SOLUTION BRIEF

Mellanox Hadoop Cluster Template Architecture

Mellanox provides a template architecture for building a scalable Hadoop cluster, based on the following characteristics.

1. Cluster can scale to more than 3,500 nodes
2. Using RDMA based interconnect for best performance
3. Top of Rack (ToR) Switches up-link bandwidth exceeds 160Gb/s
4. Installation allows customers to scale without re-cabling the initial setup or requiring cluster down time

Solution Overview

Mellanox's SX1036 Ethernet switch is based on Mellanox's industry leading SwitchX® silicon. SwitchX® is a high performance, high density and highly configurable switch silicon, supporting up to 64 ports of 10GbE, 36 ports of 40GbE or 36 ports of FDR 56Gb/s InfiniBand. The SX1036 is the building block used in this reference architecture.

The proposed solution is based on three layer, fat-tree construction. The upper two layers, the core and aggregation layers are built as a single logical group called "Core Switch Group"(CSG). The CSG is constructed from 50 units of SX1036 switches connected at 40GbE data rate. The total number of ports provided by the CSG is 576 ports each operating at 40GbE. These 576 ports are used as the aggregation layer for the ToR switches, and based on the level of oversubscription in the ToR can extend the solution to over 34,000 nodes.

Building a fat tree architecture from discrete components provides an additional layer of resiliency, protecting the cluster from a single point of failure scenario.

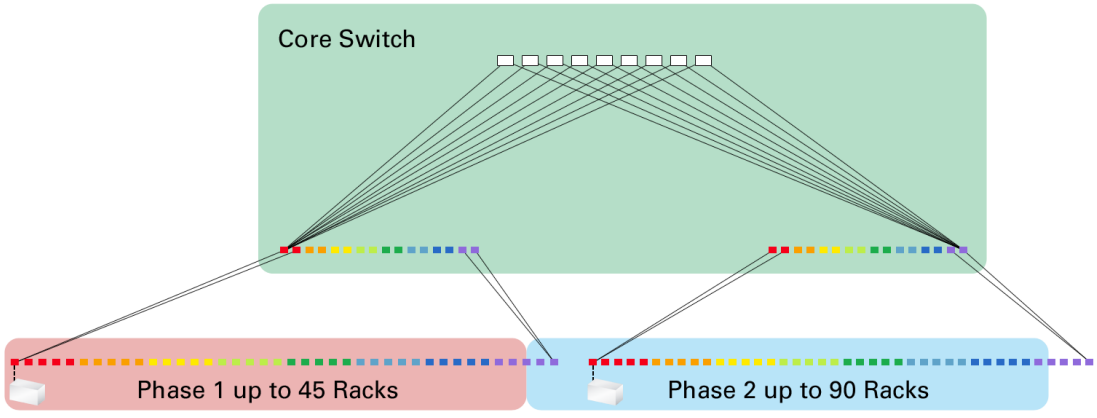


Figure 1

This solution brief calls for installation of the CSG at the initial stages. Typical installation of CSG can be accommodated within 2 standard 42 U Racks.

Unstructured Data Accelerator (UDA) to Harness RDMA Capabilities

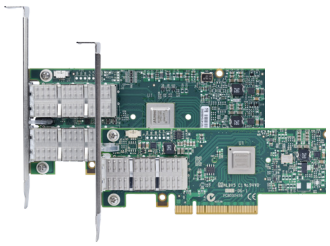
Using fat pipes for data transfer is imperative for MapReduce frameworks. Heavy data transfers take their toll on servers' CPU, reducing the effective number of cycles used for data processing.

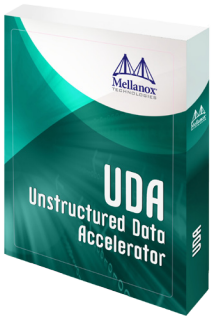
RDMA (Remote Direct Memory Access) technology releases the utilization bottleneck. RDMA provides data transfer with minimal CPU intervention and lower latency connectivity between the servers' memory.

Mellanox UDA is an Apache Hadoop, plug-in utilizing RDMA for fast and effortless data transfer.

Mellanox ConnectX®-3 adapter cards provides hardware support for the RDMA connectivity required by UDA. ConnectX®-3 VPI cards support both Ethernet and InfiniBand at various data rates from 56Gb/s to 10GbE and 1GbE.

UDA shows over 40% reduction in execution time, 150% improvement in CPU utilization and up to 40% reduction in disk writes.





Building a Scale-out Cluster for Apache Hadoop

The following reference design constructs a 3,840 node cluster in two phases, the first phase includes up to 1,920 nodes, connecting the CSG to 45 SX1036 ToR switches, with each ToR connected with 6X 40GbE links providing 240GbE bandwidth to the compute node layer. Servers using ConnectX®-3 adapter cards are connected with a 10GbE link to the ToR switches. Each ToR can support up to 58 servers.

Figure 1 shows the CSG layer construction and the ToR connectivity to the aggregation layer.

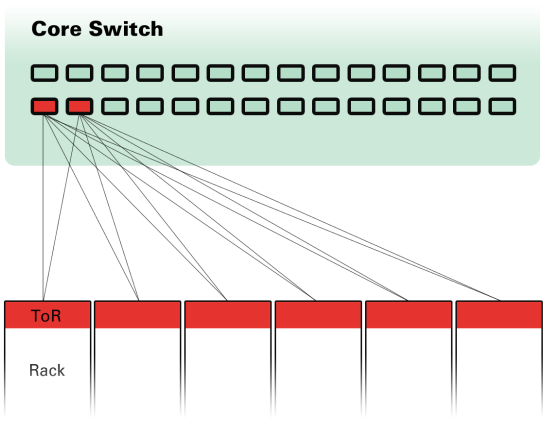
Figure 2 shows a sample of 40 servers connected to the ToR and the up-link connections from the ToR to the aggregation layer.

The installation of the second phase is simple; there is no need to reconfigure the CSG and additional racks can be directly connected without shutting down the existing cluster.

Reducing the number of servers connected to the ToR lowers the oversubscription and provide more BW per server to the CSG. Connecting 24 servers to each ToR in the configuration (Figure 1) provides a full non-blocking fat-tree fabric.

Mellanox's SX1036 utilize QSFP+ ports to connect either 40GbE or 4x10GbE breakout cables. Usage of breakout cables will enable connectivity of 4 SFP+ (10GbE) to a single QSFP+ port (40GbE).

Mellanox provides state-of-the-art cabling solutions including copper and optical cables and modules to ensure smooth installation and long lasting reliable operation.



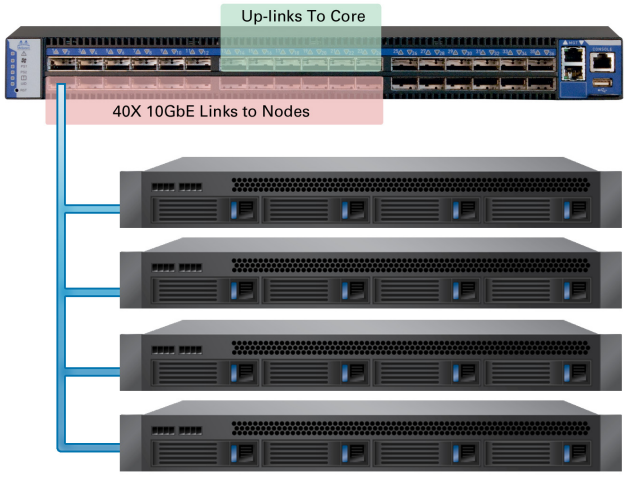


Figure 2

Power Savings Without Performance Compromise

One of the prime concerns in building a large scale cluster is its power consumption.

A full populated Mellanox SX1036 connected at 40GbE mode consumes typically 100W for the complete system.

Applying the unit power to the CSG, provides a fully non-blocking 40GbE 576 port switch which consumes less than 2.5W for a port of 10GbE. One quarter of the power used by the closest solution.

The power consumption of a 10GbE NIC is less than 2W per port.

Providing a CPU efficient solution creates additional power savings by reducing the number of servers.

Hyper-scale Designs (>10K Nodes)

Hyper-scale deployment with over 10K nodes is made simple. A cluster of 3 CSGs supports approximately 12K nodes, where each CSG supports up to 648 ports of 40 Gigabit. This configuration will have 40 compute nodes per ToR and 240Gb/s up-link connectivity to the aggregation and core layer. Larger scale is possible with a 240Gb/s up-link connectivity when a ToR is connected to a maximum of 58 servers per ToR

Other Notes

- Mellanox's SX1024 is a highly efficient 1U ToR switch, combining twelve 40GbE ports with forty-eight 10GbE ports in the front panel.
- SX1024 provides an excellent form factor for building a non-blocking 10GbE cluster fabric with a 40GbE aggregation layer.
- Mellanox single chassis switches will be available in the near future, including a 648-port non-blocking 40GbE switch.
- The single chassis switches enable a higher level of scalability using 10GbE and 40GbE networks.
- Please contact your Mellanox representative if you wish to better learn more on our multi-port chassis switches.

Mellanox TECHNOLOGIES

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

3881SB Rev 1.1