WHITE PAPER

May 2013

SX1024: The Ideal Multi-Purpose Top-of-Rack 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 SX1024 is an optimal top-of-rack switch (ToR) with 48 1/10GbE SFP+ ports and 12 QSFP interfaces which can operate at 1/10GbE, 40GbE or 56GbE speeds. The SX1024 enables non-blocking throughput between the rack and the aggregation layer, which makes it the optimal-performing switch for storage environments, high performance computing, Hadoop, and any other enterprise data center.

The SX1024 switch supports data transport at various speeds (1GbE, 10GbE, 40GbE or 56GbE), and it can operate in a rack that includes servers and storage nodes from different generations, allowing for server upgrade without the need to change the ToR. The SX1024 operates at ultra-low sub-microsecond latency and at very low power consumption.

Highest Server Density In a Rack

60 Servers with 4 x 40 GbE Uplinks

With every new generation of technology, server sizes shrink. Whereas in the past all servers occupied the entire rack width (19"), it is quite common today to find different racking solutions that allow fitting two servers on the same rack shelf, yielding a total of 56, 60 or more servers in the same rack.

Since most commercially available ToR switches have 48 10GbE ports, connecting more than 48 servers in this rack would require two ToR switches. These two switches increase the rack cost, latency and energy consumption, and occupy space that could have been allocated for servers.

This problem is solved by the port flexibility of the SX1024 switch. By simple configuration, its 40GbE ports can be configured to 10GbE and support more servers. For example, the SX1024 can be configured with 60 ports of 10GbE and 4 ports of 40GbE. This configuration connects 60 servers with an uplink of 160Gb/s, or oversubscription of 3.75:1. If configuring the four uplink ports to operate at 56GbE, the oversubscription ratio improves to 2.68:1.

Alternatively, the SX1024 can be configured with 56 ports of 10GbE and 8 ports of 40GbE. This configuration connects 56 servers with an uplink of 320Gb/s, or oversubscription of 1.75:1.

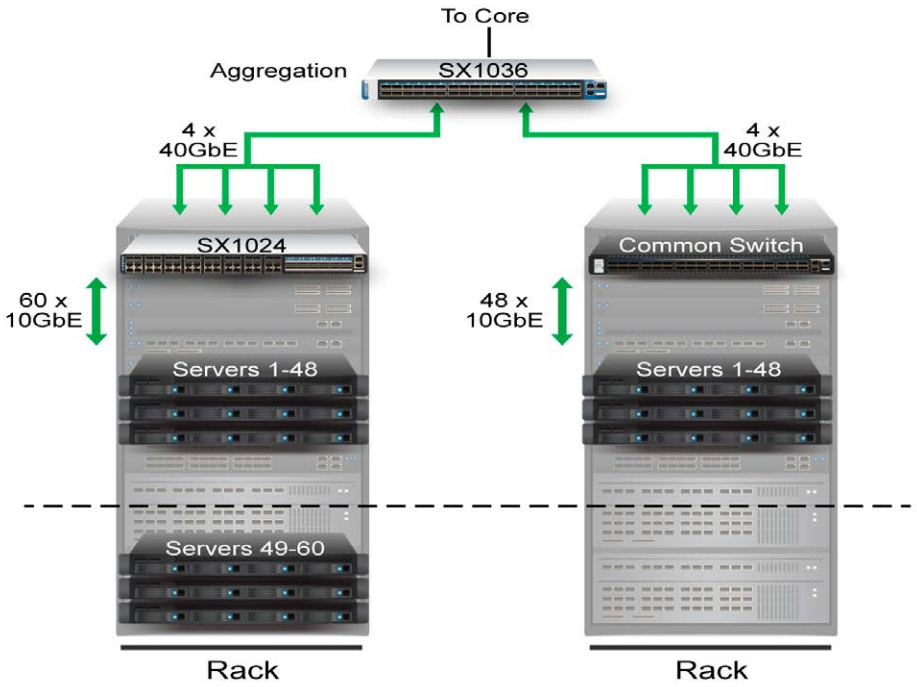

Figure 1. *SX1024 Connects More Servers Through More 10 GbE Links*

A listing of all possible port configurations can be found in the following table.

40 GbE Ports	12	10	8	6	4	2	0
10 GbE Ports	48	52	56	58	60	62	64

Storage in a Rack Enabler

Another measure for the scalability of a data center (beyond size growth) is its flexibility to upgrade or repurpose servers and storage in the same rack without compromising on the available bandwidth reserved for the uplinks.

While 10GbE port connections are very common in racks, with the trend of deploying JBODs (Just a Bunch of Disks) in a rack the ToR now needs to serve 40GbE storage nodes in addition to the 10GbE compute servers. Since the population of servers in the rack is often a gradual process, the ToR switch flexibility to support both 10GbE links and 40GbE links becomes a necessary feature, allowing for a wider selection of servers or storage without any concern for the link speeds.

Most available 10GbE switches do not support 40GbE for server connectivity. This necessitates buying a 40GbE ToR switch, even if only a single 40GbE server port exits in the rack.

Mellanox SX1024 enables the required flexibility for gradual rack population. In its default port configuration it provides 48 10GbE SFP+ ports and 12 40GbE QSFP ports. Several of the 40GbE ports can be used to connect newer generation servers while still providing a better oversubscription ratio than competitor switches. For example, SX1024 allows for building a rack with 44 10GbE server ports and 4 40GbE storage ports, representing maximum traffic of 600Gb/s. With this ToR, the uplink bandwidth of the rack will be 320Gb/s (8 ports of 40GbE), which is an oversubscription of under 2:1.

To support the same configuration of 10GbE and 40GbE server ports using other common market switches, two ToR switches would be required – one for 10GbE connectivity and another for 40GbE connectivity.

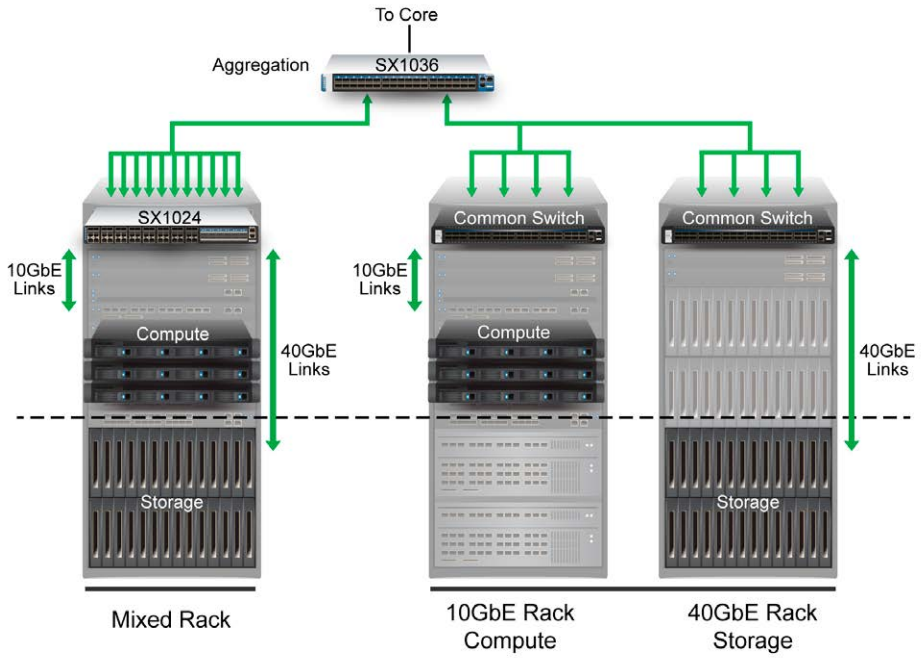

Figure 2. *SX1024 Enables Mixed Racks of Servers and Storage*

Note that for racks with only 40GbE servers, Mellanox offers the SX1036 Ethernet switch system, which can connect up to 36 40GbE servers.

Non-Blocking Rack Implementation

The common server rack size is between 42 rack units (RU) and 52RU. Typically, each of the servers occupies 1RU in height, for a total of 40 servers or more. The ToR switch is connected to each of the servers, aggregates the traffic from all of them, and provides the uplink connectivity toward the aggregation layer of the data center.

Since many of the available servers have a 10GbE port, the common ToR switch has 48 ports that are used for these connections.

A common implementation of ToR switches allows for blocking oversubscription of the rack traffic. While the maximum traffic from the servers through such a switch can reach 480Gb/s (48 x 10Gb/s), the available switch uplink connectivity provides lower bandwidth, typically up to 160Gb/s. The 160Gb/s uplink bandwidth is achieved either by operating 4 ports of 40GbE or 16 ports of 10GbE.

Such oversubscription might cause data loss in the ToR switch, which can only be recovered by retransmissions or re-computation. In other scenarios, the oversubscription adds latency to the traffic flow, which results in lower efficiency in the data center.

Using Mellanox's SX1024, this oversubscription can be eliminated. The SX1024 has 12 uplink ports that operate at 40GbE, for a total uplink bandwidth of 480Gb/s that is equal to the server link bandwidth. For high performance environments ("many-to-less" scenarios), the SX1024 can leverage Mellanox 56GbE technology for under-subscribed wider pipes.

Based on the properties of a specific data center, the SX1024 switch can be installed first in an oversubscribed environment (for example, to connect 8 uplink ports for oversubscription of 1.5:1). As the data center traffic increases, the additional uplink ports can be connected to provide non-blocking performance. This extends the life span of the rack, since the data center scales up with no need to replace the ToR switch.

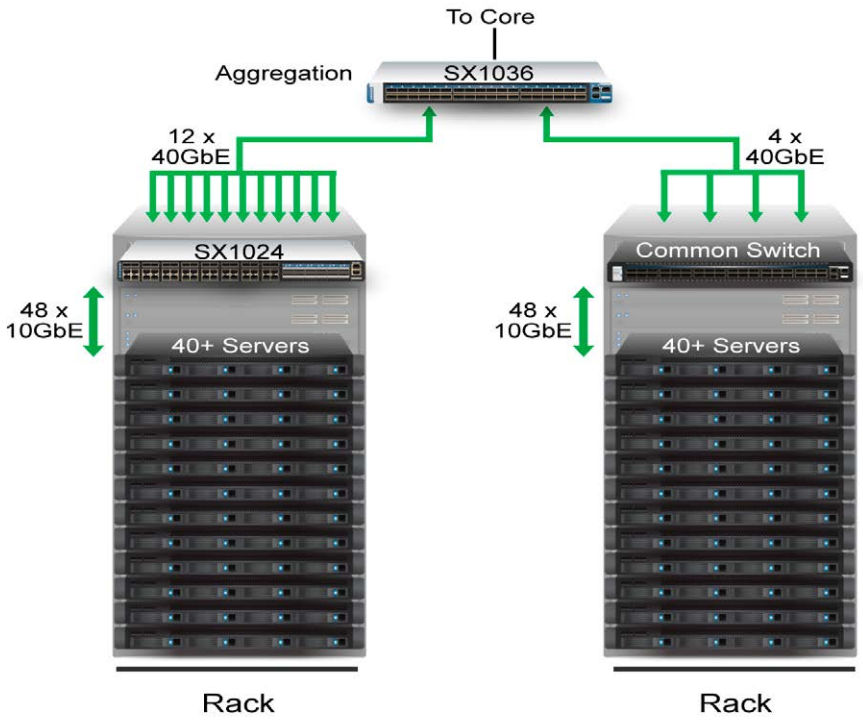

Figure 3. *Non-Blocking Setup of SX1024*

56GbE Uplink Ports

The SX1024 has thus far been considered with its uplink ports running 40GbE. There is, however, an advantage in utilizing the 56GbE capability of its ports as the uplink speed to the data center aggregation. When operating at 56GbE, the 40% higher bandwidth can be used to further reduce oversubscription or to free more ports for server connectivity. Activation of the 56 GbE mode is performed via a software-only upgrade, with no change of equipment.

In the first example above, with four uplink ports of the SX1024 configured to operate at 56GbE instead of 40GbE, the oversubscription ratio for connecting the 60 10GbE posrts (servers) improves from 3.75:1 to 2.68:1.

In the second example, by configuring 56 ports of 10GbE (56 servers) and 8 uplink ports of 40GbE, the oversubscription of the rack is 1.75:1. By configuring the uplink ports to 56GbE mode, the oversubscription is improved to only 1.25:1.

Summary

Selecting the Right ToR Switch

When installing a new rack, several ToR switch properties must be considered. A new rack may host servers with 10GbE ports and should be ready for newer servers with higher throughput and 40GbE ports. It needs to be ready for higher bandwidth connection between the rack and the aggregation and to connect more servers over time.

Mellanox's SX1024 Ethernet switch is the ideal ToR switch and brings flexibility, scalability and future-proofing. Its ability to connect up to 60 servers at different connection speeds (1GbE, 10GbE or 40GbE) and the non-blocking uplink capacity makes it the best fit for a variety of rack architectures and the best means to protect the initial investment.

Through a software upgrade, the SX1024 switch system is ready for adding InfiniBand functionality and enabling a Virtual Protocol Interconnect (VPI) gateway.

The SX1024 switch system is also 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, improved behavior of specific scenarios or shorter turnaround time to implement new specifications.

Mellanox 48-Port 10Gb and 12-Port 40/56Gb Ethernet Switches and Cables Solution Components

Ordering Part Number	Description
Switch Systems	
MSX1024B-1BFS	SwitchX®-2 based 48-port SFP+ 10GbE, 12 port QSFP 40/56GbE, 1U Ethernet Switch, 1PS, Short depth, PSU-side to Connector-side airflow, Rail kit and RoHS-6
MSX1024B-1BRS	SwitchX®-2 based 48-port SFP+ 10GbE, 12 port QSFP 40/56GbE, 1U Ethernet Switch, 1PS, Short depth, Connector-side to PSU-side airflow, Rail kit and RoHS-6
Cables and Modules	
MC3309130-002	Passive copper cable, ETH 10GbE, 10 Gb/s, SFP+, 2m
MC3309130-003	Passive copper cable, ETH 10GbE, 10 Gb/s, SFP+, 3m
MC2207312-010	Active fiber cable, VPI, up to 56Gb/s, QSFP, 10m
MAM1Q00A-QSA	Cable module, ETH 10GbE, 40Gb/s to 10Gb/s, QSFP to SFP+
MC3208011-SX	Optical module, ETH 10GbE, 1Gb/s, SFP, LC-LC, SX 850nm, up to 500m
MC3208411-T	Module, ETH 1GbE, 1Gb/s, SFP, Base-T, up to 100m

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

15-628WP
Rev 1.0