WHITE PAPER

July 2015

Mellanox Virtual Modular Switch®

Introduction

As new applications are constantly evolving, data centers must be flexible and future proof to meet the demand for higher throughput and higher scalability, while protecting the original capital investment and without increasing operating costs such as power consumption.

Traditionally, aggregation switching in data centers of Cloud providers, Web 2.0 providers, and large-scale enterprises has been based on modular switches. These switches are usually both expensive to purchase and to operate in addition to being optimized for specific sizes of clusters. They do not provide the flexibility required of today's data centers and have lagged behind the technological progress achieved by the latest solutions.

To overcome this flexibility limitation of modular switches, users are shifting to fixed switches (or top-of-rack switches) to increase the efficiency in data center aggregation.

The Mellanox Virtual Modular Switch® solution (VMS), comprised of Mellanox 10, 40, 56, and 100GbE fixed switches, provides an ideal, optimized approach for aggregating racks. When compared to competing offerings, VMS excels with higher flexibility, better scalability and energy efficiency, while future-proofing the investment and reducing expenses.

Mellanox switches leverage the unique advantages of SwitchX®-2 and Spectrum™ chips. It is the highest density switching chipset, supporting 36 40/56GbE ports (SwitchX-2) at a low power consumption of 1.5W per port and 32 100GbE ports (Spectrum) at a power consumption of only 4W per port. Configurations of up to 64 10/25GbE ports are also available. The chipset provides the highest bandwidth in the industry, up to 6.4Tb/s of non-blocking switching and routing at a very low latency of under 3000ns for all packet sizes.

Considerations for Data Center Aggregation Switching

When defining a solution for data center aggregation switching, several aspects should be considered:

- **Scalability** – While a data center has a specific cluster size at the time of definition, it should be flexible enough to grow over time. Offerings that do not allow an increase to port count or that require major investment in equipment for even a modest increase in port count should not be considered.

- **Resiliency** – Aggregation switching must recover gracefully from hardware failures to minimize the impact on data center performance. If such a failure causes bandwidth degradation of 50 percent or more for a long period of time, it will result in high maintenance expenses. Offerings that provide at least 90 percent of the optimal bandwidth at all times are preferred.
- **Performance** – The aggregation layer must perform in a non-blocking manner. In addition, low latency is desired to improve application efficiency.
- **Cost of equipment** – Keeping costs low, both at first installation as well as during any future upgrade, is of utmost importance.
- **Standard technology** – Implementation that does not rely on proprietary vendor features is preferred. Proprietary features tend to increase complexity and limit upgradability. Standard technology scales easily and is easier to maintain.
- **Energy efficiency** – The low energy consumption of Mellanox switches allows more servers to be populated in the racks. Each Mellanox switch can save up to 250W compared to competing fixed switches.
- **Ease of use** – The switches should be based on standard management tools and utilities and be managed from a single, centralized management utility. Using proprietary features or distributed management tools limits flexibility, increases maintenance efforts and costs, and severely impacts upgradability.

Virtual Modular Switch Architecture - Dual-Tier 40/56/100GbE Aggregation

To address the need to scale out over time, a well-designed data center should require minimal redesign and as little oversubscription and bottlenecks as possible.

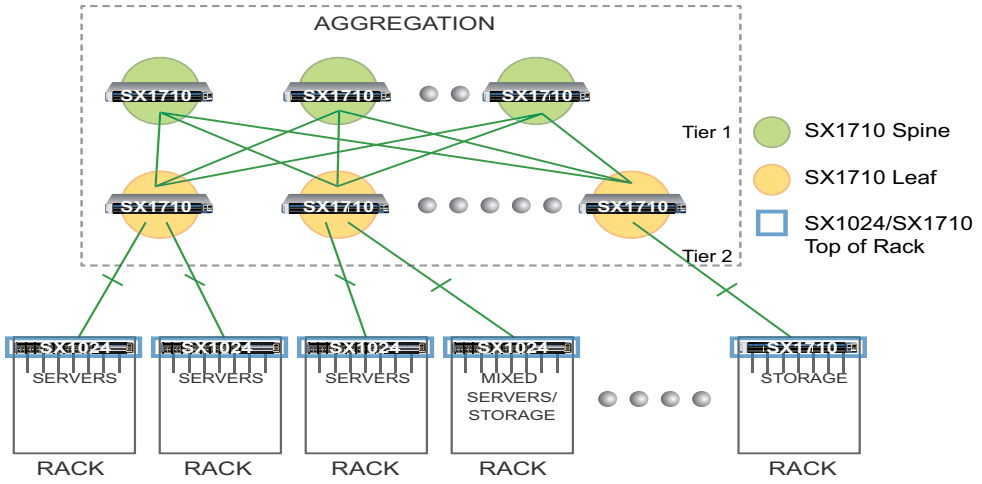

Figure 1. *40/56GbE VMS Architecture*

The VMS dual-tier leaf-spine topology, illustrated in Figure 1 and Figure 3, is commonly adopted for many data centers. In this topology, fixed switches build a dual-tier aggregation layer, which yields a significant data center scale out. The 1st-tier aggregation switches (spines) connect to the 2nd-tier switches (leafs), and the 2nd-tier switches interconnect with both the 1st-tier switches and the data center racks.

Utilizing the high density SX1710 switches, with 36 40GbE ports each and based on a single chip (Mellanox SwitchX-2), the 40/56GbE VMS-based data center can scale from a few ports up to 648 ports for top-of-rack (ToR) connections, up to five times more than the typical capabilities of the competition. This is achieved by connecting 18 SX1710 switches in the 1st-tier and 36 switches in the 2nd-tier.

A significant additional advantage is achieved by utilizing the 56GbE capability of the SX1710 ports as the interconnect speed between the 1st and 2nd tiers of aggregation switches. When operating the interconnect at 56GbE, fewer ports are needed to provide non-blocking performance. This frees up more ports for ToR connections. This configuration allows for up to 16 tier-1 switches and up to 36 tier-2 switches. Each 2nd-tier switch can connect 20 ports toward ToR (at 40GbE) and one port toward each of the 16 1st tier switches (at 56GbE). This provides non-blocking performance for up to 720 40GbE ports.

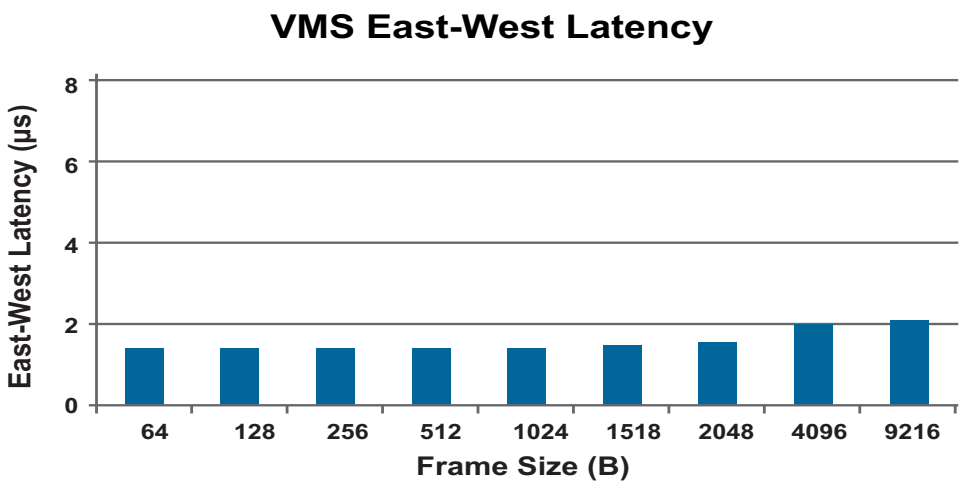

Figure 2. *VMS Latency*

The non-blocking operation is coupled with ultra-low latency. VMS operates at "zero jitter" latency regardless of the traffic profiles or packet sizes. As shown in Figure 2, the VMS east-west latency is up to 2 microseconds, compared to up to 30 microseconds for different packet sizes by competing offerings.

100GbE VMS is similar in its architecture. Utilizing the high density SN2700 switches, with 32 100GbE ports each and based on a single chip (Mellanox Spectrum), the VMS-based data center can scale from a few ports up to 512 ports for top-of-rack (ToR) connections. This is achieved by connecting 16 SN2700 switches in the 1st-tier and 32 switches in the 2nd-tier.

Similar to the 40GbE VMS, the non-blocking operation is coupled with ultra-low latency. VMS operates at "zero jitter" latency regardless of the traffic profiles or packet sizes. In the case of the SN2700, the VMS east-west latency is under 1 microsecond, compared with up to 30 microseconds for various packet sizes by competing offerings.

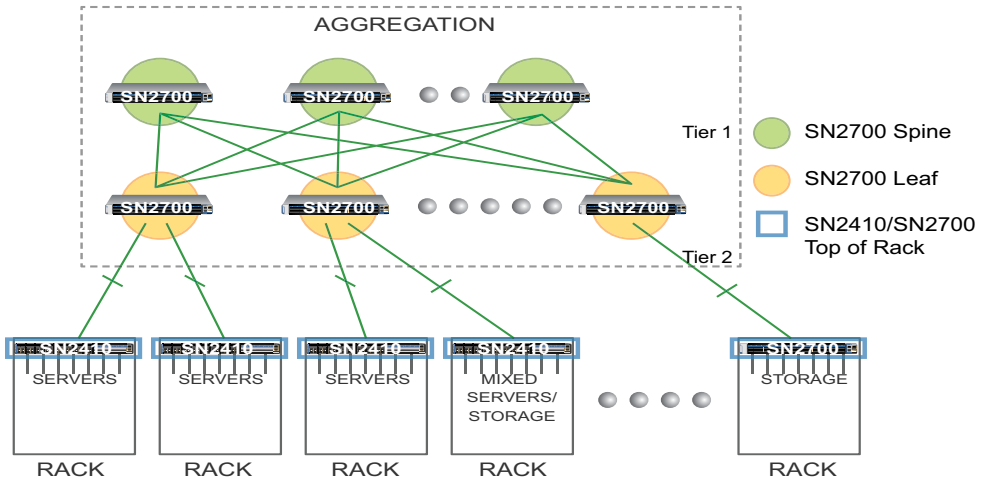

Figure 3. *100GbE VMS Architecture*

VMS Configuration and Management

Mellanox VMS is practically a small scale IP network and therefore is managed and configured using standard IP protocols. These protocols contribute to the ease-of-use of VMS and provide a robust, standard, and resilient solution.

Each of the VMS switches runs the Open Shortest Path First (OSPF) protocol or the Border Gateway Protocol (BGP). The OSPF/BGP protocol gathers link state information from the switches (routers) in the data center and constructs a topology map of the network to determine the IP routing tables. By building the routing tables, routing loops are prevented. It automatically detects topology changes such as link failures and updates the routing structure accordingly within seconds.

To complement the OSPF/BGP routing decisions, the switches also run Equal-Cost Multi-Path routing (ECMP), which provides dynamic load balancing between routes of equal cost. By leveraging ECMP, congestion within the VMS and in the data center is avoided.

Configuration of the VMS infrastructure is performed by the VMS Wizard automation software andapplications such as Puppet (http://www.puppetlabs.com/). These applications automate the configuration and provisioning of the switches, such as software upgrades, VLAN provisioning, port configuration, OSPF and BGP, and provide an easy way to scale the VMS. After configuration, the VMS Wizard and Puppet can report errors and mismatches.

VMS CAPEX and OPEX Savings

Mellanox VMS excels in saving money for its users, both in capital and in operating expenses.

Since VMS is scalable, only the required number of switches at the time of definition needs to be installed. From then on, the VMS scales in a pay-as-you-grow manner and provides additional savings compared to modular switches. With VMS, there is no need to purchase extra equipment to compensate for future growth.

VMS scalability also contributes to its best-in-class resiliency. Upon failure of a switch within VMS, the spine-leaf topology maintains the data flow at over 90 percent of the total bandwidth, depending on the VMS size. This allows the failure to be fixed and the malfunctioning unit to be replaced with almost no impact on the performance and behavior of the data center. This is a tremendous safeguard against financial loss from performance degradation during such a failure.

VMS is the most energy-efficient solution in the industry and provides the best power-per-port performance. For example, the VMS solution for 192 nodes of 40GbE consumes only 1,250W, or 6.5W per 40GbE port, delivering even further OPEX savings.

Mellanox VMS Advantage

The following table compares VMS against commercially available products in a 192 40GbE port fabric:

Products	Fixed Switches by Vendor A	Fixed Switches by Vendor B	VMS Solution	VMS Advantage
Scalability Beyond 192 Nodes	No	Yes	Yes	
Power Consumption (KW)	7.7-19	9.6-19.8	1.25	Reduces power consumption by 80-90%
Street Price ($K)	550-720	1,240-1,500	225	Reduces cost by 60-85%
Additional Software Fees	Yes	Yes	No	No software fees
Max. East-West Solution Latency (us)	29-30	3	< 2	Reduces latency by up to 93%
Bandwidth After Hardware Failure	0-50%	83-97%	93%	

The following table compares VMS against commercially available products in a 512 100GbE port fabric:

Products	Modular Switches by Other Vendors	Fixed Switches by Other Vendors	VMS Solution	VMS Advantage
Power Consumption (KW)	14.9	23.3	1.25	Reduces power consumption by up to 50%
Performance	Packet loss	Packet loss	Non-blocking	Guaranteed performance
Additional Software Fees	Yes	Yes	No	No software fees
Max. East-West Solution Latency (us)	Varies	Varies	< 1	Guaranteed performance

Please contact a Mellanox representative for a detailed competitive analysis of VMS.

Mellanox provides a complete VMS solution, which includes Ethernet switches, software application and management software stack, as well as copper cables or fiber cables.

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

15-691WP
Rev 1.1