Cluster Interconnects: Single Points of Performance

Gilad Shainer, MSc., Mellanox Technologies

High-performance computations are rapidly becoming a critical tool for conducting research, creative activity, and economic development. In the global economy, speed-to-market is an essential component to get ahead of the competition. High-performance computing utilizes compute power for solving highly complex problems, perform critical analysis, or run computationally intensive workloads faster and with greater efficiency. During the time needed to read this sentence, each of the Top10 clusters on the Top500 list would have performed over 150,000,000,000,000 calculations.

HPC clusters have become the most common building blocks for high-performance computing. Clusters provide the basic needs of flexibility and deliver superior price/performance compared to other proprietary symmetric multiprocessing (SMP) systems, with the simplicity and value of industry-standard computing. A cluster's ability to perform depends on three elements – the central processing unit, the memory on a single server, and the interconnect that gathers the single servers into one cluster.

The TOP500, an industry respected report since 1993, ranks the most powerful computer systems and provides an indication on usage trends in computing and interconnect solutions. According to the TOP500, InfiniBand is emerging as the most used high-speed interconnect, replacing the proprietary or low-performance solutions. As shown below, InfiniBand shows higher adoption rate in the TOP500 comparing to Ethernet. The graph compares the adoption rate of InfiniBand and Ethernet, from the first year they were introduced to the TOP500 list (InfiniBand – June 2003, Ethernet – June 1996).

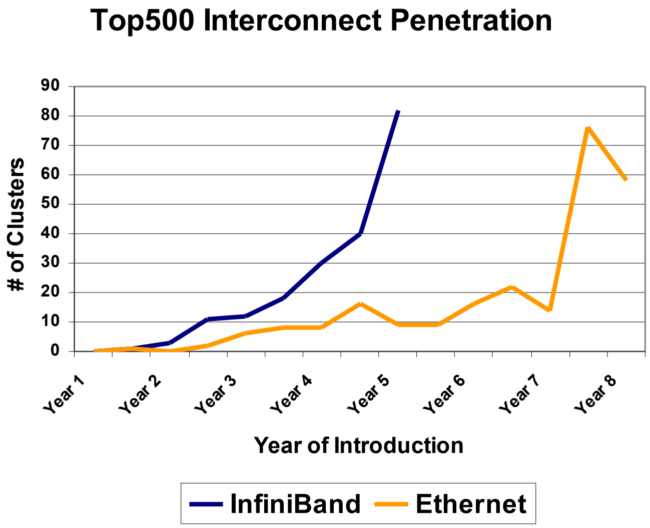


Figure One: Top500 Interconnect Penetration of Inifiniband and Ethernet

Single-points of performance

The most common approach for comparing different interconnect solutions is the "single-points" approach. Latency, bandwidth, N/2 and message rate are known as single-points of performance used to make the initial determination. When referring to bandwidth, it is typical to report the maximum available bandwidth. Message rate, defined as the numbers of messages transmitted in a period of time, is yet another single-point on the bandwidth graph. Message rate bandwidth (bytes/second) divided by the length of the message (typically done for 0 or 2 bytes) resulting in messages per second metric. Note: that a zero byte message still requires data to be sent.

N/2 benchmark provides a better understanding of the bandwidth characteristics, as it indicates the point where the interconnect achieves half of its maximum capability, therefore provides an indication on the bandwidth curve. However, the N/2 number must be followed with the corresponding bandwidth number or else it can be misleading. A higher bandwidth solution can present higher N/2 compared to a lower bandwidth one, while its bandwidth graph is higher for a single-point. An example is shown later in this paper.

Any set of single-points relate to a specific application or application interface, and each application interface has a different set of parameters associated with it. The most common application interface in high-performance computing is Message Passing Interface (MPI). Among the single-points of performance, the N/2 benchmark provides an additional indication on the interconnect's capabilities – how fast the bandwidth curve rises. Lower N/2 results in higher actual or effective bandwidth delivered by the interconnect to the applications. The following graphs compare the N/2 MPI benchmark between the two InfiniBand providers, Mellanox and QLogic.

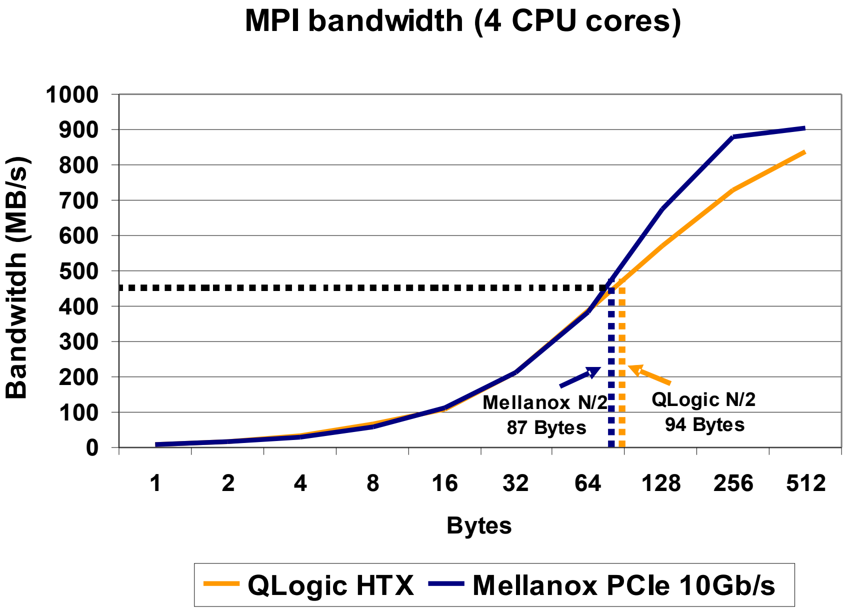


Figure Two: Mellanox 10Gb/s and Qlogic HTX N/2 Data

Mellanox's InfiniBand 10Gb/s PCIe adapter delivers lower N/2 than the QLogic HTX InfiniPath adapter, and therefore delivers overall higher, effective bandwidth. QLogic had reported N/2 of 88 Bytes on an 8 core server, yet it is still higher than the Mellanox 10Gb/s adapter on a 4 core server, as shown in the graph.

The second graph compares between the best performance solutions from Mellanox and QLogic – Mellanox's InfiniBand 20Gb/s and QLogic's InfiniPath HTX. The result is a good example of how N/2 can be misleading, if not followed by the corresponding bandwidth figures. Mellanox's 20Gb/s achieves higher N/2 of 110 Bytes versus QLogic's N/2 of 94 Bytes, but as the graph shows, Mellanox has a superior bandwidth curve, and therefore has an overall higher, effective bandwidth.

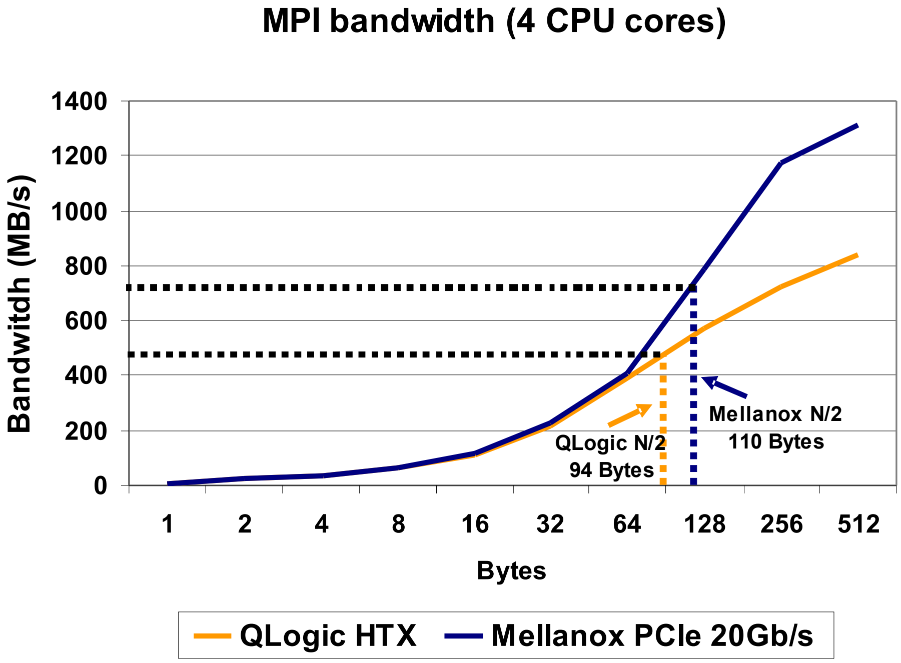


Figure Two: Mellanox 20Gb/s and Qlogic HTX N/2 Data

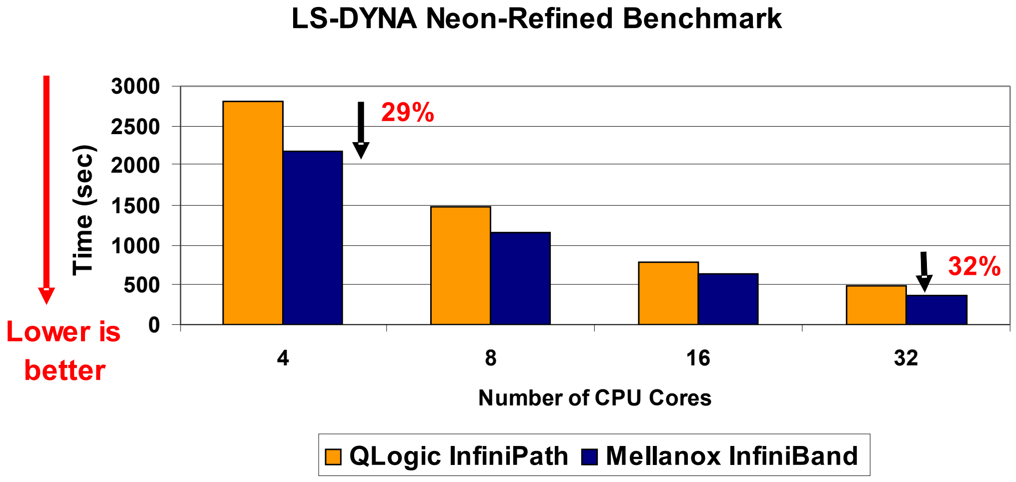
The bandwidth advantage of Mellanox increases with message size to about 50%. Mellanox achieves QLogic's maximum BW (of 954MB/s) in message size of ~200 Bytes and provides maximum uni-directional bandwidth of more than 1400 MB/s (Mellanox adapter is capable of almost 2000MB/s uni-directional bandwidth and limited by the current system components).

Real world application performance

The sets of single-points have traditionally been used as the prime metric for assessing the performance of the system's interconnect fabric. However, this metric is typically not sufficient to determine the performance of real-world applications. Real-world applications use a variety of message sizes and a diverse mixture of communication patterns. Moreover, the interconnect architecture becomes a key factor and greatly influences the overall performance and efficiency.

Mellanox's adapter architecture is based on a full offload approach with RDMA capabilities, reducing the traditional protocol overhead from the CPU and increasing processor efficiency. QLogic's architecture is based on an on-load approach, where the CPU needs to deal with the transport layer, error handling etc., and therefore increases the overhead on the CPU and reduces processor efficiency, leaving less cycles for useful application processing.

The following chart compares Mellanox InfiniBand and QLogic InfiniPath interconnects using LD-DYNA Neon-Refined benchmark (frontal crash with initial speed at 31.5 miles/hour, model size 535k elements, simulation length: 150ms - model created by National Crash Analysis Center (NCAC) at George Washington University).



Mellanox on Intel Xeon 5160, QLogic on AMD Opteron 2.6GHz

Figure Three: LS-DYNA Neon-Refined Benchmark for Mellanox and Qlogic

Using a best case hardware scenario, Mellanox InfiniBand shows higher performance and better scaling compared to QLogic InfiniPath (Note the improvement from 4 cores to 32 cores). Livermore Software Technology Corporation's (LSTC) LS-DYNA (general purpose transient dynamic finite element program capable of simulating complex real world problems) is a latency-sensitive application. While QLogic shows lower latency, as a single-point of performance, Mellanox's architecture delivers higher system performance, efficiency and scalability.

It is difficult, and sometimes misleading, to predict real-time application performance with just single-points of data. In order to determine the system's performance and the interconnect of choice, one should take into consideration a set of metrics including the single-point of performance (bandwidth, latency etc.), architecture characteristics (CPU utilization, overlap capabilities of computations and communications scalability etc.), applications results, field proven experience and hardware reliability.

In order to provide better applications sight, Mellanox has created the Mellanox Cluster Center. The Mellanox Cluster Center offers an environment for developing, testing, benchmarking and optimizing products based on InfiniBand technology. The center, located in Santa Clara, California, provides on-site technical support and enables secure sessions onsite or remotely. More details can be achieved through Mellanox web site (*http://www.mellanox.com*).

The author would like to thank Sagi Rotem, Gil Bloch and Brian Sparks for their input during reviews of this article.

Gilad Shainer is a senior technical marketing manager at Mellanox technologies focusing on high performance computing. He joined Mellanox Technologies in 2001 to develop Mellanox's InfiniHost PCI-X Host Channel Adapter (HCA) device and later led the development of Mellanox's InfiniHost III Ex PCI Express HCA device. Gilad Shainer holds MSc. degree (2001, Cum Laude) and a BSc. degree (1998, Cum Laude) in Electrical Engineering from the Technion Institute of Technology in Israel. He is also a member of the PCISIG PCI-X and PCI Express Working Groups and has contributed to the definition of the PCI-X 2.0 specifications.