WHITE PAPER

May 2013

SX1036: The Ideal 40GbE Aggregation Switch

Introduction

As new data center applications are used, new solutions are required to meet the demand for higher throughput, without increasing the power consumption or cost.

Mellanox's SX1036 36-Port 40/56GbE Switch System provides the highest-performing fabric solution in a 1U form factor by delivering over 4Tb/s of non-blocking throughput.

The SX1036 switch can be configured to provide 36 ports running at 40GbE or 56GbE, providing the highest available speed on the market and allowing for higher-density, faster, and more efficient data centers. The SX1036 can also be configured to provide up to 64 1GbE or 10GbE ports. In all configurations, the SX1036 operates at an ultra-low latency: 230nsec for 40GbE and 56GbE and 250nsec for 10GbE. The SX1036 also features very low power consumption, typically 2.3 Watts per port at 40GbE.

The SX1036 switch system is ready for Mellanox Open Ethernet, which provides the ability to terminate the dependency on the operating system and protocol stack of the switch vendor. With Open Ethernet, customers gain control over the switch, and use it as a customizable platform, optimized for their special requirements. This customization can allow, for example, improving the behavior of specific scenarios or shortening the turnaround time of implementation of new specifications.

Single 40GbE Aggregation Switch

Small data centers typically use a single switch as the aggregation switch for several racks. Most available top-of-rack (ToR) 40GbE switches use four ports of 40GbE as the uplink toward aggregation. Other available switches mostly have 16 40GbE ports, allowing up to 4 racks to be connected. With typical ToR switches each connecting to 48 servers via 10GbE links, the maximum scale-out of such a data center allows for 192 servers.

Using Mellanox's SX1036 higher-density switch allows for a 2.25 scale-out of data center size compared to other solutions. With its 36 ports running 40GbE, the aggregation switch can connect up to 9 racks, effectively allowing for a 432-server data center.

Leaf-Spine Architecture In Data Centers

A key requirement in many data centers is the ability to scale out over time as more racks are added and the total traffic increases. In a well-designed data center, scaling up should involve minimal redesign and as little oversubscription and bottlenecks as possible.

The leaf-spine topology illustrated in Figure 1 is commonly adopted for many data centers implementing aggregation using 40GbE hardware. In this topology, ToR switches serve as the leafs and the aggregation switches as the spines. Each leaf is connected to the servers in its rack and also to each of the spine switches.

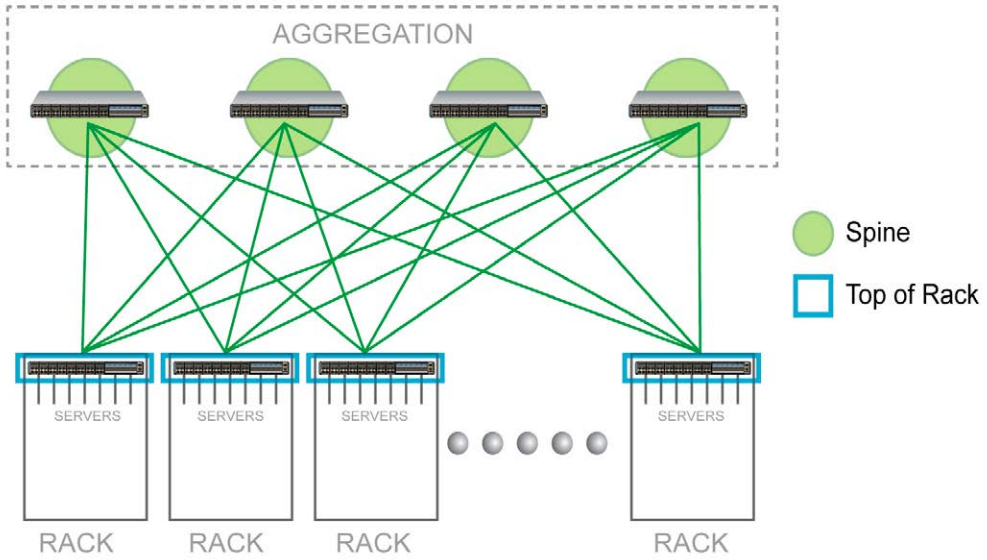

Figure 1. *Leaf-Spine Architecture*

The ability to scale greatly depends on the number of Ethernet ports of the switches. The number of ports in a spine switch dictates the maximum number of leaf switches. Correspondingly, the number of ports in each leaf switch that are allocated for spine connections dictates the maximum number of spine switches. The obvious conclusion is that more switch ports enable higher scalability.

Mellanox Single-Tier 40GbE Aggregation

In a single-tier 40GbE aggregation layer such as the one described in Figure 1, ToR switches use four 40GbE ports as the uplink toward the aggregation layer, each link connecting to one of the four switches comprising the aggregation layer (spines). With typically 16 ports per 40GbE switch, this allows up to 16 leaf switches. As the typical ToR switch connects to 48 servers via 10GbE links, such a data center scales up to a maximum of 768 servers.

Using Mellanox's SX1036 switch allows for a 2.25 scale-out of data center size compared to other solutions. With 36 leaf switches connecting to 48 servers each, an SX1036-based data center can scale up to 1728 servers.

Dual-Tier 40GbE Aggregation

Building a dual-tier aggregation layer yields a significant data center scale out. Figure 2 illustrates a dual-tier topology, where the 1st-tier aggregation switches (spines) connect to the leafs and the 2nd-tier switches interconnect with 1st-tier switches and connect to other leafs.

With 16-port 40GbE switches, each spine on one layer connects over one link to each of the 8 spines on the other layer, leaving 8 40GbE links to connect to leafs, or 128 40GbE ports to connect to leafs. Each leaf needs four 40GbE connections toward (spine) aggregation, therefore the architecture allows for 32 racks. As such, the maximum number of servers this data center can have is 1536.

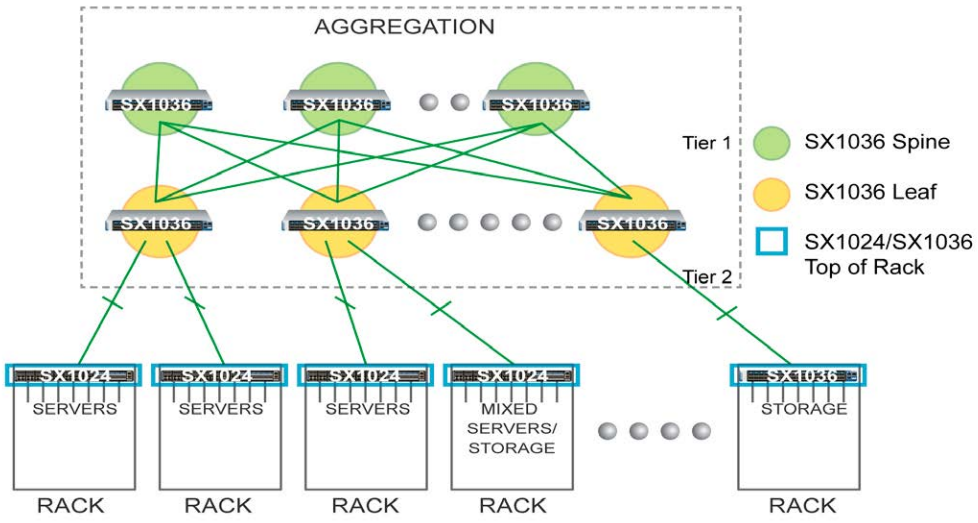

Figure 2. *Dual-Tier Architecture*

Using SX1036 switches with 36 40GbE ports, the data center can be scaled up 4.5 times in the number of servers. This is achieved by having 18 switches in the 1st-tier and 36 in the 2nd. Each switch in tier 1 interconnects with all 36 switches in tier 2. This way, tier-2 switches each remain with 18 ports to ToR (leaf) switches, or 4 leafs each. In other words, the data center can scale up to 144 ToR switches, for a total of 6912 servers.

56 GbE Interconnection In Dual-Tier Aggregation Topologies

The SX1036 has thus far been considered with all its ports running 40GbE. There is a big advantage in utilizing the 56GbE capability of its ports as the interconnect speed between the 1st and 2nd tiers of aggregation switches. When operating at 56GbE, fewer ports are needed for the interconnection, freeing up more ports for ToR connections.

The aggregation switches must be non-blocking. When all ports operate at 40Gb/s, the 1st-tier switches must be configured with 18 ports toward ToR and 18 ports toward the 2nd tier. When the interconnects are 56GbE, the 1st-tier switches can be configured with 20 ports toward ToR (at 40GbE) and 16 ports toward the 2nd tier, providing non-blocking performance.

With 36 2nd-layer switches operating at 40GbE on each port, and with connections to 180 ToR switches, leveraging Mellanox's SX1036 56GbE capabilities allows for scaling up the data center size significantly to include 8,640 servers.

SX1036 as the Top-of-Rack Switch

The high port count of SX1036 and its support for both 10GbE and 40GbE make it an excellent ToR switch that provides a low oversubscription ratio.

With the increasing need for bandwidth, more and more servers integrate 40GbE network interface cards. To scale up the capacity of the data center so as to meet the bandwidth requirements without an explosion in complexity and footprint, increasing the density of 40GbE ports of ToR switches becomes a must.

Most 40GbE switches available on the market can only support 12 servers and 4 uplink ports concurrently (3:1 oversubscription).

In comparison, the SX1036 can support, for similar configurations, up to 24 servers and 12 uplink ports with a lower oversubscription of 2:1, making it a superior choice for this domain of applications.

The SX1036 is also ideal as the ToR switch in mixed racks, with 40GbE server links and 10GbE server links (storage and compute), and gives the flexibility to align the port split in any way desired with the use of breakout cables.

Summary

Mellanox's SX1036 Ethernet switch is the ideal switch for a variety of applications. With 36 ports that can operate at 1GbE, 10GbE, 40GbE or 56GbE, a 4Tb/s capacity, low power consumption, and compact size, SX1036 is a perfect fit from the small-scale to the high-scale data center aggregation.

Moreover, with the increased bandwidth at the server racks, SX1036 makes the ToR switching very efficient with the lowest oversubscription.

Solution Components

Ordering Part Number	Port
MSX1036B-1SFS	SwitchX®-2 based 36-port QSFP 40GbE 1U Ethernet Switch, 36 QSFP ports, 1 PS, Standard depth, PSU side to Connector side airflow, Rail Kit and RoHS6
MSX1036B-1BRS	SwitchX®-2 based 36-port QSFP 40GbE 1U Ethernet Switch, 36 QSFP ports, 1 PS, Short depth, Connector side to PSU side airflow, Rail Kit and RoHS6

Mellanox Ethernet Switch Portfolio

Switch	Description
SX1036	Ideal ToR/Core switch using 36 40/56GbE from server to aggregation
SX1024	Ideal ToR/Core switch using 48 10GbE from server and 12 40/56GbE to aggregation layer
SX1012	Ideal ToR switch for storage and database, using 12 40/56GbE or 48 10GbE ports, compact size to fit 2 x SX1012 in 1 RU (19" racks)

Mellanox TECHNOLOGIES

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

15-589 WP
Rev 1.0