TECHNOLOGY BRIEF

April 2010

ConnectX®-2 EN with RoCE

1.0 Opportunities with Evolution of Ethernet

The two commonly known RDMA (remote DMA) technologies are InfiniBand and iWARP (Internet Wide Area RDMA Protocol). InfiniBand has enjoyed significant success to date in HPC applications. iWARP solutions over Ethernet have seen limited success because of implementation and deployment challenges. Recent enhancements to the Ethernet data link layer under the umbrella of IEEE data center Bridging (DCB) open significant opportunities to proliferate the use of RDMA technology into mainstream data center applications by taking a fresh and yet evolutionary look at how such services can be more easily and efficiently delivered over Ethernet. The proposed DCB standards include: IEEE 802.1bb – Priority-based flow control, 802.1Qau – Congestion Notification, and 802.1az – Enhanced Transmission Selection (ETS) and DCB Capability Exchange. The lossless delivery features in DCB, enabled by Priority-based Flow Control (PFC), are analogous to those in the InfiniBand data link layer. As such, the natural choice for building RDMA services over PFC-based DCB Ethernet is to apply use InfiniBand-based native RDMA transport services. The IBTA (InfiniBand Trade Association) has recently released a specification called RDMA over Converged Ethernet (RoCE, pronounced as "Rocky") that applies the InfiniBand-based native RDMA transport services over Ethernet. ConnectX-2 EN with RoCE (RDMA over Ethernet) implements the RoCE standard to deliver InfiniBand-like ultra low latency and high scalability over Ethernet fabrics.

1.1 How ConnectX-2 RoCE Works

ConnextX-2 EN with RoCE is born out of combining InfiniBand native RDMA transport with Ethernet per the IBTA RoCE specification. The data link InfiniBand-based layer 2 is replaced by Ethernet layer 2, as shown in the figure below. The InfiniBand transport is applied over a PFC-based loss less Ethernet data link.

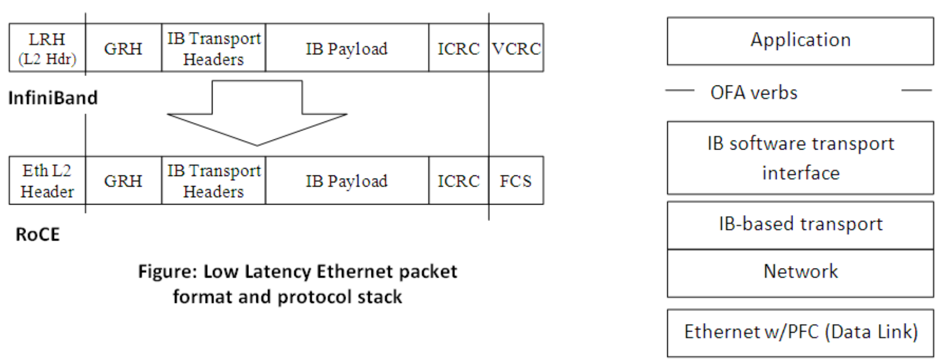

Figure: Low Latency Ethernet packet format and protocol stack

Software Interface: ConnextX-2 EN with RoCE is compliant with the Open Fabrics Alliance OFED verbs definition and is interoperable with the OFA software stacks (similar to InfiniBand and iWARP). ConnextX-2 EN with RoCE uses the proven and feature rich InfiniBand verbs interface available in the OFA stacks. OFED v1.5.1 supports RoCE and ConnextX-2 EN with RoCE.

Transport Layer: ConnextX-2 EN with RoCE uses the InfiniBand transport layer, as defined in the IBTA RoCE specification. The adaptation from InfiniBand data link to Ethernet data link is straight forward because the InfiniBand transport layer was designed ground up to be data link layer agnostic. The InfiniBand transport layer expects certain services from the data link layer related to lossless delivery of packets, and these are delivered by a PFC enabled Ethernet data link layer. ConnextX-2 EN with RoCE inherits a rich set of transport services beyond those required to support OFA verbs including connected and unconnected modes, reliable and unreliable services. Built on top of these services is a full set of verbs-defined operations including kernel bypass, send/receive, RDMA read/write, and atomic operations.

Network Layer: ConnextX-2 EN with RoCE relies on the InfiniBand defined GRH (Global Route Header) based Network Layer. When necessary, ConnextX-2 with RoCE requires InfiniBand GRH-based network layer functions. The GRH carries GID (Global Identifier) which is equivalent to IPv6 addressing and can be adapted to IPv4 addressing.

Data Link Layer: At the data link layer level, standard layer 2 Ethernet services are needed, and 802.1bb Priority flow control (PFC) or 802.3x Pause at a minimum to ensure lossless packet delivery. 802.1au congestion notification is desirable but not mandatory unless server to server or server to storage connectivity fabrics are oversubscribed and are prone to congestions. L2 Addressing is based on source and destination MAC addresses. The 802.1Q header priority field alongside 802.1az (ETS) and other Ethernet practices provide a way to implement of QoS. Finally, an IEEE assigned Ethertype is used to indicate that the packet is of type RoCE. The following table summarizes how Ethernet layer 2 header fields are mapped to functions provided by the InfiniBand layer 2 header fields to enable seamless operation of the InfiniBand transport layer over Ethernet data link layer.

Function	InfiniBand L2 Header Field	Ethernet L2 Header Field
Addressing	SLID and DLID	SMAC and DMAC
Priority Queues	Service Level (SL)	802.1Q header priority
Partitioning or VLAN	Partition Key (P-Key)	802.1Q header VLAN ID
Congestion notification	IBTA defined FECN and BECN	802.1Qau QCN

Converged Traffic: A RoCE packet is identified by an Ethertype number in the L2 header. This allows differentiation among different packet types to occur low in the stack and allows different types of Ethernet traffic, including RDMA traffic to simultaneously co-exist on a single physical Ethernet wire. ConnextX-2 EN with RoCE uses linear look up on the destination queue pair number (DQPN) in the transport header to de-multiplex traffic into queue pairs.

Management: ConnextX-2 EN with RoCE does not require an SM (InfiniBand subnet manager), and can operate using standard Ethernet network management practices for L2 address assignments, L2 topology discovery, and switch filtering data base (FDB) configuration. For example spanning tree and learning can be used. QoS management for RoCE can be accomplished using Ethernet management practices for 802.1Qaz (ETS). For congestion management features RoCE relies on 802.1au congestion management features in Ethernet. PFC priority configuration and negotiation with PFC-capable switches can do done statically using VLANs (associating RDMA traffic to VLANs in hosts and assigning high PFC priority to those VLANs in switches) or dynamically using DCB exchange protocols between the NIC and the switch. ConnectX-2 EN with RoCE supports both modes of PFC configuration. Finally, performance monitoring, baseboard and device management can be done by using standard SNMP/RMON MIBs.

The following table summarizes how network management characteristics expected by the InfiniBand transport layer and applications using the InfiniBand transport layer can be seamlessly delivered over Ethernet using standard Ethernet management practices and eliminating the need for the InfiniBand Subnet manager. Data center IT managers can continue to use their familiar Ethernet-based management tools making deployment of ConnextX-2 EN with RoCE in the data center easy like deployment of any other Ethernet-based technology.

Management Feature Required by IB Transport Layer and Apps using IB Transport Layer	How InfiniBand delivers them in the InfiniBand subnet	How Ethernet (and DCB) delivers them using standard Ethernet management practices
L2 address assignment	Subnet Manager L2 address assignment	Fixed assigned L2 address or other Ethernet mechanisms
L2 topology discovery and switch FDB configuration	Subnet Manager topology discovery using direct routed subnet management packets (SMP). Subnet Manager path computation and path distribution	Spanning Tree and Learning mechanisms. Also IETF Transparent Interconnection of Lots of Links (TRILL) when available and other eth practices
Address resolution	SA based path resolution	Address Resolution Protocol (ARP) or direct mapping
QoS	QoS Manager extension to Subnet Manager	Standard Ethernet QoS management practices. Local API to access fabric policy settings
Congestion management	Congestion Manager for IB	802.1Qau congestion management features
Performance management	IB Performance Manager	SNMP/RMON MIBS
Device/baseboard management	IB Baseboard Manager	SNMP/RMON MIBS

ConnextX-2 EN with RoCE adapters based on the IBTA RoCE specification are available today from Mellanox Technologies and have been demonstrated to deliver end to end application level latencies of as low as 1.3 microseconds. Mellanox and other industry leaders are collaborating on growing the ecosystem of RoCE-based adapters and independent software vendor applications that capitalize on the benefits of ConnextX-2 EN with RoCE. Some examples of target applications are financial services, business intelligence, data warehousing, cloud computing and Web 2.0.

1.2 ConnectX-2 EN with RoCE Advantages

Based on the discussion above, it is obvious that ConnextX-2 EN with RoCE comes with many advantages and holds the promise to enable widespread deployment of RDMA technologies in mainstream data center applications.

1. ConnextX-2 EN with RoCE utilizes advances in Ethernet (DCB) to enable efficient and low cost implementations of RDMA over Ethernet.
2. ConnextX-2 EN RDMA traffic can be classified at the data link layer which is faster and requires less CPU overhead.
3. ConnextX-2 EN with RoCE delivers 1.3usec application to application latency, which is 1/10th of other industry standard implementations over Ethernet. Benchmarking with popular financial services applications show more than 60% lower latency applicable to capital market data processing and trade executions.
4. ConnextX-2 EN with RoCE supports the entire breath of RDMA and low latency features. This includes reliable connected service, datagram service, RDMA and send/receive semantics, atomic operations, user level multicast, user level I/O access, kernel bypass, and zero copy.
5. The OFA verbs used by ConnextX-2 EN with RoCE are based on InfiniBand and have been proven in large scale deployments and with multiple ISV applications, both in the HPC and EDC sectors. Such applications can now be seamlessly offered over ConnextX-2 EN with RoCE without any porting effort required
6. ConnextX-2 EN with RoCE based network management is the same as that for any Ethernet and DCB-based network management, eliminating the need for IT managers to learn new technologies.

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