WHITE PAPER

February 2017

RoCE vs. iWARP Competitive Analysis

Executive Summary

Remote Direct Memory Access (RDMA) provides direct access from the memory of one computer to the memory of another without involving either computer's operating system. This technology enables high-throughput, low-latency networking with low CPU utilization, which is especially useful in massively parallel compute clusters.

RDMA over Converged Ethernet (RoCE) is the most commonly used RDMA technology for Ethernet networks and is deployed at scale in some of the largest "hyper-scale" data centers in the world. RoCE is the only industry-standard Ethernet-based RDMA solution with a multi-vendor ecosystem delivering network adapters and operating over standard layer 2 and layer 3 Ethernet switches. The RoCE technology is standardized within industry organizations including the IBTA, IEEE, and IETF.

Mellanox Technologies was the first company to implement the new standard, and all of its product families from ConnectX-3 Pro and onward implement a complete offload of the RoCE protocol. These solutions provide wire-speed throughput at up to 100Gb/s throughput and market-leading latency, with the lowest CPU and memory utilization possible. As a result, the ConnectX family of adapters has been deployed in a variety of mission critical, latency sensitive data centers.

RoCE's Advantages over iWARP

iWARP is an alternative RDMA offering that is more complex and unable to achieve the same level of performance as RoCE-based solutions. iWARP uses a complex mix of layers, including DDP (Direct Data Placement), a tweak known as MPA (Marker PDU Aligned framing), and a separate RDMA protocol (RDMAP) to deliver RDMA services over TCP/IP. This convoluted architecture is an ill-conceived attempt to fit RDMA into existing software transport frameworks. Unfortunately this compromise causes iWARP to fail to deliver on precisely the three key benefits that RoCE is able to achieve: high throughput, low-latency, and low CPU utilization.

In addition to the complexity and performance disadvantages, only a single vendor (Chelsio) is supporting iWARP on their current products, and the technology has not been well adopted by the market. Intel previously supported iWARP in its 10GbE NIC from 2009, but has not supported it in any of its newer NICs since then. No iWARP support is available at the latest Ethernet speeds of 25, 50, and 100Gb/s.

iWARP is designed to work over the existing TCP transport, and is essentially an attempt to patch up existing LAN/WAN networks. The Ethernet data link delivers best effort service, relying on the TCP layer to deliver reliable services. The need to support existing IP networks, including wide area networks, requires coverage of a larger set of boundary conditions with respect to congestion handling, scaling, and error handling, causing inefficiency in hardware offload of the RDMA and associated transport operations. RoCE, on the other hand, is a purpose-built RDMA transport protocol for Ethernet, not as a patch to be used on top of existing TCP/IP protocols.

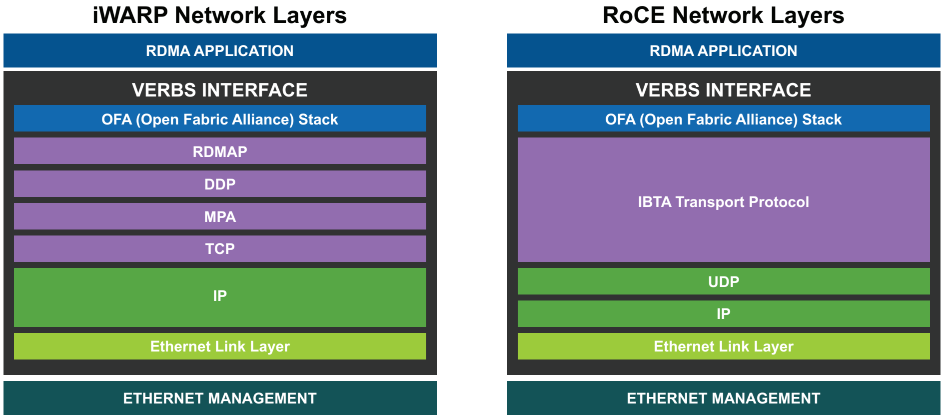


Figure 1. *iWARP's complex network layers vs. RoCE's simpler model*

Because TCP is connection-based, it must use reliable transport. iWARP, therefore, only supports reliable connected transport service, whch also means that is is not an appropriate platform for multicast. RoCE offers a variety of transport services, including reliable connected, unreliable datagram, and others, and enables user-level multicast capability.

iWARP traffic also 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, for example within the fabric itself. Sharing of TCP's port space by iWARP makes using flow management impossible, since the port alone cannot identify whether the message carries RDMA or traditional TCP. iWARP shares the protocol number space with legacy TCP traffic, so context (state) is required to determine that a packet is iWARP. Typically, this context may not fit in the NIC's on-chip memory, which results in much more complexity and therefore longer time in traffic demultiplexing. This also occurs in the switches and routers of the fabric, where there is no such state available.

In contrast, a packet can be identified as RoCE simply by looking at its UDP destination port field. If the value matches the IANA assigned port for RoCE then the packet is RoCE. This stateless traffic identification allows for quick and early demultiplexing of traffic in a converged NIC implementation, and enables capabilities such as switch or fabric monitoring and access control lists (ACLs) for improved traffic flow analysis and management.

Similarly, because iWARP shares port space with the legacy TCP stack, it also faces challenges integrating with OS stacks. RoCE, on the other hand, offers full OS stack integration.

These challenges limit the cost-effectiveness and deployability of iWARP products, especially in comparison to RoCE.

RoCE includes IP and UDP headers in the packet encapsulation, meaning that RoCE can be used across both L2 and L3 networks. This enables layer 3 routing, which brings RDMA to networks with multiple subnets.

"Resilient RoCE" enables running RoCE on Lossy fabrics, which do not enable Flow Control or Priority Flow Control. RoCE's advanced hardware mechanisms deliver RDMA performance on lossy networks on par with that of lossless networks.

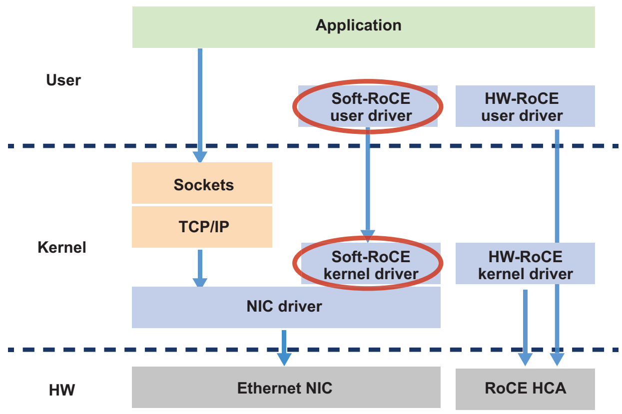


Figure 2. *Soft-RoCE Architecture*

Finally, by deploying Soft-ROCE (Figure 2), the implementation of RoCE via software, RoCE can be expanded to devices that do not natively support RoCE in hardware. This enables greater flexibility in leveraging RoCE's benefits in the Data Center.

Performance and Benchmark Examples

EDC latency-sensitive applications such as Hadoop for real-time data analysis are cornerstones of competitiveness for Web2.0 and Big Data providers. Such platforms can benefit from Mellanox's ConnectX-3 Pro, as its RoCE solution delivers extremely low latencies on Ethernet while scaling to handle millions of messages per second.

Benchmarks comparing the performance of Chelsio's T5 and T6 messaging applications running over 25, 40, and 100Gb Ethernet iWARP against the ConnectX-3 Pro with RoCE shows that RoCE consistently deliver messages significantly faster than iWARP (Figure 3).

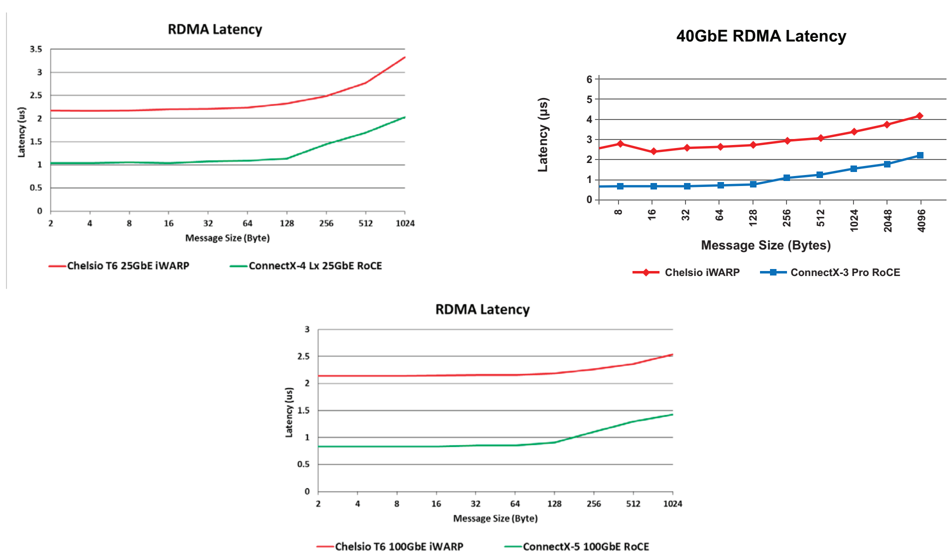


Figure 3. *25, 40, and 100Gb Ethernet Latency Benchmarks*

When measuring ConnectX-3's RoCE latency against Intel's NetEffect 020 iWARP, the results are even more impressive. At 10Gb, RoCE showed an 86% improvement using RoCE at 64B message size, and a 64% improvement at 2048B (Figure 4).

Message Size (Bytes)	NetEffect 020 iWARP	ConnectX-3 10GbE RoCE	ConnectX-3 40GbE RoCE
64B	7.22µs	1µs	.78µs
2048B	10.58µs	3.79µs	2.13µs

Figure 4. *Mellanox RoCE and Intel iWARP Latency Benchmark*

Meanwhile, throughput when using RoCE at 40Gb on ConnectX-3 Pro is over 2X higher than using iWARP on the Chelsio T5 (Figure 5), and 5X higher that using iWARP at 10Gb with Intel (Figure 6).

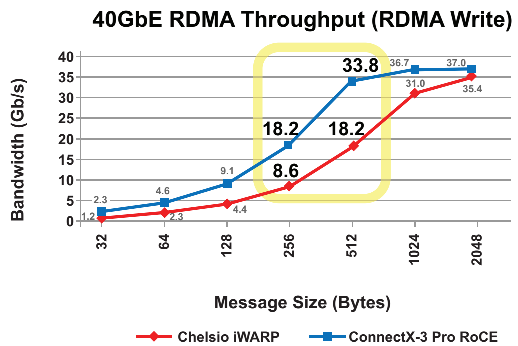


Figure 5. *40Gb Ethernet Throughput Benchmark*

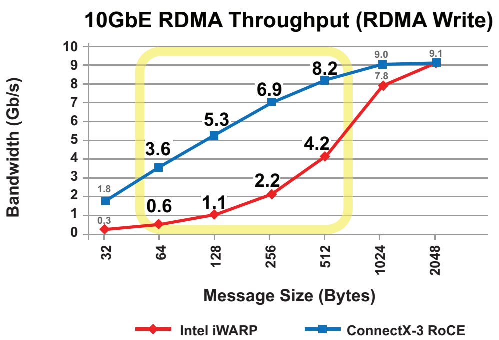


Figure 6 *10Gb Ethernet Throughput Benchmark*

The performance advantages are maintained whether using RoCE over a lossless or a lossy network (Figure 7). With Resilient RoCE, Mellanox can provide consistent, top performance with congestion control in lossy environments.

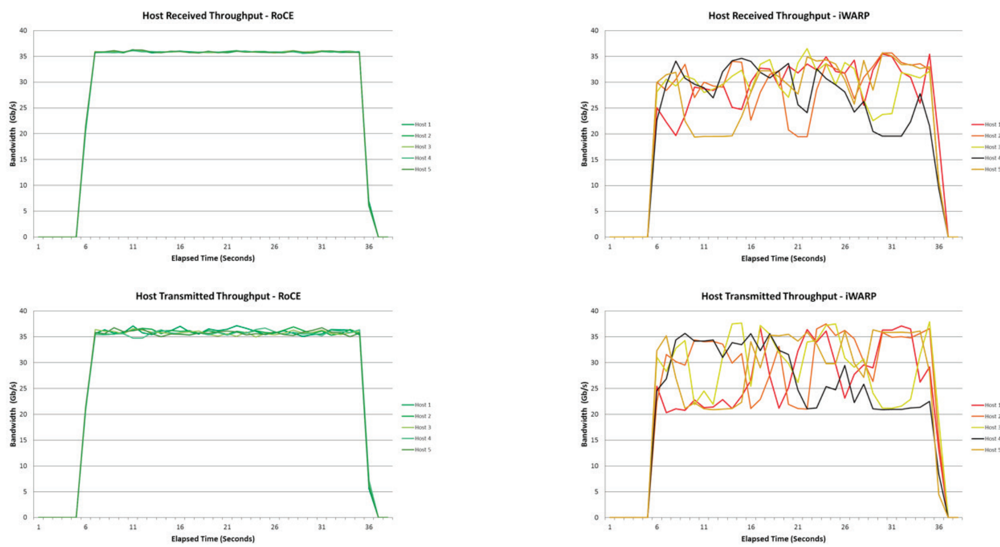


Figure 7. *Host send and receive throughput (500KB packets) in lossy network with RoCE and iWARP*

Best Performance for Virtualization

A further advantage to RoCE is its ability to run over SR-IOV, enabling RoCE's superior performance of the lowest latency, lowest CPU utlilization, and maximum throughput, in a virtualized environment. RoCE has proven it can provide less than 1 us latency between virtual machines while maintaining consistent throughput as the virtual environment scales. Chelsio's iWARP does not run over multiple VMs in SR-IOV, relying instead on TCP for VM-to-VM communication. The difference in latency is astounding (Figure 8).

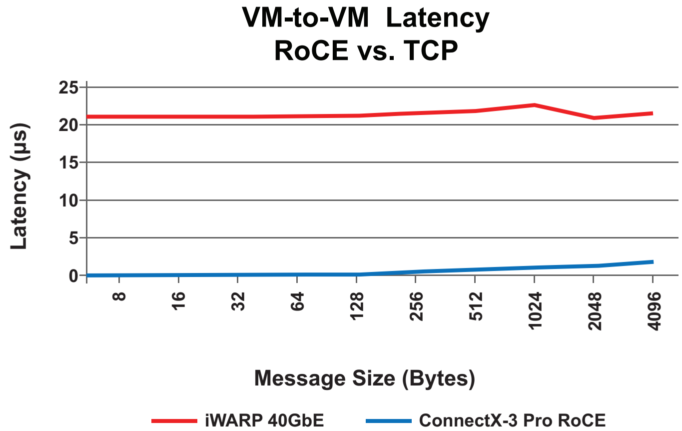


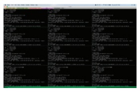
Figure 8. *40Gb Ethernet Latency from VM to VM*

Summary

RoCE simplifies the transport protocol; it bypasses the TCP stack to enable true and scalable RDMA operations, resulting in higher ROI. RoCE is a standard protocol, which was built specifically with data center traffic in mind, with consideration paid to latency, performance, and CPU utilization. It performs especially well in virtualized environments.

When a network runs over Ethernet, RoCE provides a superior solution compared to iWARP. For the enterprise data center seeking the ultimate in performance, RoCE is clearly the choice, especially when latency-sensitive applications are involved.

Moreover, RoCE is currently deployed in dozens of data centers with up to hundreds of thousands of nodes, while iWARP is virtually non-existent in the field. Simply put, RoCE is the obvious way to deploy RDMA over Ethernet.



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