Performance Evaluation of the RDMA over Ethernet (RoCE) Standard in Enterprise Data Centers Infrastructure

Motti Beck
Director, Marketing
motti@mellanox.com

Michael Kagan
Chief Technology Officer
michaelk@mellanox.com

Mellanox Technologies, 350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085, USA

Abstract:

RDMA or Remote Direct Memory Access, communications using Send/Receive semantics and kernel bypass technologies in server and storage interconnect products permit high through-put and low-latency networking. As numbers of cores per server and cluster sizes servicing enterprise datacenters (EDC) applications have increased, the benefits of higher performance - aka completing the job faster – are being increasingly complemented by the efficiency factor - being able to do more jobs with fewer servers. Data Center efficiency is synonymous with Return on Investment (ROI) has ever been a critical goal of the EDC, 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 efficient RDMA products that is broadly deployable across market and application segments has become critical.

Recent enhancements to the Ethernet data link layer under the umbrella of IEEE Converged Enhance Ethernet (CEE) open significant opportunities to proliferate the use of RDMA, SEND/RECEIVE and kernel bypass into mainstream datacenter applications by taking a fresh and yet evolutionary look at how those services can be more easily and efficiently delivered over Ethernet. The CEE new standards include: 802.1Qbb – Priority-based flow control, 802.1Qau – End-to-End Congestion Notification, and 802.1Qaz – Enhanced Transmission Selection and Data Center Bridge Exchange. The lossless delivery features in CEE enables a natural choice for building RDMA, SEND/RECEIVE and kernel bypass services over CEE is to apply RDMA transport services over CEE or in short RoCE.

978-0-9836283-2-3

This paper was peer reviewed by subject matter experts for publication in the Proceedings of DC–CaVES 2011

In April 2010, the RoCE – RDMA over Converged Ethernet standard that enables the RDMA capabilities of InfiniBand™ to run over Ethernet was released by the InfiniBand® Trade Association (IBTA). Since then, RoCE has received broad industry support from many hardware, software and system vendors, as well as from industry organizations including the OpenFabrics Alliance and the Ethernet Alliance.

Introduction

Converged Enhance Ethernet (CEE):

The set of standards, defined by the Data Center Bridging (DCB) task group within IEEE 802.1 is popularly known as Converged Enhanced Ethernet (CEE). The primary target of CEE is the convergence of Inter Process Communication (IPC), networking and storage traffic in the data center. To accomplish this, CEE introduces the notion of Ethernet as a lossless wire, accomplished through link level flow control and improved congestion control. In addition, CEE introduces differentiated classes of traffic and the ability to assign traffic to unique priority levels. The lossless CEE functionality is conceptually similar to the features offered by the InfiniBand data link layer and includes:

- IEEE 802.1Qbb Priority flow control (PFC) standardizes a link level flow control that recognizes 8 traffic classes per port (analogous to InfiniBand virtual lanes). 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 (analogous to InfiniBand congestion notification mechanisms)

In addition, CEE includes IEEE 802.1Qaz standardizes a scheduling function, Enhanced Transmission Selection (ETS), and a capability exchange, Data Center Bridge Exchange (DCBX). ETS allocates bandwidth to groups sharing the same traffic class. DCBX is used to learn and propagate the datacenter bridging features of connected devices. A network device can learn its optimal settings without being manually configured.

RDMA over Converged Ethernet (RoCE):

ROCE is borne out of combining IB native RDMA transport with Ethernet-based CEE. The data link IB-based layer 2 is replaced by Ethernet-based layer 2, as shown in the figure below. This combination is made possible by the unique set of features included in CEE, such as its lossless characteristics.

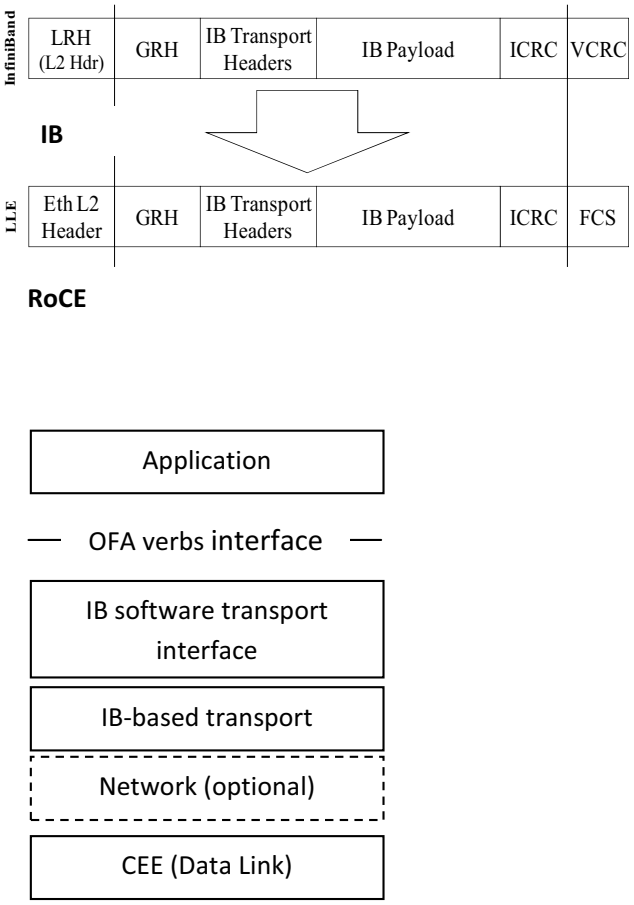

Figure 1: Low Latency Ethernet packet format and protocol stack

Software Interface and Transport Layer: ROCE is compliant with the OFA verbs definition and is interoperable with the OFA software stack (similar to InfiniBand and iWARP). The features provided by the IB software transport interface and the IB-based transport layer are analogous to what was presented earlier in the "Introduction to InfiniBand" section. The IB transport layer, as defined in the IBTA specification, is data-link layer agnostic. Hence, it does not make any assumptions about the lower layers, especially the data link layer except in one case that requires a small adaptation to work over Ethernet-based layer 2: the InfiniBand transport layer checks layer 2 address match between incoming packet and the queue pair context entry; hence the InfiniBand transport layer needs to be adapted to check Ethernet layer 2 address instead of the InfiniBand layer 2 address. The IB transport layer expects certain services from the data link layer, especially related to lossless delivery of packets, and these are delivered by a CEE based data link layer. ROCE inherits a rich set of transport services beyond those required to support OFA verbs including connected and unconnected modes and 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. Also, UDP and multicast operations are fully supported.

Network Layer: The network layer can be used for routing even though, as explained earlier, routing fo ROCE packets is undesirable when latency, jitter and throughput are the biggest considerations. When necessary, ROCE requires InfiniBand GRH-based network layer functions. In GRH, routing is based on GID (Global Identifier) which is equivalent to IPv6 addressing and can be adapted to IPv4 addressing. Layer 3 addressing is GID based. End nodes are

referred to by their IP addresses, where the GID is derived from the IP address. In the context of the layer 3 header, it is important to note that ROCE uses the InfiniBand Communication Management (CM) and packet format which makes the existence of layer 3 information in the packet mandatory. Also, layer 3 input modifiers are mandatory for the support of and compatibility with relevant OFA verbs (such as verbs for creation, modification, and query of address handle etc).

Data Link Layer: At the data link layer level, standard layer 2 Ethernet services are needed, as well as IEEE 802.1Qbb Priority flow control (PFC) at a minimum. IEEE 802.1Qau congestion notification is desirable but not mandatory unless server to sever or server to storage connectivity fabrics are oversubscribed and are prone to congestions. Addressing is based on source and destination MAC addresses (replacing SLID and DLID in InfiniBand). PFC is implemented using IEEE 801.p based priority queues (instead of virtual lanes in InfiniBand). The IEEE 802.1Q header priority fields provide the necessary service levels (replacing SLs used in InfiniBand). Finally, an IEEE assigned Ethertype is used to indicate that the packet is of type ROCE.

RoCE is implemented in and downloadable today in the latest OpenFabrics Enterprise Distribution [3] (OFED) stack. Many Linux distributions, which include OFED, support a wide and rich range of middleware and application solutions such as IPC, sockets, messaging, virtualization, SAN, NAS, file systems and databases, which enable RoCE to deliver all three dimensions of unified networking on Ethernet – IPC, NAS and SAN.

RoCE Performance Evaluation

Immediately after the RoCE standard was published, companies started the implementation and the integration of the technology into their Ethernet controllers. Mellanox Technologies was the first to implement the standard and in April 2010 the company introduced its ConnectX-2 10 GigE with RoCE product. Since then several latency sensitive applications providers already ported their applications to run over RoCE and published performance results.

The first to adopt RoCE is the financial market segment. Data volumes in the financial services industry are seeing dramatic growth, bringing existing systems to their limits. In a business where profits are directly measured by system speed, low latency, high volume infrastructures are needed with higher speeds and greater scalability.

IBM's WebSphere MQ Low Latency Messaging (WMQ LLM) is a transport fabric product engineered for the rigorous latency and throughput requirements of today's financial trading environments. The transport provides one-to-one, one-to-many and many-to-many data exchange. It also exploits the IP multicast infrastructure to ensure scalable resource conservation and timely information distribution.

Figure 2 shows the results of a benchmark that was done, comparing the performance

of IBM MQ Low Latency Messaging application running over 10Gig Ethernet with and without RoCE shows that in average RoCE delivers the message 2.5 times faster that 10Gig Ethernet.

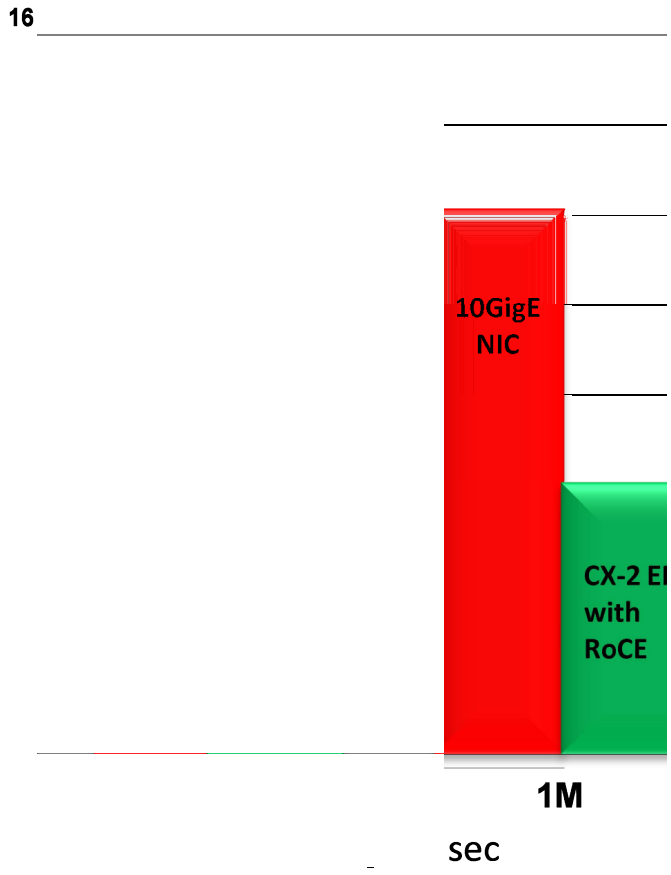

Figure 2: WMQ LLM over RoCE shows 250% latency reduction

The test results demonstrate that the WebSphere MQ Low Latency Messaging product running over low latency Ethernet can be used to implement the high performance, highly available messaging infrastructure needed for the next generation implementations of exchange systems. Support for next-generation communication fabrics allows applications to achieve the lowest possible latencies at high message volumes required to meet the trading targets that their participants require. This benchmark ran over Mellanox ConnectX-2 EN with RoCE adapters with OFED 1.5.1 and RoCE support, permitting efficient RDMA communications over a 10 GbE network.

RoCE performance has been also tested running MRG Messaging over the Red Hat Enterprise Linux 6.1 (RHEL) operating system using various interconnects and supported protocols include RoCE and RDMA. MRG Messaging was designed to provide a way to build distributed applications in which programs exchange data by sending and receiving messages. A message can contain any kind of data. Middleware messaging systems allow a single application to be distributed over a network and throughout an organization without being restrained by differing operating systems, languages, or network protocols. Sending and receiving messages is simple, and MRG Messaging provides guaranteed delivery and extremely good performance. For more information refer to http://www.redhat.com/mrg.

Figure 3 shows the performance comparison between different networking technologies. The averages for 1024-Bytes message sizes is plotted for each of the interconnects/protocols. The 1-GigE has the highest latency. Considerably less, but consistently second, is 10 GigE. The IPoIB results follow next. The 10GigE RDMA and IB RDMA both provide the lowest latency in all tests.

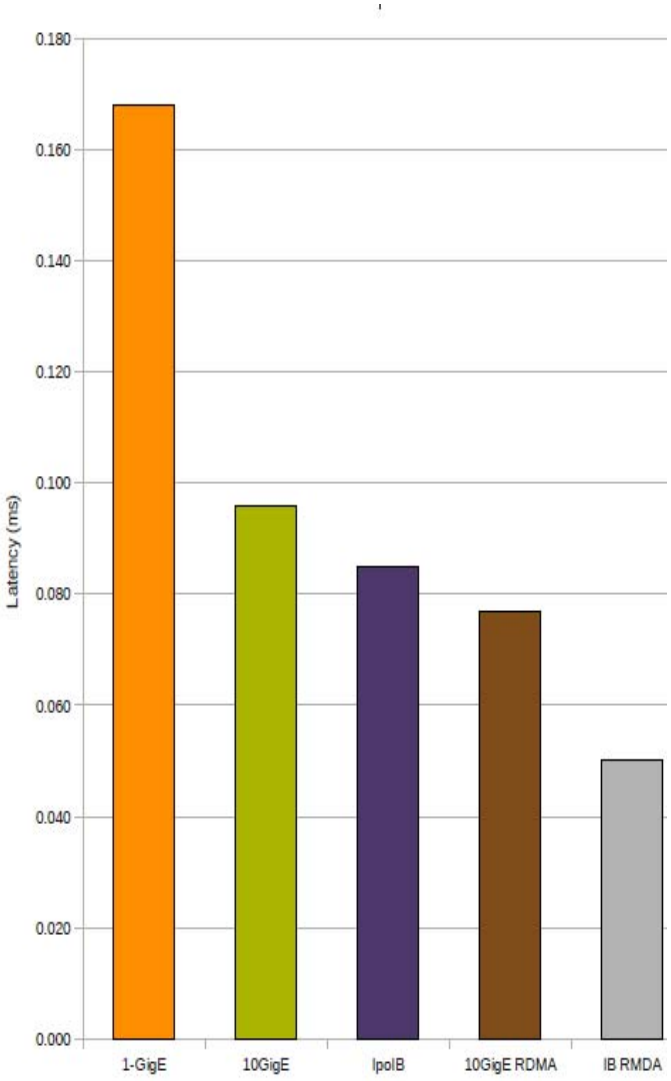

Figure 3: MRG over RH 6.1 performance comparison (Msg ZSize 1034-Bytes)

Conclusion

The importance of RDMA is growing in the industry, driven by increased use of clustered computing and the need to scale such clusters efficiently, both in node counts and performance. RDMA enables low latency, which is a cornerstone for delivering efficient computing and linear scaling of clusters, resulting in higher ROI.

The evolution of Ethernet and the development of CEE based enhancements to Ethernet open new opportunities for delivering RDMA over Ethernet.

The RoCE standard combines the proven and well deployed InfiniBand transport layer over a CEE based data link layer that deliver lossless services. Since the InfiniBand transport is data link layer agnostic and is designed for lossless fabrics, it is a natural fit above CEE to deliver RDMA services.

The RoCE advantages include:

- RoCE utilizes the advances in Ethernet (CEE) to enable efficient and lower cost implementations of RDMA over Ethernet.
- RoCE, focuses on short range server to server and server to storage networks, delivering the lowest latency and jitter characteristics and enabling simpler software and hardware implementations
- RoCE supports the OFA verbs interface seamlessly. The OFA verbs used by RoCE are based on IB 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 RoCE without any porting effort required
- RoCE based network management is the same as that for any Ethernet and CEE-based network management, eliminating the need for IT managers to learn new technologies.

In summary, RoCE comes with many advantages and holds the promise to enable widespread deployment of RDMA

technologies in mainstream datacenter applications.

References

[1] InfiniBand Trade Association, www.IBTA.org, InfiniBand Architecture Specification, Release 1.2.

[2] RoCE – RDMA over Converged Ethernet, InfiniBand Trade Association, http://www.infinibandta.org/content/pages.php?pg=press_room_item&rec_id=663

[3] OpenFabrics RDMA Protocols through OFED software, www.openfabric.org