Migrating Legacy IB Networks to Ethernet

With the advent of 40GbE, and the imminent arrival of 100GbE, Ethernet today can match or exceed InfiniBand in raw speed. Coupled with mature iWARP (RDMA over Ethernet) implementations, this sets the stage for migrating compute clusters from legacy IB networks to Ethernet, without any performance penalty, while realizing all the economies of scale that an all-Ethernet environment allows.

In order to conduct a side-by-side benchmark comparison, an 8-node cluster was configured in a dual-connected fashion using an IB-FDR fabric and a 10Gbps Chelsio T440 based Ethernet fabric. The Ethernet fabric provides a 40Gbps pipe over one QSFP cable, aggregating traffic from multiple 10Gb Ethernet ports of each Chelsio T440 card. Since the T440 attaches to the host with a PCIe Gen2 x8 bus, its maximum throughput is PCIe bus limited to 28Gbps, i.e. effectively 3x10Gbps ports. The figure below shows the topology.

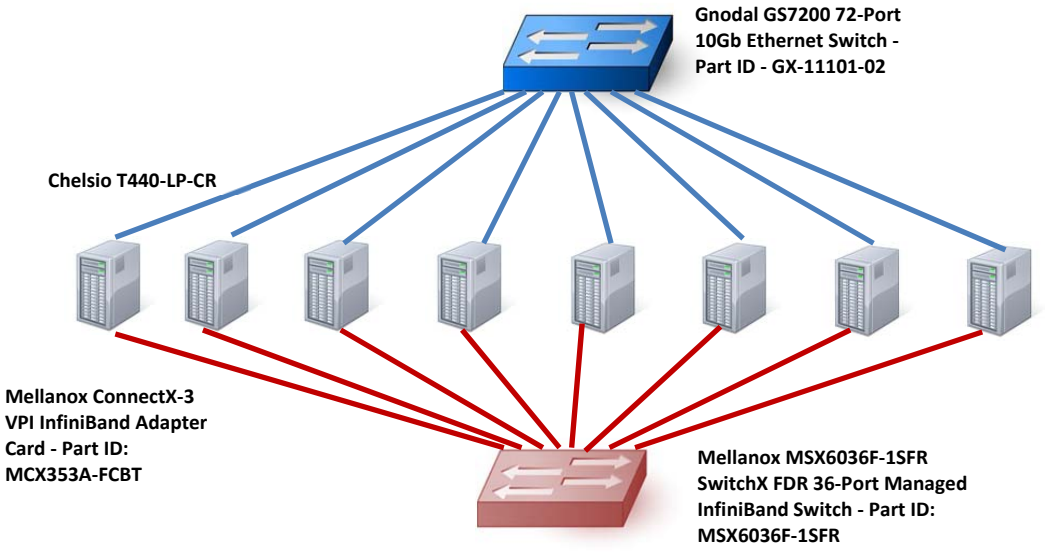

Figure 1 – Cluster Topology

Initially, in order to baseline the performance, one-way and bidirectional bandwidth on both fabrics were measured using OSU's Open-MPI benchmarks. The command line parameters are listed below.

Next, a projection was made on where the upcoming Chelsio 40GbE equivalent numbers would be based on core clock frequencies and next generation hardware simulation data.

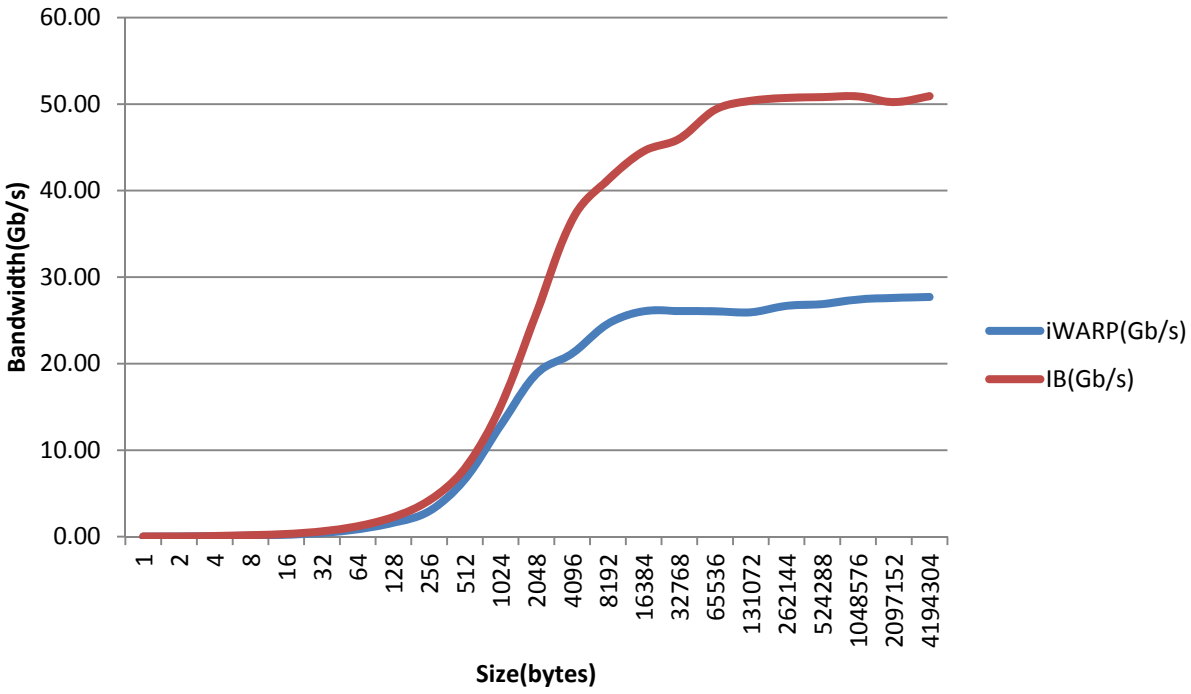

Figure 2 – IB-FDR vs. 3x10GbE iWARP RDMA

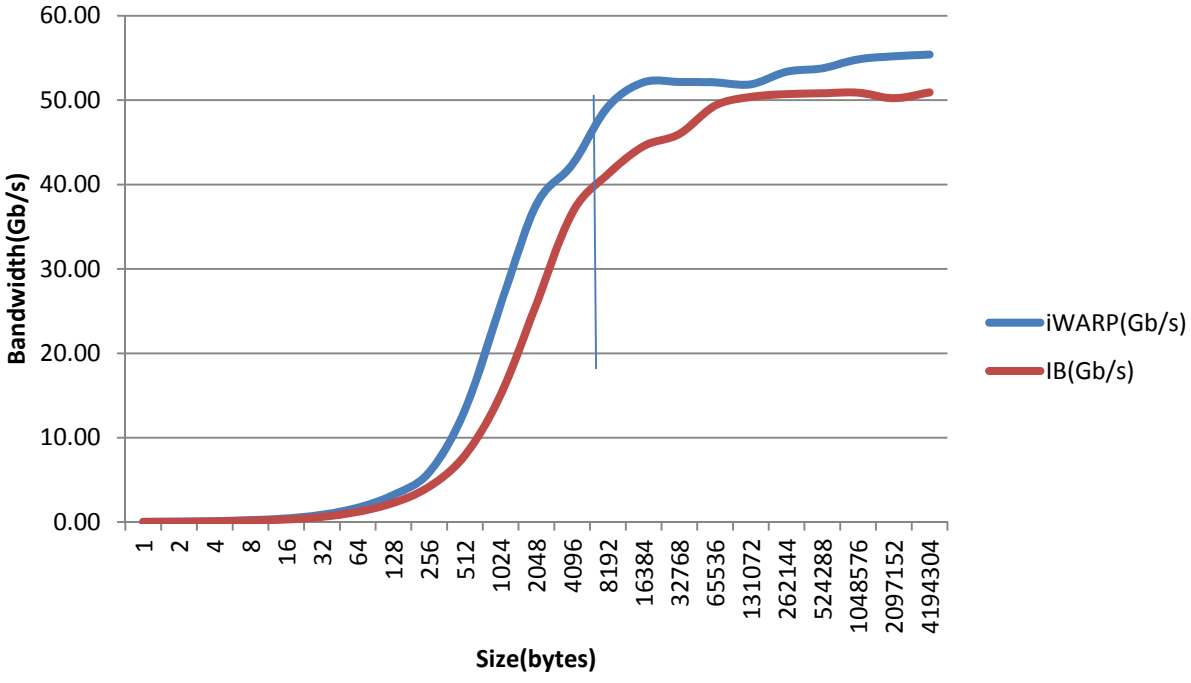

Figure 3 – Projected IB-FDR vs. 2x40GbE iWARP

OSU MPI Benchmarks
Open-MPI 1.43
CPU E5-2687W, 3.1GHz SandyBridge + Romley, 8 nodes only

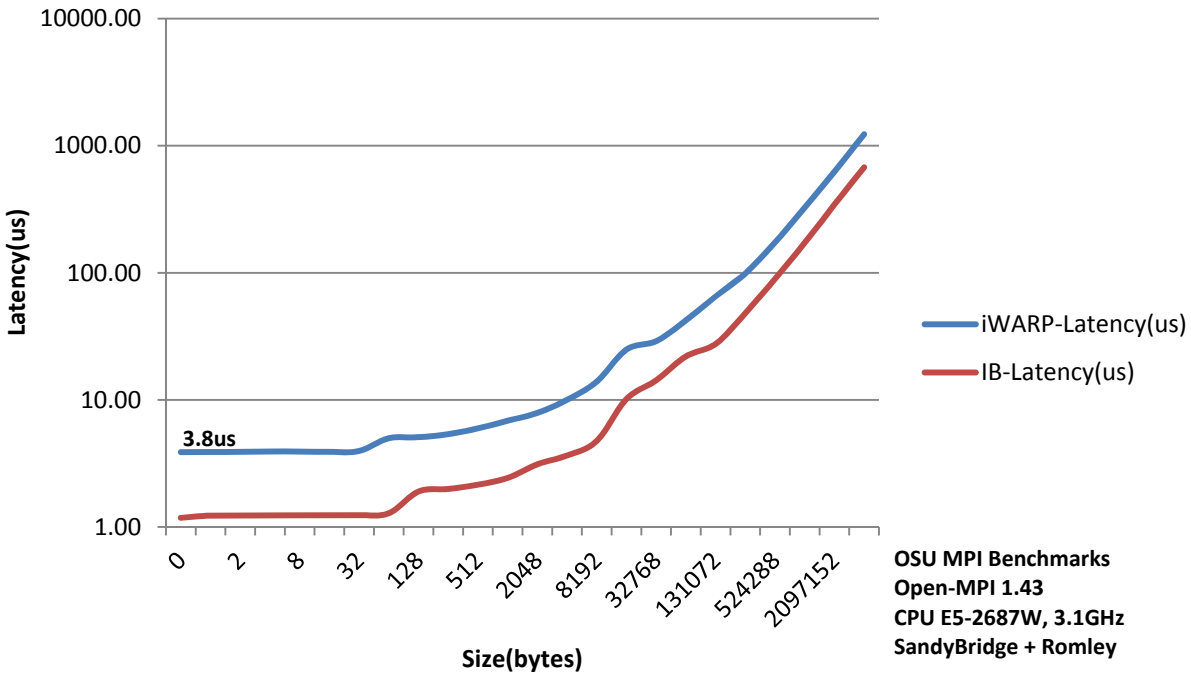

Figure 4 – Latency, IB-FDR vs. 10Gb iWARP RDMA

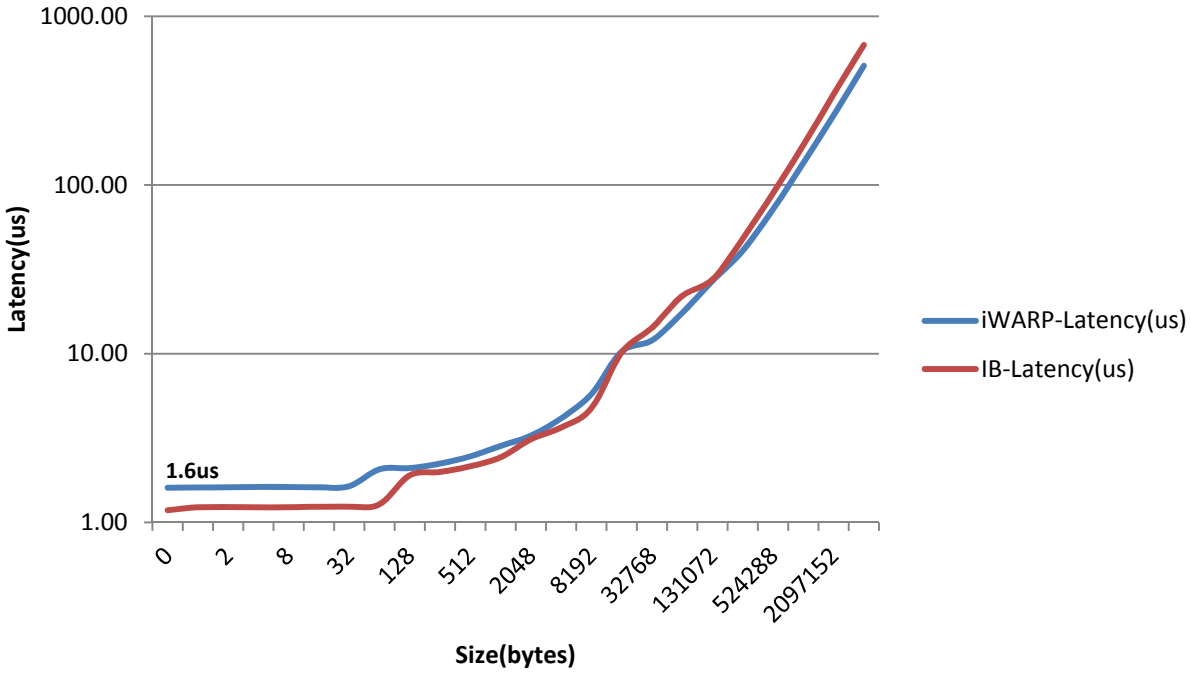

Figure 5 – Projected Latency, IB-FDR vs. 40Gb iWARP RDMA

There are several observations that are noteworthy:

1. IB-FDR is limited to 50Gb of application level data throughput, not 56Gb as advertised.
2. For up to 8K IO sizes (most common storage IO size), a single 40GbE iWARP spigot produces more bandwidth than a single IB-FDR spigot.
3. The latency gap of 40GbE vs. IB is closed past 128B-sized IO.
4. Ethernet iWARP RDMA data has been collected using the same middleware and application as IB RDMA, which clearly demonstrates the ability of iWARP to run IB applications without modification over Ethernet (utilizing the Open Fabrics OFA interface)

Also evident in the iWARP numbers, is its unique ability to saturate the communication channel by leveraging the underlying TCP Offload Engine congestion and flow control (via performing all retransmissions and exception handling at silicon speed).

In contrast, an IB based infrastructure can be severely handicapped by intrinsic noise on the cables, which results in slow path recovery. This problem becomes more significant as increases in physical link speed bring about higher relative noise levels. Because IB reliability is handled at a high layer in the IB protocol stack, a very large amount of data may have to be retransmitted to recover from the loss of one packet. This results in exacerbating the congestion condition, and as a result once congestion sets in, it causes a steep drop in performance. Studies have shown a difference of almost 6x between the best case and worst case IB performance in the presence of packet loss or congestion. In contrast, this is something that TCP/IP, and hence iWARP, does not suffer from thanks to 3 decades of refinement to handle exactly such issues. Furthermore, because all exception processing associated with iWARP or TCP is handled in silicon, the application code is more likely to remain cache resident. This allows for better application performance under load.

Going forward, given that the underlying physical layers for IB-EDR and 100GbE are the same, and given that IEEE bodies have already started to look at 400GbE as the next speed up, it is clear that there is no longer any raw bandwidth advantage to an IB fabric, to balance out the cost of specialized IT staff, gateway products, and esoteric infrastructure. Finally, IB does not take advantage of Ethernet economies of scale, and ancillary technical developments such as EEE for power management. Therefore, the only cost effective and enduring approach is to choose iWARP and Ethernet for new deployments, and to replace existing IB fabrics with Ethernet.

Figure 6 – T440-CR: 4x10GbE Half Size SFP+

Figure 7 – T440-LP-CR: 4x10GbE Low Profile QSFP