handle a large portion of the packet processing operations in hardware. These operations include VXLAN encapsulation/decapsulation, packet classification based on a set of common L2 – L4 header fields, QoS and Access Control List (ACL). Built on top of these enhanced NIC hardware capabilities, ASAP2 feature provides a programmable, high-performance and highly efficient hardware forwarding plane that can work seamlessly with SDN control plane. It overcomes the performance degradation issues associated with software VTEP, as well as complexity issues of coordinating between server and TOR devices in case of hardware VTEP.

There are two main ASAP2 deployment models: ASAP2 Direct and ASAP2 Flex



ASAP² Direct

In this deployment model, VMs establish direct access to Mellanox ConnectX-4 NIC hardware through SR-IOV Virtual Function (VF) to achieve the highest network I/O performance in virtualized environment.

One of the issues associated with legacy SR-IOV implementation is that it bypasses hypervisor and virtual switch completely, and the virtual switch is not aware of the existence of VMs in SR-IOV mode. As a result, SDN control plane could not influence the forwarding plane for those VMs using SR-IOV on the server host.

ASAP2 Direct overcomes this issue through enabling rules offload between the virtual switch and the ConnectX-4 eSwitch forwarding plane. In this case, we use Open Virtual Switch (OVS), one of the most commonly used virtual switch for illustration. The combination of SDN control plane through OVS who communicates with a corresponding SDN controller, and NIC hardware forwarding plane offers the best of both world, software-defined flexible network programmability, and high network I/O performance for the state-of-art speeds from 10G to 25/40/50/100G. By letting the NIC hardware taking the I/O processing burden from the CPU, the CPU resources can be dedicated to application processing, resulting in higher system efficiency.

ASAP2 Direct offers excellent small packet performance beyond the raw bit throughput. Our benchmark shows that on a server with 25G interface, ASAP2 Direct achieves 33 million packets per second (MPPS) with ZERO CPU cores consumed for a single flow, and about 25 MPPS with 15000 flows performing VXLAN encap/decap in ConnectX-4 Lx eSwitch.

ASAP² Flex

In this deployment model, VMs run in para-virtualized mode and still go through the virtual switch for its network I/O needs. But through a set of open APIs such as Linux Traffic Control (TC), or Data Path Development Kit (DPDK), the virtual switch can offload some of the CPU intensive packet processing operations to the Mellanox ConnectX-4 NIC hardware, including VXLAN encapsulation/decapsulation and packet classification. This is a roadmap feature and availability date will be announced in the future.

OpenFlow support on Spectrum Switches

Spectrum is Mellanox's 10/25/40/50 and 100Gb/s Ethernet switch solution 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.

From the ground up, at the switch silicon level, Spectrum is designed to have a very flexible processing pipeline so that it can accommodate programmable OpenFlow pipeline that allow packets to be sent to subsequent tables for further processing and allow metadata information to be communicated between OpenFlow tables. In addition, Spectrum is an OpenFlow-hybrid switch that supports both OpenFlow operation and normal Ethernet switching operation. 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. In addition, Spectrum also provides a classification mechanism to direct traffic within one switch port to either the OpenFlow pipeline or the normal Ethernet processing pipeline.

Here is a summary of the OpenFlow features supported on Spectrum:

- OpenFlow 1.3 Control packet parsing
  - Mapping interfaces to OpenFlow hybrid ports
- Interoperability with Open Daylight and ONOS controllers.
- Set OpenFlow bridge DATAPATH-ID
- Show configured flows by controller
- Supporting flex rules in hardware using ACLs
- Query table features
- Table Actions – Add, Delete, Modify, Priority, Hard Timeout
- Configurable remote controllers:
  - Selective send to controller
- Per port counters
- Per port state:
  - STP
  - Operation
  - Speed

VTEP support in Spectrum Switches

- 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 and 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.
- Layer 3 VTEP gateway that provide 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, and the software to enable it is still under development.

Spectrum is an Open Ethernet switch and can support multiple switch operating system running over it. The Layer 2 VTEP gateway features will first be available in Cumulus Linux over Spectrum, and subsequently in MLNX-OS.

This achieves line rate performance while saving on compute and storage costs.

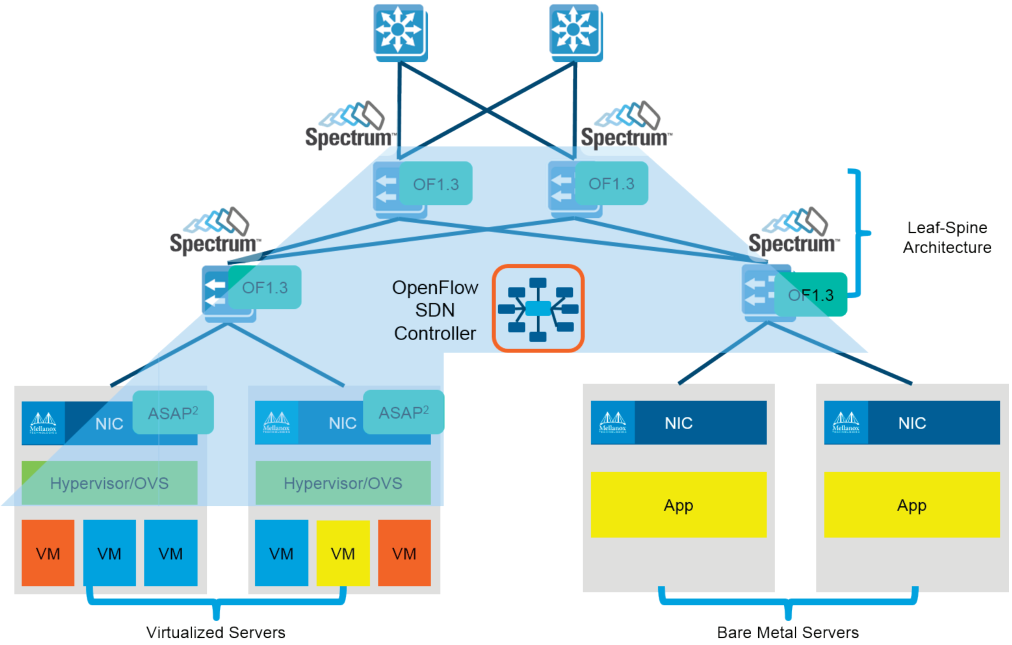
Some of the major storage protocols and platforms are RDMA-enabled today. Examples include iSER (iSCSI over RoCE), SMB Direct, iSER within OpenStack's Cinder, Ceph RDMA, and other.

iSER and SMBDirect performance advantages below:

Building the Most Efficient SDN Networks with Mellanox Interconnect

In this section, we recommend best practice when it comes to building the most efficient SDN networks for OpenFlow and Overlay based models.

Recommended deployment for OpenFlow based SDN networks



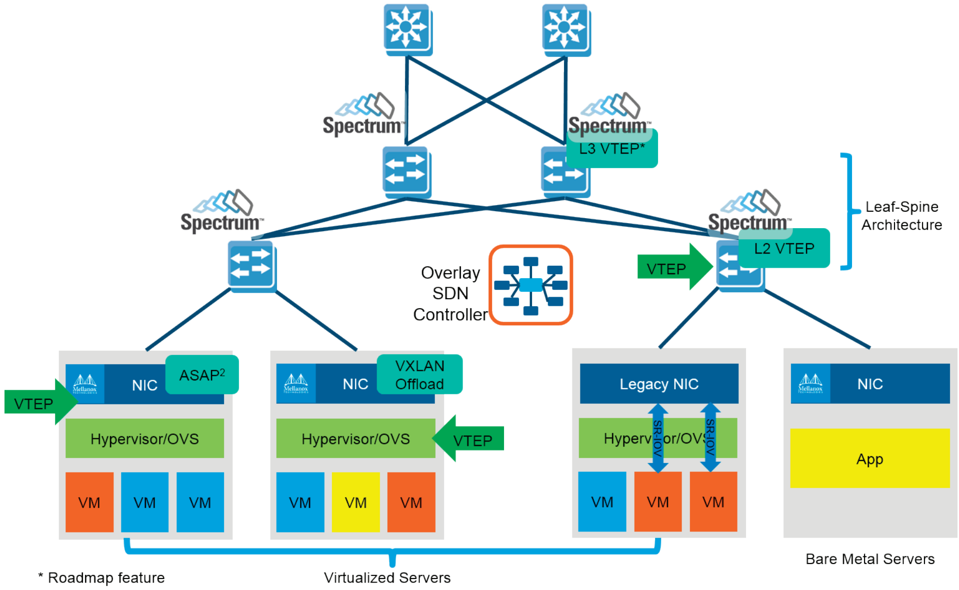
Highlights:

- Leaf-Spine architecture built using Spectrum switches with OpenFlow 1.3 support
- For physical + virtual fabric, leverage ASAP2 to offload virtual switch data plane to Mellanox ConnectX-4 NICs
- Advanced flow monitoring with SFLOW capabilities on Spectrum

Key benefits:

- High performance, line rate performance at any speed from 10G to 100Gb/s, including on virtualized servers
- Most flexible OpenFlow switch implementation
- In-depth visibility into the OpenFlow fabric

Recommended Deployment for Overlay SDN Networks



Highlights:

- Multiple ways for virtualized server deployments:
  - » Virtual switch as VTEP + Mellanox VXLAN stateless offload
  - » Mellanox NIC as VTEP (Leverage ASAP2 to offload virtual switch data plane to Mellanox ConnectX-4 NICs, while keeping SDN control plane operations in virtual switch)
  - » For VMs who needs SR-IOV, and the legacy NIC does not support ASAP2, use Mellanox Spectrum ToR as VTEP
- High-performance hardware VTEP on Mellanox Spectrum ToR for bare metal server or storage provisioning
- (Roadmap Feature) High-performance hardware Layer 3 VTEP on Mellanox Spectrum spine switches for VXLAN routing.
- Advanced underlay fabric monitoring with SFLOW capabilities on Spectrum

Key benefits:

- High performance, line rate performance at any speed from 10G to 100Gb/s, including on virtualized servers;
- Most advanced and future-proof VTEP implementation with flexibility to do VTEP either at NIC or switch level, with potential to extend to Layer 3 hardware VTEP without forklift upgrade;
- In-depth visibility into the SDN underlay fabric to facilitate correlation of stats from both layers and achieve easy troubleshooting.

Conclusion

SDN is an evolving technology, and Mellanox is the only data center networking solution vendor that can provide the most comprehensive, flexible and efficient SDN support through our end-to-end interconnect and associated software.

Mellanox
TECHNOLOGIES

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