RoCE vs. iWARP Competitive Analysis Brief

1.0 Executive Summary

Internet Wide Area RDMA Protocol (iWARP) and RDMA over Converged Ethernet (RoCE) are the two commonly known Remote Direct Memory Access (RDMA) technologies over Ethernet. Introduced in 2002, iWARP solutions over Ethernet have seen limited success due to implementation and deployment challenges. Benchmarks results published by lead iWARP technology providers, show 6usec latency that Mellanox standard Connectx-2 Ethernet technology supports today. Since the scaling and ROI needs in Web 2.0 and Cloud Computing have not been able to be met by iWARP, a more advanced RDMA over Ethernet technology was required.

Recent enhancements to the Ethernet data link layer under the umbrella of IEEE Data Center Bridging (DCB) enabled the application of advanced RDMA transport services over DCB. In early 2010, this technology, now known as RDMA over Converged Ethernet (RoCE) was standardized by the IBTA. Mellanox Technologies was the first company to implement the new standard and its ConnectX-2 EN with RoCE supports RDMA natively and completely offloads the transport layer. It provides a wire-speed throughput and 1.3usec latency at the lowest CPU and memory utilization over 10Gig Ethernet. As a result, in less than six months after its introduction, ConnectX-2 EN with RoCE has been deployed in a variety of mission critical, latency sensitive data centers.

RDMA communications using Send/Receive semantics and kernel bypass technologies in server and storage interconnect products permit high through-put and low-latency networking, and the benefits of high throughput and low-latency networking became predominant today in EDC applications. Efficiency is synonymous with ROI, an ever critical goal of the EDC market especially with the scaling needs of Web 2.0 and Cloud Computing applications. As such, the importance of low latency technologies such as RDMA has grown, and the need for viable RDMA products that is broadly deployable across market and application segments has become critical.

iWARP uses IETF defined RDDP (Remote Direct Data Placement) to deliver RDMA services over standard, unmodified IP network and standard TCP/IP Ethernet data link services and the RDDP is layered over existing IP transports. While this helps support a broad range of network characteristics, from short range to long range networks, it brings some inherent disadvantages:

1. RDDP traffic cannot be easily managed and optimized in the fabric itself, leading to inefficiency in deployments. It does not provide a way to detect RDMA traffic at or below the transport layer, e.g. within the fabric itself. Sharing of TCP’s port space by RDDP makes using flow man-

1.1 An Introduction to iWARP

iWARP uses IETF defined RDDP (Remote Direct Data Placement) to deliver RDMA services over standard, unmodified IP network and standard TCP/IP Ethernet data link services and the RDDP is layered over existing IP transports. While this helps support a broad range of network characteristics, from short range to long range networks, it brings some inherent disadvantages:

1. RDDP traffic cannot be easily managed and optimized in the fabric itself, leading to inefficiency in deployments. It does not provide a way to detect RDMA traffic at or below the transport layer, e.g. within the fabric itself. Sharing of TCP’s port space by RDDP makes using flow man-

1.0 Executive Summary ...............................................................................................................1
1.1 An Introduction to iWARP ....................................................................................................1
1.2 Introduction to RoCE ............................................................................................................. 2
1.3 Performance and Benchmarks Examples ..............................................................................2
1.4 Summary ............................................................................................................................... 3
1.2 An Introduction to RoCE

The above challenges have significantly delayed the availability of cost-effective and easily deployable iWARP products. As such, iWARP’s usage in the industry is very limited, and has not been able to leverage the ubiquity of Ethernet to proliferate its use.

Recent enhancements to the Ethernet data link layer open significant opportunities to proliferate the use of RDMA into mainstream data center applications by taking a fresh and evolutionary look at how such services can be more easily and efficiently delivered over Ethernet:

- IEEE 802.1Qbb Priority flow control (PFC) standardizes a link level flow control that recognizes 8 traffic classes per port. While traditional Ethernet pause is flow controlled at the granularity of physical ports, with priority flow control, pause is at the granularity of a traffic class. PFC is the first per priority flow control mechanism for Ethernet.
- IEEE 802.1Qau standardizes congestion notification through an admission control algorithm called Quantized Congestion Notification (QCN). The QCN algorithm is inspired by congestion control in TCP, but implemented at layer 2.

With these new enhancements the market took a fresh look at iWARP. As noted earlier, the wider scope of iWARP to deliver RDMA services over existing IP networks resulted in unnecessary burdens and product and deployment failures and delays. Many of the services in the iWARP stack become redundant when applied over DCB. Since DCB provides congestion control, the congestion functions provided in TCP, SCTP and RDDP become redundant. Over DCB, TCP slow-start behavior can be relaxed or removed. The question is should the weathered and beaten iWARP stack be redesigned with a new approach to make it more suitable to operate over DCB and reduce the implement and deployment complexities it continues to suffer. From both the technology and business perspectives, the natural choice for building RDMA services over DCB is to apply more efficient RDMA transport services over Ethernet. RDMA over Converged Ethernet or RoCE has quickly become the sought after solution.

EDC latency sensitive applications like messaging platforms for trade execution are cornerstones of competitiveness for financial services institutions. Such platforms can benefit from Mellanox’s ConnectX-2 EN with RoCE solution to deliver extremely low latencies on Ethernet while scaling to handle millions of messages per second.

1.3 Performance and Benchmarks Examples

agreement impossible, since the port alone cannot identify whether the message carries RDMA or traditional TCP.

2. Application integration is harder. Because RDDP is layered over existing transports, it shares their port space, and the necessary RDMA capability detection and service resolution or negotiation is an additional, RDMA-specific step which applications must make.

3. Hardware offload implementation is harder. RDDP is designed to work over the existing TCP transport and Ethernet data link layers. The Ethernet data link delivers best effort service and is prone to packet loss, relying on the TCP layer to deliver reliable services. The need to support existing IP networks, including wide area networks require coverage of a larger set of boundary conditions with respect to congestion handling, scaling and error handling, causing hardware offload of the RDMA and associated transport operations significantly harder.

4. iWARP is relatively an old technology that was introduced in 2002. Most advanced 10Gig Ethernet that support iWARP shows 6usec latency which can be achieved by a variety of today’s 10GigE NICS over common TCP/IP.
A benchmark that was done comparing the performance of a lead low latency messaging application running over 10Gig Ethernet iWARP and Connectx-2 EN with RoCE shows that in average, RoCE delivered 100Byte message 2.5 times faster than iWARP.

Similar performance results have been achieved when measuring the performance of IBM MQ Low Latency Messaging (MQ LLM).

The latency benefits showed that RoCE cut the latency by more than 2X, using lower CPU utilization and memory. The above complete benchmarks results and other are available on www.mellanox.com

Over Ethernet, RoCE provides a superior solution compared to iWARP. Many latency-sensitive applications have been ported to run over RoCE and RoCE has been already deployed in mainstream data centers.

1. RoCE utilizes advances in Ethernet (DCB) to enable efficient and lower cost implementations of RDMA over Ethernet. RoCE eliminates the need to modify and optimize iWARP to run efficiently over DCB.
2. RoCE focuses on server-to-server and server-to-storage networks, delivering the lowest latency and jitter characteristics and enabling simpler software and hardware implementations
3. RoCE supports the OFA verbs interface seamlessly. The OFA verbs used by RoCE have been proven in large-scale deployments and with multiple ISV applications in the EDC sectors. Such applications can now be seamlessly offered over RoCE without any porting effort required.
4. RoCE based network management is the same as that for any Ethernet and RoCE-based network management, eliminating the need for IT managers to learn new technologies.

For information on the best solution for accelerating your applications with Mellanox’s 10Gig Ethernet with RoCE technologies, please visit www.mellanox.com