WHITE PAPER

August 2015

RoCE vs. iWARP – The Facts You Should Know

Introduction 1
Outdated Information 1
Inaccurate Information 2
Use of Logarithmic Scales 2
Lab Tests vs. Real-Life Scenarios 3
Summary 4

Introduction

Chelsio has published several papers that compare its 40Gb Ethernet products with Mellanox's 40GbE and FDR 56Gb/s InfiniBand solutions, or that compare its iWARP RDMA to Mellanox's RDMA over Converged Ethernet (RoCE). In these publications, the data presented by Chelsio uses outdated information and presents data in unconventional ways. When you delve beyond Chelsio's FUD, it is clear that Mellanox solutions, whether 40Gb Ethernet or RoCE – and even more so FDR 56Gb/s and 100GbE – perform much better than Chelsio's adapters.

Outdated Information

In its attempts to promote its iWARP RDMA solution over Mellanox's RoCE solution, Chelsio's Website continues to promote documents with outdated information to bolster its claims.

For example, in its RoCE FAQ document[1] (hereafter, referred to as "the RoCE FAQ"), Chelsio claims that RoCE is not the standard RDMA over Ethernet protocol. While iWARP may have been standardized first, RoCE is an open IBTA standard[2] that runs on top of IETF standard UDP using an IANA assigned port number (4791).

Similarly, in the RoCE FAQ, Chelsio posits that RoCE does not scale and has issues of interoperability with switches from other vendors. These claims have been overcome or proven incorrect long ago. Today virtually all advanced data center network equipment supports data center bridging technology that is required to fully take advantage of RDMA. Furthermore, there is no pricing premium for these switches and RoCE NICs are smaller, lower powered, and priced significantly lower than iWARP offerings. There are deployed RoCE-based networks with tens of thousands of nodes, and interoperability across multiple vendors has been demonstrated and documented in these deployments.

Chelsio also indicates that RoCE is not routable and is unrecognized by standard traffic management and monitoring tools. Again, both claims are based on outdated information and play on antiquated fears of potential customers.

RoCE does support routable networks. The routable version of the standard was released September 2014 by the IBTA[3] with multiple vendors supporting the announcement. While the standard was only ratified in Q3 of 2014, products supporting routable RoCE have been shipping in high volume for much longer. Furthermore, RoCE fully supports standard IPv4 and IPv6 encapsulations, as well as traffic management and monitoring tools. Data centers with hundreds of thousands of nodes take advantage of these capabilities.

[1] http://www.chelsio.com/wp-content/uploads/2011/05/RoCE-FAQ-1204121.pdf
[2] https://cw.infinibandta.org/document/dl/7781
[3] http://www.infinibandta.org/content/pages.php?pg=press_room_item&rec_id=803

Inaccurate Information

Chelsio makes statements that seem to contradict demonstrated real-world results. For example, in the RoCE FAQ, Chelsio claims that the positive performance numbers seen in RoCE's micro-benchmarks do not match its real application performance. Yet public presentations[4] have demonstrated a 10X performance improvement using RoCE in real-world applications such as virtual machine migration. Moreover, RoCE's performance advantages have been proven to induce higher level performance across a broad range of storage and compute-centric applications.

In that same document, Chelsio posits that RoCE is limited to operation over short distances (of a few hundred meters). However, this is easily overcome with Layer 3 networking and various switch configurations.

Furthermore, Chelsio suggests that RoCE has no congestion management layer, depending entirely on the Priority Flow Control (PFC) Pause feature instead. In fact, the Pause feature is a Layer 2 mechanism that is unrelated to congestion management. The latest update to the RoCE specification (RoCEv2) defines all the necessary mechanisms to address congestion, and there are multiple schemes used in practice to manage congestion that are very effective in avoiding packet loss and retransmission.

Use of Logarithmic Scales

Another way that Chelsio plays with the data in its published papers is to display information on graphs that use a logarithmic scale instead of the more commonly used linear scale. This either provides an incorrect visual representation of the data, or it seems to reduce the gap between Mellanox's superior performance and Chelsio's.

For instance, Chelsio published a paper entitled "Net Direct Chelsio 40GbE vs Mellanox 56G IB,"[5] in which it presents two graphs that chart latency and bandwidth, respectively, for the two technologies. In the latency graph (Figure 1a), Chelsio claims that the performance numbers are similar, especially with IO sizes of 1288 bytes or larger.

But when the graph is set on a more conventional linear scale, the difference is much more obvious than Chelsio would have the reader believe. As Figure 1b shows, Mellanox FDR InfiniBand has consistently lower latency of 73%, and the performance gap remains steady at larger IO sizes.

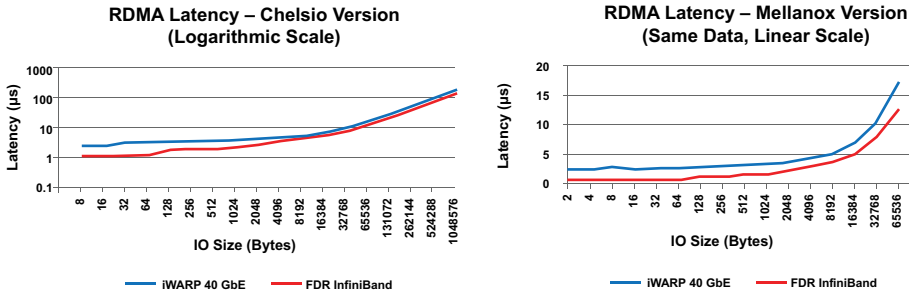

Figure 1. (a) *Chelsio RDMA latency graph using logarithmic scale* ***(b)*** *Mellanox RDMA latency graph with the same data using linear scale*

By analyzing bandwidth using the same unorthodox method of logarithmic scales, Chelsio suggests that the performance numbers are similar, and that the numbers are even closer as IO size increases. In this case, presenting the numbers on a more realistic linear scale shows the exact opposite of what Chelsio claims, with similar bandwidth for small packets but a wide gap between performance results at larger IO sizes. The distortion by using a logarithmic scale is clear, as the Chelsio graph (Figure 2a) highlights the difference of less than 0.1 Gb/s between bandwidths at small packet transfers at the expense of showing the much more significant difference of over 10 Gb/s (nearly 40% more bandwidth with FDR InfiniBand) at the larger packet sizes (Figure 2b).

[4] https://www.youtube.com/watch?v=8Kyoj3bKepY
[5] http://www.chelsio.com/wp-content/uploads/2013/11/Netdirect-Chelsio-40G-ETH-vs-Mellanox-56G-IB.pdf

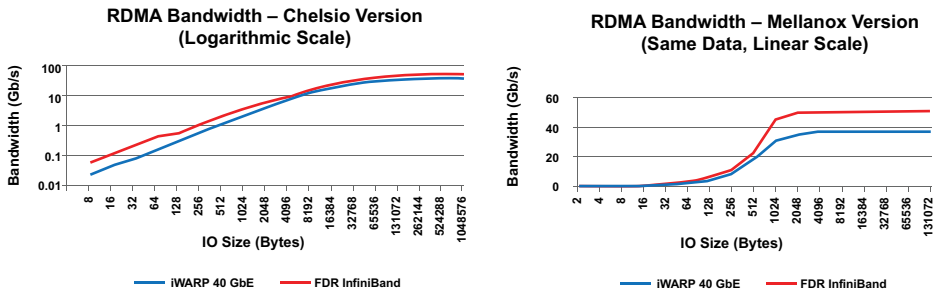

Figure 2. (a) *Chelsio RDMA bandwidth graph using logarithmic scale* ***(b)*** *Mellanox RDMA bandwidth graph with the same data using linear scale*

Lab Tests vs. Real-Life Scenarios

Another method that Chelsio uses to attack Mellanox performance is to present the results of synthetic, unrealistic tests that do not mimic real-life customer scenarios.

A perfect example of this is in Chelsio's published paper, "Chelsio vs. Mellanox 40GbE Performance."[6] Chelsio attempts to show that its 40Gb Ethernet performance is superior to that of Mellanox by showing a graph with a 34% improvement in transactions per second and 2-3 times less CPU utilization when using a TCP Offload Engine.

What Chelsio does not mention, however, is that this test was run using only a single thread of TCP_CRR Chelsio T-5 40GbE. This is a synthetic lab result, as data centers seldom use only one thread to run their applications. And yet, Chelsio's supposed performance advantage is presented as if it could scale to the demands of the world's largest data centers.

When running multiple threads, the results are drastically different. Mellanox testing showed that Chelsio's technology crashed once it reached 80 parallel threads, hardly enough to supply a reasonably sized data center. Mellanox's ConnectX®-3 Pro 40GbE adapter, on the other hand, showed consistent performance across thousands of parallel threads.

By limiting its test to a single thread, Chelsio was able to claim better performance than Mellanox on paper, but when faced with realistic scenarios, Chelsio's technology crashed and exposed its severe limitations.

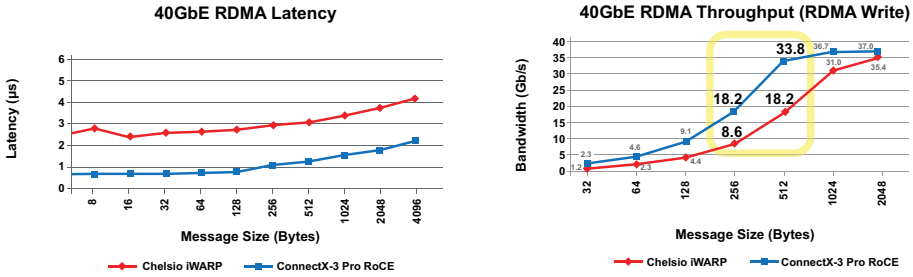

Figure 3. *Mellanox 40GbE with RoCE shows clear advantages over Chelsio 40GbE with iWARP in both latency and bandwidth*

But even when comparing at levels that Chelsio's adapter could withstand, Mellanox's 40GbE RoCE offering outperformed Chelsio's 40GbE iWARP in both latency and bandwidth (Figure 3).

[6] http://www.chelsio.com/wp-content/uploads/2013/11/40G-TOE-vs-NIC-Perfromance.pdf

Summary

Mellanox technologies are proven to provide the best performance in the marketplace for interconnect solutions. Moreover, Mellanox makes good on its claims by demonstrating its world-class performance, robustness, and scalability in the world's largest supercomputers and data centers. Chelsio's attempts to undermine Mellanox do not change the fact that their products do not match Mellanox in quality and performance.

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

15-5348WP
Rev 1.0