WHITEPAPER

April 2010

Cisco Nexus Switches and UCS Servers with Mellanox ConnectX®-2 EN with RoCE for a Reliable and Low Latency Data Center Infrastructure

High Frequency Trading, Business Intelligence, Web Analytics Applications Can Be Accelerated For Higher Productivity, Competitiveness and Customer Satisfaction

1.0 The Networking Reliability Challenge in Demanding Data Centers

Data center networks are being increasingly stretched to deliver performance in the face of growing users and data, usage spikes and unpredictable workloads. High frequency trading is experiencing exponential growths in the amount of capital market data that needs to be processed as well as growth in trading volumes – this means that more trades need to be executed faster and more data needs to be analyzed and processed for each trade. Business intelligence applications such as data warehousing and online transaction processing are challenged with the need to maintain low table scan times with exponentially growing data sets that need to be processed for each table scan request. Page uploads that require web analytics face the same challenges – more customized page uploads (requiring analytics) in the face of growing user base and more data to process. Use of cloud computing for such applications only exacerbate the challenges further with unpredictable data volume and usage spikes.

The above are a few examples of where the IT networking infrastructure needs to deliver reliable and low latency services without compromise. The need to scale server and storage capacity efficiently is equally important. A reliable infrastructure means a lossless network where data is guaranteed to be delivered. A low latency infrastructure implies that data is delivered from the sender to the receiver in the lowest possible time. Reliability with low latency goes hand-in-hand in delivering fast and deterministic performance. Low latency server to server messaging, as in clustered applications, is a key ingredient of efficient scaling of compute and server capacity. This solution brief provides an overview of how the Cisco Unified Computing System, the Cisco Nexus switches and the Mellanox ConnectX EN with RoCE (RDMA over Ethernet) low latency 10GigE adapters when used together can meet the demanding performance needs of data center fabrics.

1.1 Elements of a Lossless Network

IEEE DCB enables Ethernet fabrics to support lossless transmission to increase network scalability, support I/O consolidation, ease management of multiple traffic flows, and optimize performance. The applicable IEEE DCB and related standards to deliver a lossless and reliable data center fabric include:

- Priority-based Flow Control (PFC) provides a link level flow control mechanism that can be controlled for independently for each priority. The goal of this mechanism is to ensure zero loss due to congestion in networks.
- Enhanced Transmission Selection (ETS) provides a common management framework for assignment of bandwidth to traffic classes.
- A discovery and capability exchange protocol that is used for conveying capabilities and configuration of the above features between neighbors to ensure consistent configuration across the network. This protocol is expected to leverage functionality provided by 802.1AB (LLDP).

1.2 Elements of End-to-End Lowest Network Latency

Delivering low latency means moving data between senders and receivers of data at the lowest possible time. Server to server, server to LAN or server to storage latency is measured in milliseconds or microseconds and it encompasses the application that sends and receives data and everything in between that is involved in moving the data. This includes the operating system user and kernel space networking stack, the PCI bus on the server, the power of the CPU, the network interface card (NIC), and the connectors and the cables that connect the NICs to the switches, and the switches themselves. Most latency delays occur in the operating system and the NIC, followed by the switch.

On the NIC and operating system sides, technologies such PCI Express Gen 2, RDMA (remote DMA), kernel bypass, transport offload, and the RDMA verbs API are applicable. There are two RDMA over Ethernet standards in the industry – the IETF iWARP standard that uses the TCP or SCTP transports, and the IBTA RoCE standard that uses the InfiniBand transport. RDMA and kernel bypass implementations require offload of the transport layer to the NIC while maintaining scalability for applications and costs and power budgets in the NICs. The RDMA verbs API is defined and maintained by the OpenFabrics Alliance and the verbs API; the associated driver and protocol software is distributed using the OpenFabrics Enterprise Distribution (OFED). Switches need to implement cut-through switching to deliver the lowest latency. The ability to identify RDMA or low latency flows and assign the appropriate quality of service levels at cut-through speeds is critical. Finally, at the connector and cable levels, the differences in latencies are only barely perceptible with SFP+ performing better than 10G-BaseT.

1.3 The Cisco Nexus Switches Deliver a Reliable Data Center Fabric

The Cisco Nexus 5000 Series meets business, service, application, and operational requirements of such data centers wanting to solve the above challenges. The switch family, using cut-through architecture, supports line-rate 10 Gigabit Ethernet on all ports while maintaining consistently low latency independent of packet size and services enabled. It supports the IEEE Data Center Bridging (DCB) features that increase the reliability, efficiency, and scalability of Ethernet networks. These features allow the switches to support multiple traffic classes over a lossless Ethernet fabric, thus enabling consolidation of LAN, SAN, and cluster environments. The combination of high port density, lossless Ethernet, wire-speed performance, and extremely low latency makes the Cisco Nexus 5000 Series family ideal for meeting the needs of demanding data centers.

Figure 1: The Cisco Nexus Switch Family

1.4 High Performance and Scalable Cisco UCS Rack Servers

The Cisco UCS C-Series Rack Mount Servers feature Intel Xeon 5500 series processors and PCI Express Gen 2 bus that deliver intelligent performance, automated energy efficiency, and flexible virtualization. Intel Turbo Boost Technology automatically boosts processing power through increased frequency and use of hyperthreading to deliver high performance when workloads demand and thermal conditions permit. The patented Cisco Extended Memory Technology offers twice the memory footprint (384 GB) of any other server using 8-GB DIMMs, or the economical option of a 192-GB memory footprint using inexpensive 4-GB DIMMs. Both choices for large memory footprints can help speed performance of data intensive applications by allowing more data to be cached in memory.

Figure 2: The Cisco UCS Product Family

1.5 Mellanox ConnectX EN with RoCE (RDMA over Ethernet)

Mellanox ConnectX-2 EN NICs support the IEEE DCB standards for lossless Ethernet over 10GigE and 40GigE data rates. The ConnectX-2 EN with RoCE NIC solution delivers the lowest application-to-application level latency utilizing the RoCE industry standard specification. It supports the entire breadth of the OpenFabrics Enterprise Distribution (OFED) API. ConnectX-2 EN with RoCE incorporates a purpose-built, most deployed, field proven and scalable RDMA transport technology to deliver 1.3 microseconds application level latency, bringing InfiniBand-like performance and clustering efficiency to lossless Ethernet fabrics. The use of a simple, purpose-built transport in the hardware results in cost and power economies as well. ConnectX-2 EN with RoCE requires a lossless Ethernet fabric to deliver such low latencies in a reliable and deterministic way to ensure demanding data center applications are able to scale and perform consistently in the face to massive growth in data and fabric usage spikes. Because it supports the standard OFED API, applications that utilize that API today can now seamlessly run over ConnectX-2 EN with RoCE in UCS servers, making the benefits of this product readily available to end users, independent software vendors, appliance vendors and original equipment manufacturers.

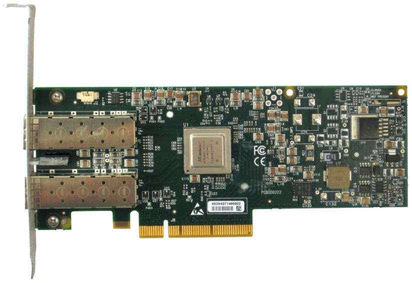

Figure 3: Mellanox ConnectX EN with RoCE NICs (Dual SFP+ Card shown)

1.6 Cisco UCS Rack Servers with Mellanox ConnectX EN (RoCE)

Mellanox ConnectX-2 EN with RoCE used with the Cisco UCS Rack Servers perfectly complements the lossless reliability and deterministic latency benefits in the Cisco Nexus series switches. Mellanox ConnectX-2 EN with RoCE when installed in Cisco UCS Rack Servers enables the Cisco UCS servers to participate in a lossless and ultra low latency end-to-end data center network. Support of discovery and capability exchange protocol, PFC and ETS in the ConnectX-2 EN with RoCE NIC enables the Cisco USC server to negotiate the appropriate level of priority for latency sensitive flows with the Cisco

Nexus switches in the fabric to deliver reliable and deterministic performance.

The efficient RDMA implementation in ConnectX-2 EN with RoCE helps reduce the cost and power requirements in the server. Kernel bypass and support of the entire breath of the OFED API enables applications running on UCS servers to deliver the maximum performance.   The balanced resources of the Cisco Unified Computing System and the use of high throughput, low latency, lossless connectivity with ConnectX-2 EN with RoCE allow the system to easily process intensive high frequency trading, data warehousing, online transaction processing (OLTP) and decision-support system (DSS) workloads with no server or network resource saturation.

1.7 Summary

Demanding data center applications require network infrastructures that can deliver reliable and deterministic performance. High performance, low latency and lossless Ethernet fabrics serve as cornerstones in the delivery of mission critical application performance in the face of growing data and transaction volumes, and their ability to handle sudden usage spike without compromising performance. Cisco UCS Servers with Mellanox ConnectX EN with RoCE and the Cisco Nexus family of switches deliver the performance, reliability, lossless and ultra-low latency features demanded by today's data centers.

Mellanox
TECHNOLOGIES

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