Mellanox TECHNOLOGIES

SOFTWARE

ConnectX®-2 EN with RDMA over Ethernet (RoCE)

An Efficient, low cost, zero copy implementation of RDMA over Ethernet

Introduction

Remote Direct Memory Management (RDMA) is the remote memory management capability that allows server to server data movement directly between application memory without any CPU involvement. RDMA over Ethernet (RoCE) is a mechanism to provide this efficient data transfer with very low latencies on lossless Ethernet networks. With advances in data center convergence over reliable Ethernet, ConnectX-2 EN with RoCE uses the proven and efficient RDMA transport to provide the platform for deploying RDMA technology in mainstream data center application at 10GigE and 40GigE link-speed.

RDMA over Ethernet technology

RoCE with ConnectX®-2 EN is based on the IBTA RoCE specifications. ConnectX-2 EN with its hardware offload support takes advantage of this efficient RDMA transport (InfiniBand) services over Ethernet to deliver ultra low latency for performance-critical and transaction intensive applications such as financial, data base, storage, and content delivery networks.

RoCE utilizes the Open Fabrics Enterprise Distribution (OFED) verbs interface as the software interface between application layer and ConnectX-2 EN hardware. RoCE takes advantage of transport services support of various modes of communication, such as reliable connected services and datagram services. RoCE uses well defined verbs operations including kernel bypass, send/receive semantics, RDMA read/write, user-level multicast, user level I/O access, zero copy and atomic operations.

ConnectX-2 EN with RoCE uses InfiniBand Global Route Header (GRH) for network layer. GRH uses Global Identifier (GID), equivalent of IPv6 addressing for network addresses.

Data Link Layer utilizes standard layer 2 services with end to end Priority Flow Control (PFC, IEEE 802.1Qbb) or 802.3x Pause for a lossless packet delivery. Ethernet traffic differentiation on the same wire is done by using a RoCE specific ether-type (IEEE provided) enabling convergence over Ethernet.

Data Center Convergence with RoCE

RoCE provides the important third dimension for converging on Ethernet along with TCP (LAN) and FCoE (SAN) completing the traffic consolidation on a single wire with ConnectX-2 EN adapter. RoCE is suited for clustered, grid and utility computing and addresses an ever growing component of service oriented infrastructure where low-Latency between server nodes directly translates to doing more with fewer serv¬ers, through higher efficiency, and deliver on the promise of "time is money" where every microsecond delay in executing an algorithmic or derivative transaction can result in millions of dollars in losses.

Improve Application Performance

RoCE with ConnectX-2 EN delivers the lowest latency of 1.3 microseconds allowing very high-volume, transaction intensive applications typical of financial market firms and other industries where speed of data delivery is paramount to take advantage. In data mining or web crawl applications, RoCE provides the needed boost in performance to search faster by solving the network

BENEFITS

- Utilize advances in lossless Ethernet (DCB) for an efficient RDMA over Ethernet
- A low cost, low power solution for performance centric Data Center applications
- Traffic Classification at the Link Layer (layer 2) improving network efficiency
- Lowest latency of 1.3 microseconds on lossless Ethernet
- RDMA Transport offload witFh zero copy for low CPU utilization
- Improves performance in financial, data-warehouse, data mining, storage, data base, Web 2.0 and Business intelligence applications
- Ethernet management infrastructure can be leveraged "as-is"
- Single Ethernet wire for IPC, LAN and SAN to complete Ethernet convergence

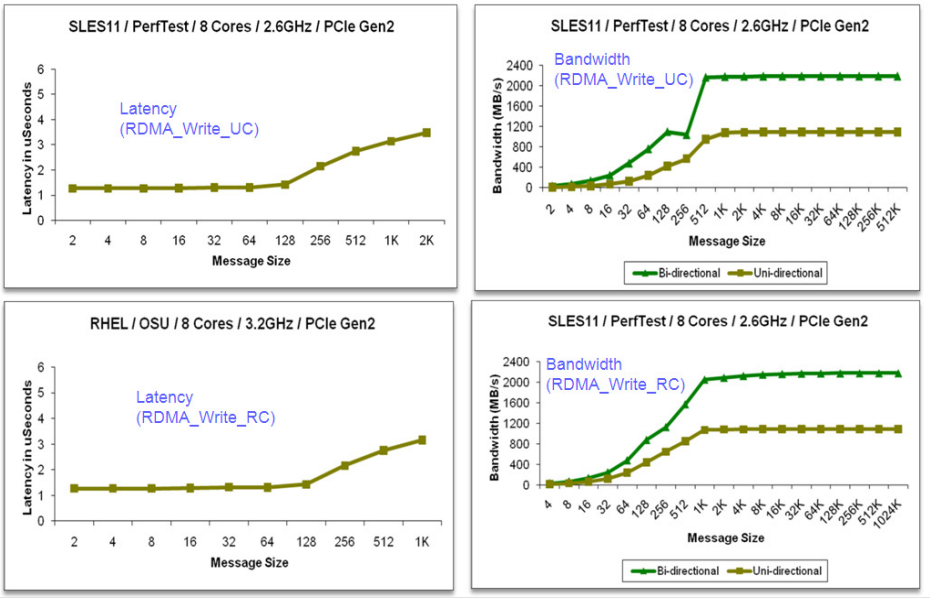
KEY FEATURES

- IBTA 1.2.1 RoCE standards compliant
- Supports IEEE 802.1Qau, 802.1Qbb and, 802.1Qaz (DCB) standards
- OFED verbs compliant with OFED software stack interoperability
- RoCE utilizes mature RDMA transport layer based on IBTA specification
- Traffic differentiation at Link Layer with IEEE defined ether-type
- SNMP based network management with MIB II support
- Interoperable with any industry standard 10GigE (DCB) switches in a large cluster

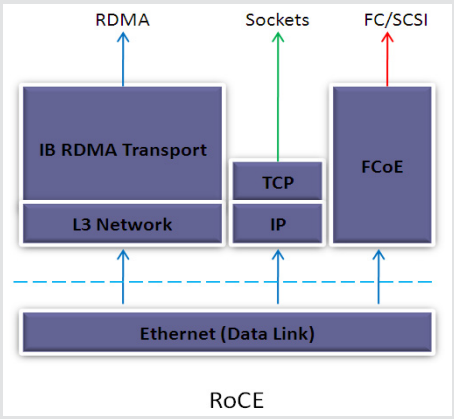
latency bottleneck associated with I/O cards and the corresponding transport technology in the cloud. Various other applications that benefit from RoCE with ConnectX-2 EN include,Web 2.0 (Content Delivery Network), Business intelligence, data base transactions and various Cloud computing applications.

Easy network management

With data center trends evolving into enterprise clouds, ConnectX-2 EN with RoCE's scalability addresses the elastic compute needs with no change in 10GigE network management. IT Managers continue to manage ConnectX-2 EN with RoCE the same way Ethernet and DCB-based networks are managed today. It ensures interoperability on existing Ethernet infrastructure and takes advantage of virtual-links with per-priority pause support.



The figure shows the low level RDMA (write) benchmarks for latency and bandwidth. These benchmarks measurements are on a 2 node cluster connected back-to-back and show latency of ~1.3 microseconds. The bandwidth reaches line rate at 512 byte message size making RoCE with ConnectX-2 EN the best Clustering solution for various transaction sensitive applications



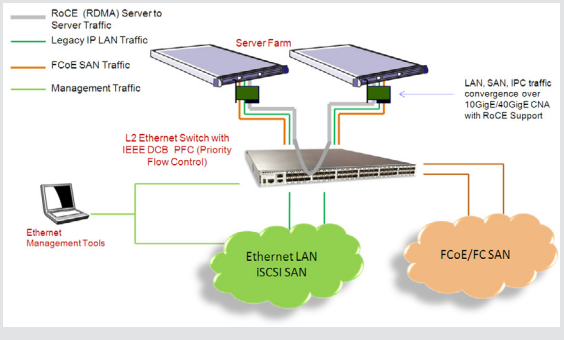


Figure 2 & 3: ConnectX-2 EN (TCP/RoCE/FCoE)

Shows the flexibility of ConnectX-2 EN and its support for Convergence using the existing stack

- Lossless Layer 2 (Ethernet)
  - o Full DCB compliant
  - o Converged I/O (LAN, SAN, IPC)
- Leverage the existing TCP/IP for LAN
- RDMA over Converged Ethernet (RoCE)
  - o Mature IB RDMA transport
  - o Proven OFED IB Verb semantics
  - o Kernel bypass, SEND/RECEIVE
  - o Atomic Operations
  - o Scales to 40Gbps for 40GigE
  - o Large scale clustering efficiency
- FCoE T-11 frame format support
- RDMA and FCoE full hardware offload

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