White Paper

Mellanox and HPC Clustering

Greater than 850 MB/sec of MPI Performance

1.0 Introduction

Clusters of low cost industry standard architecture servers have emerged as a viable alternative to proprietary Symmetrical Multi Processor (SMP) server systems for high performance computing applications. Building such clusters requires a high performance interconnect technology offering high bandwidth, low latency, and direct communication between user space processes on individual nodes. Until now available clustering solutions have been confined to proprietary technologies, which suffer from inadequate product robustness and reliability, poor range of product offerings, limited choice of vendors, small market scale, and lack of competition. Also, standardized, Ethernet solutions have proven woefully inadequate in providing delivered bandwidth due to CPU overhead and high latencies. The InfiniBand℠ Architecture is a powerful new industry standard technology that advances I/O connectivity for enterprise database and high performance computing clusters.

Today Mellanox HCAs deliver MPI bandwidth 8 times or more than that of Ethernet and up to 3 times the bandwidth of the current HPC clustering technologies.

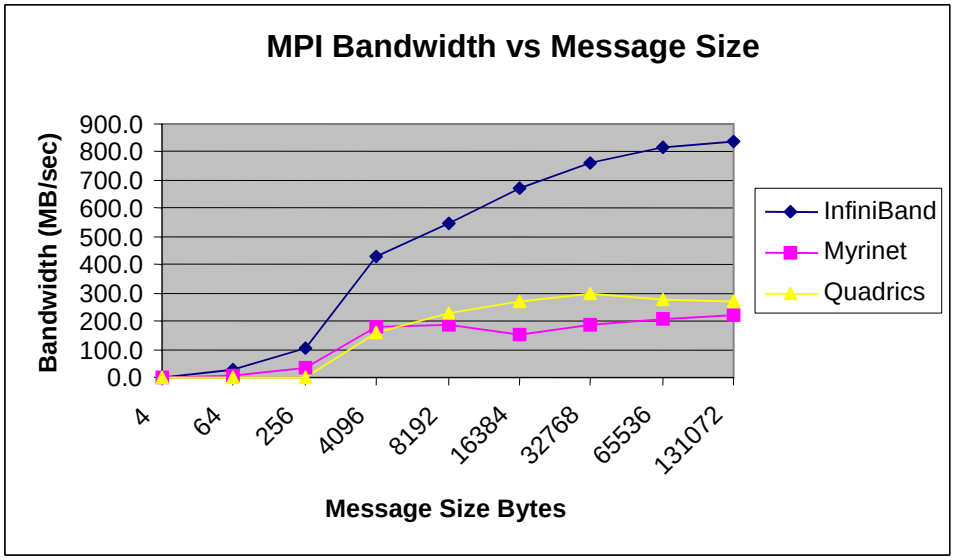


Figure 1. MPI Bandwidth (Source Ohio State University)

Updated March 2003

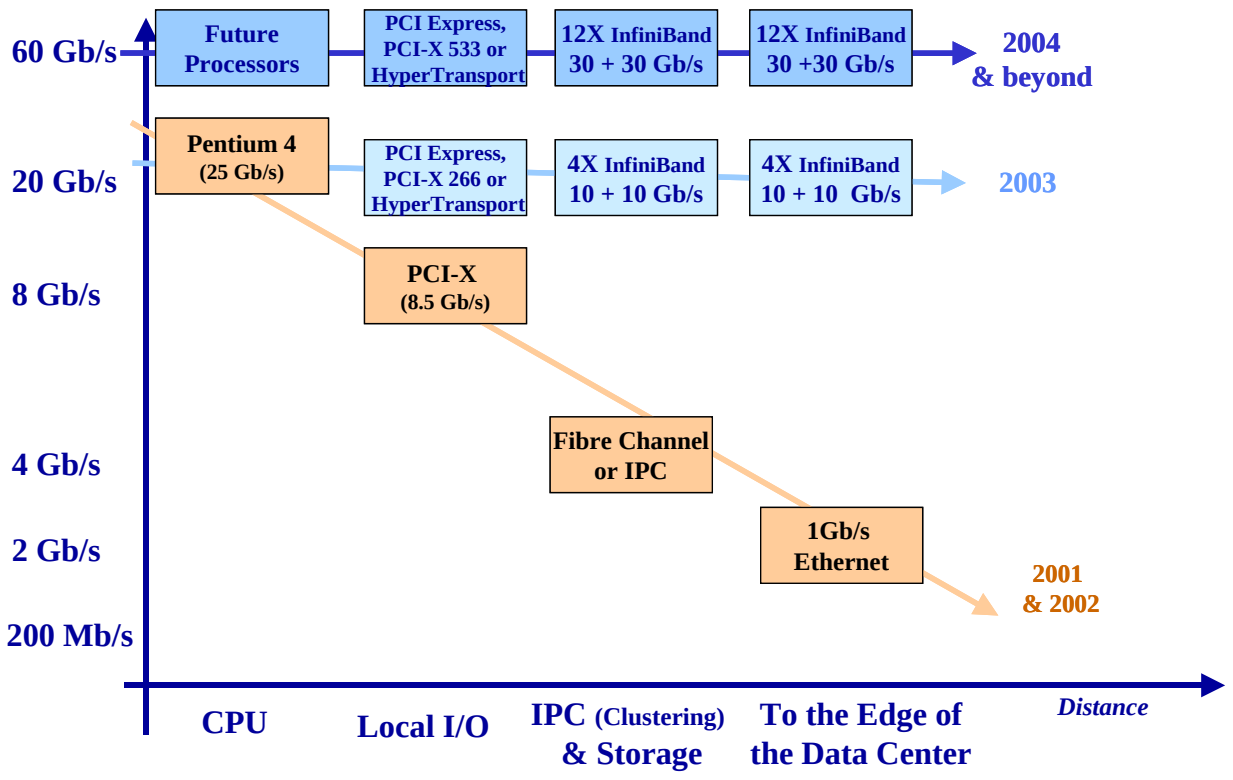
2.0 Why InfiniBand: “Bandwidth Out of the Box”

There are a myriad of reasons to use InfiniBand as a server clustering interconnect for data centers or HPC. The architecture is based on a specification of over 2,000 pages of hard work from some of the top engineering talent from most of the major server, software and storage OEMs, as well as, HPC research and development teams from major labs, university research departments, and the top InfiniBand companies. At the highest level, InfiniBand offers all the key features needed for a highly reliable clustering, storage and communication interconnect while providing a key feature for the future, delivered bandwidth that keeps pace with the processor.

Today’s data centers shown in brown in Figure 2, “Mellanox and HPC Clustering,” use the latest processors and server sub systems, but the reality is that there is more than an order of magnitude loss of bandwidth as data is moved from the processor to the edge of the data center.

The InfiniBand Architecture overcomes this limitation and, when combined with new server chip sets available in 2003, InfiniBand will enable full duplex bandwidth communication of up to 20Gb/s all the way from the processor to the edge of the data center. Thus InfiniBand delivers **“Bandwidth Out of the Box.”** This is the key feature that enables high performance computing clusters to be assembled from industry standard server architectures.

Figure 2. Mellanox and HPC Clustering



InfiniBand is Future Proof: The modular design of Mellanox InfiniHost HCA enables rapid adaptation to match future server chip sets, whether PCI-X 2.0, PCI-Express, or other interfaces. Furthermore, the InfiniBand industry standard has already approved the 12X standard for 60 Gb/sec

of bandwidth, enabling Mellanox and the other major InfiniBand silicon suppliers to keep InfiniBand a generation ahead of the competition.

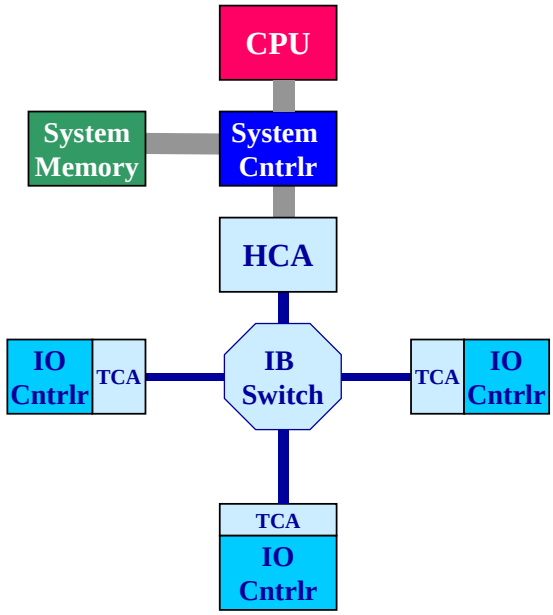
3.0 InfiniBand Architecture (IBA) Overview

Based on high-speed switched serial links that scale to 30Gb/s, InfiniBand delivers a complete clustering solution that overcomes the bottlenecks of traditional server networking and offers all the benefits of a widely supported industry standard. The InfiniBand Architecture offers kernel bypass mechanisms that allow Host Channel Adapters (HCAs) to transmit and receive data directly from user space processes without the requirement to involve the operating system kernel. The InfiniBand Architecture offers both message passing “send” mechanisms optimized for small data transfers, as well as remote direct memory access (RDMA) capabilities providing more efficient transfers of large data blocks. The wire protocols, software mechanisms and low-level electrical, as well as, physical details of InfiniBand are rigorously defined by the 1.0a specification thereby facilitating multi-vendor interoperability. The InfiniBand Trade Association Compliance and Interoperability Work Group holds frequent PlugFests to ensure this interoperability priority is achieved.

A wide range of industry standard test equipment and off-the-shelf management software and test equipment is available (protocol analyzers, signal generators, logic analyzers, etc.) to evaluate InfiniBand fabrics. These sophisticated tools make it possible for system architects to monitor and fine-tune cluster performance.

There are three basic building blocks used in creating an InfiniBand switched computing fabric: HCA (Host Channel Adapter), TCA (Target Channel Adapter) and Switches. HCAs are installed into the server and initiate communications within the fabric. TCAs are generally native InfiniBand storage units, InfiniBand to Ethernet or Fibre Channel I/O devices. Switches can be either managed or unmanaged but unlike many other fabrics, the management for the fabric can be implemented either on the switch or as host software controlling and monitoring the fabric through an HCA.

Figure 3. Three Basic Building Blocks of InfiniBand Fabric



4.0 Mellanox Products

Mellanox offers a complete line of InfiniBand silicon devices to enable an InfiniBand fabric. Our OEM partners use these devices to produce HCA cards, routers and switches.

Figure 4. Mellanox Silicon Product Family

- InfiniHost™: Second Generation Dual Port 10Gb/sec HCA
- InfiniScale™: Second Generation Eight Port 10Gb/sec Switch
- InfiniBridge™: First generation multi-purpose device: Dual Port 10 Gb/sec channel adapter or 8-Port 2.5Gb/sec switch.

The above mentioned three products offer industry leading features and performance. For detailed information on these products visit: www.mellanox.com/products.

To enable early evaluation of our silicon, Mellanox also offers a number of reference design cards, switches and InfiniBand blades to improve time-to-market for OEM partners. These reference designs are ideal to develop InfiniBand clusters or fabrics for testing and evaluation.

5.0 Reference Designs:

5.1 HCA Cards:

Figure 5. InfiniHost MTPB23108 Reference Design

Mellanox offers the InfiniHost MTPB23108 reference design card. This PCI-X card offers dual 4X ports, hardware transport, low latencies and 128 MB (or more) of HCA memory.

5.2 Switches:

Mellanox offers three switches for building HPC clusters:

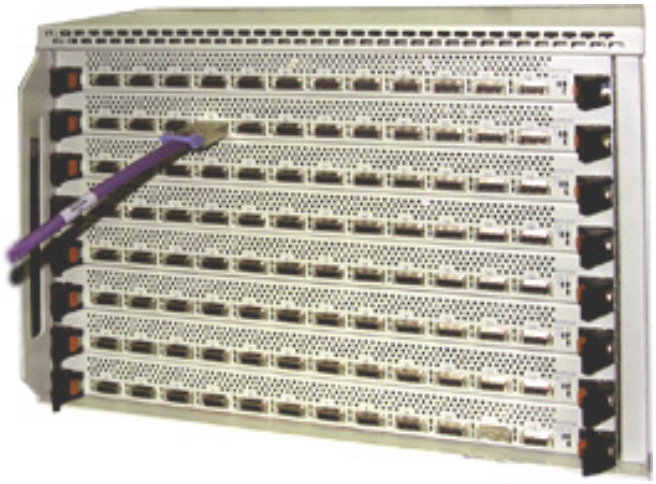
- MTEK43132-C08-S: This is an 8-Port 10Gb/sec 1U switch based on the highly integrated InfiniScale device. The switch implements a full wire speed, non-blocking design that features latencies of less than 200 ns.

Figure 6. MTEK43132-C08-S Switch.

- MTEK43132-M16-S: Modular 16-Port InfiniScale Switch Evaluation Kit in a1U chassis. The switch implements a two-stage constant bi-directional bandwidth (CBB) fat tree topology. The modular design allows for up to 16 copper or optical ports or a combination of the two.

Figure 7. MTEK43132-M16-S Switch.

- 96-Port HPCC Switch Design: This modular switch design features a 7U chassis that accommodates up to 8 leaf boards with twelve 10Gb/sec ports each that are arranged in a CBB (Constant Bi-sectional Bandwidth) or fat tree topology. The design can scale from as few as 12 ports (a single leaf) up to a total of 96 ports (with eight leafs). All ports are 4X or 10Gb/sec. Each port is capable of up to 20 Gb/sec of cross sectional bandwidth that realizes an unprecedented 1.92 Terabits of total bandwidth.



6.0 "HPC Cluster in a Box" a Vision for HPC Blade Computing

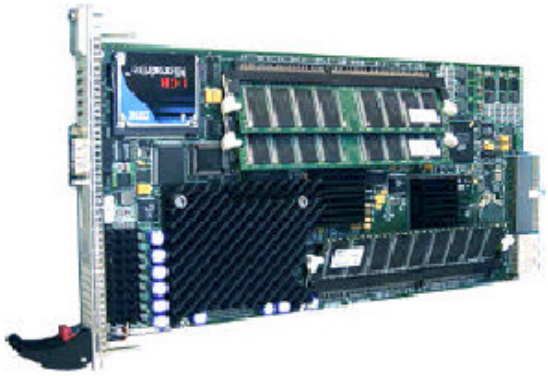
Figure 8. Nitro II Server Blade

"HPC in a Box" is the concept of delivering all the best attributes of the InfiniBand architecture in a self-contained server blade form factor that offers simplified cabling, reduced floor space, higher density and higher reliability. Mellanox Nitro II server blade reference design provides a vision of simplified high performance computing "HPC in a Box" by delivering dramatically improved clustering performance in a highly integrated, reliable, and compact package. SC2002 (Super Computing Conference, Baltimore Ma, November 2002) was the watershed event for the InfiniBand Trade Association as a multitude of vendors demonstrated native MPI InfiniBand benchmark performance results, running various applications and offering many InfiniBand card, switch, system and software solutions.

The Nitro II reference design provides a vision of the future; demonstrating high-bandwidth low-latency clustering with 10Gb/sec blades, backplane and switch all in an efficient and easy to use compact blade design able to support a 48 node Pentium 4 cluster in less than half the space of traditional server form factors. An entire 12-node HPC cluster is housed in a single 4U enclosure that can seamlessly scale to 48 nodes without the need for any Mellanox and HPC Clustering external switch. Using 8 or 16 port CBB switches enables scaling to clusters of node counts of

128, 256 and beyond. The chassis includes fully redundant switches and backplane connectivity, enabling high reliability. At SC2002 Mellanox Nitro II InfiniBand server blades were being demonstrated by Sandia Labs in the Ohio State University booth, MPI Software Technology, Microway and Abba.

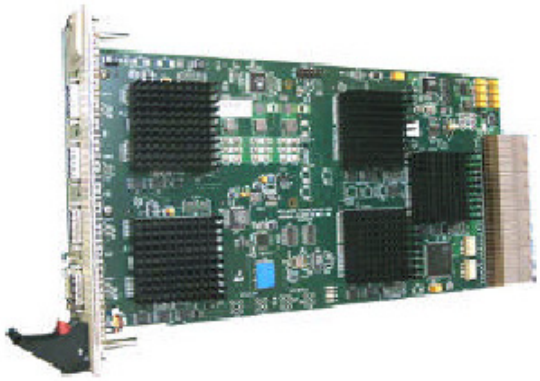
Figure 9. Nitro II Server Blade



The Nitro II server blade is a 4X InfiniBand Reference Design featuring: the Mellanox MT23108 InfiniHost HCA, 2.2Ghz Intel® Xeon® Processor, ServerWorks™ GC-LE chipset, 256 MB of HCA memory and 1Gbyte of system memory.

The 16-Port 10Gb/sec InfiniBand switch blade reference design (Nitrex II) features four front panel 4X ports and twelve 4X backplane ports.

Figure 10. Nitrex Switch



7.0 InfiniBand MPI Support

Four software sources have announced InfiniBand MPI support for Mellanox InfiniHost device at SC2002:

- MPI Software Technology Inc.: MPI/Pro is providing high performance MPI-1.2 parallel middleware for InfiniBand. MPI/Pro for the InfiniHost HCA is optimized for both low-latency and low-overhead configurations, offering maximum bandwidth for both settings of the library.
- Ohio State University: OSU is providing a version of the OSU MPI on the InfiniBand Mvapich library.
- NCSA: NCSA is providing a version of NCSA's MPI for InfiniBand.
- Scali: Next generation cluster management, Scali Manage™ and message passing interface, Scali MPI Connect ™ for InfiniBand.

These sources offer the HPC clustering community a choice in their MPI selection that are all engineered to run over Mellanox InfiniHost HCA.

8.0 HPC Environments

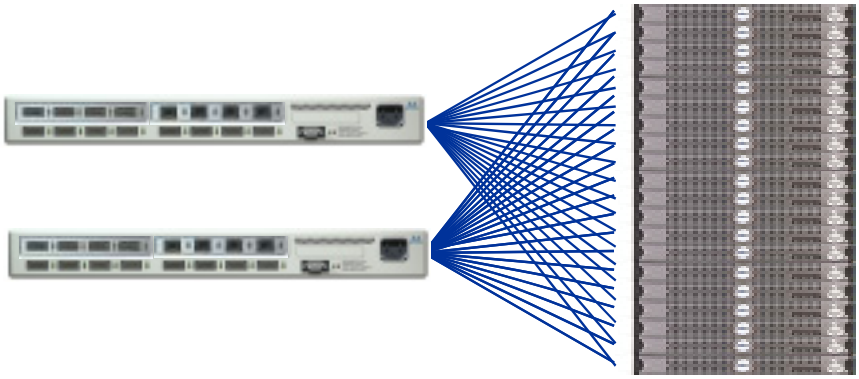
Using the previously described InfiniBand reference designs it is straightforward to build InfiniBand clusters for HPC applications. The three most common ways to achieve InfiniBand clusters are:

1. Upgrade Existing Servers with InfiniBand HCA cards and switches.

Figure 11. Switch and HCA Card

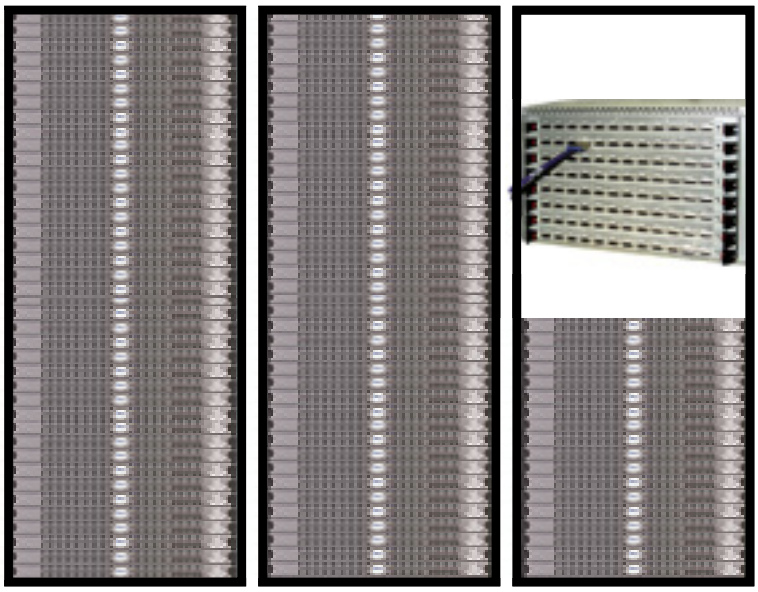
2. Create fully redundant InfiniBand Clusters with HCAs, switches and new server deployment.

Figure 12. Sixteen Node Fully Redundant IB Cluster



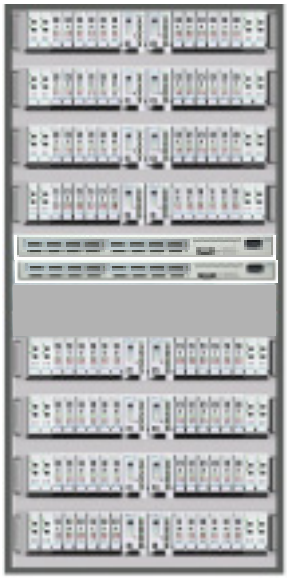
3. Create 32 to 64 to multi-thousand node HPC clusters with the 96-Port switch design.

96-Node Example



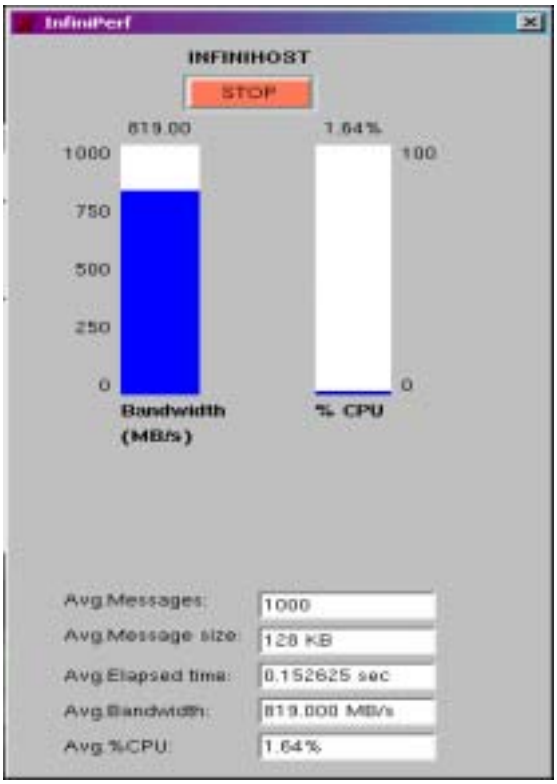
4. “HPC in a Box” utilize InfiniBand Server Blades in groups of 12 to 96 blades in a single rack. Shown here are 8 Nitro II enclosures with 96 server blades using two 16-Port switches to enable a CBB cluster of the server blades.

Figure 13. HPC in a Rack



9.0 Hardware Performance Results

Figure 14. Hardware Performance Results

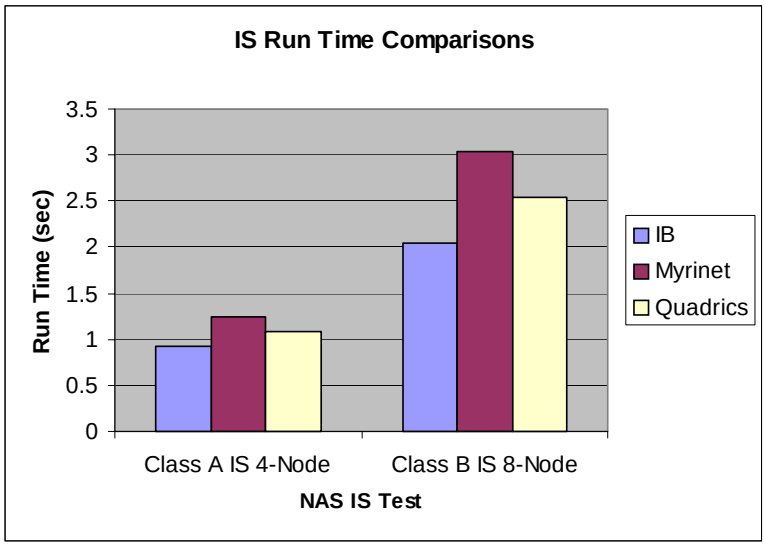


Mellanox has achieved excellent HCA hardware performance. At the Intel Developer Forum (IDF) in September 2002, Mellanox demonstrated greater than 6.5 Gb/sec (~ 820 MB/sec) of data bandwidth in a Verbs level performance test. Shown is a screen shoot that provides both the bandwidth and the processor utilization of less than 2% based on a 128KB message size.

9.1 Application and MPI results over InfiniBand

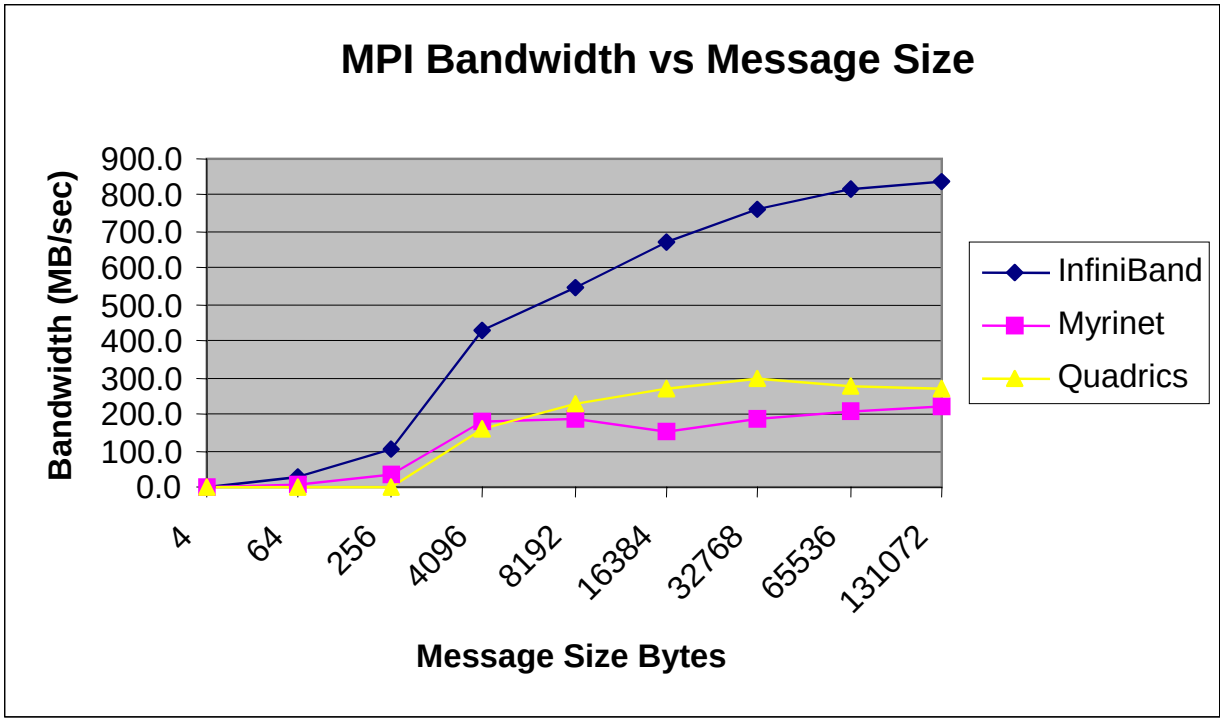
HPC clustering has achieved a new level of performance. Mellanox and our partners, including MSTI, Lane 15, OSU and NCSA are experiencing impressive Application level results. Shown here is the NAS IS Test. It can be seen that Myrinet runs 49% slower on an 8-node cluster than InfiniBand on the NAS IS Test.

Figure 15. IS Run Time Comparisons



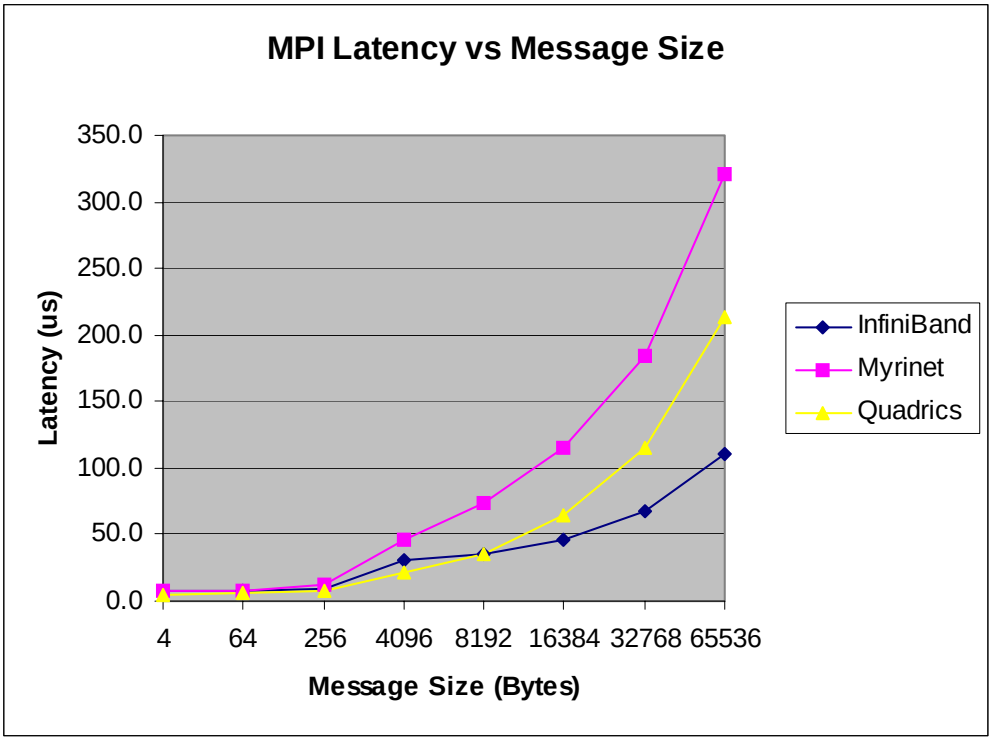
MPI bandwidth results shown in Figure 16, “Bandwidth vs. Message Size,” have topped 850 MB/sec.

Figure 16. Bandwidth vs. Message Size



InfiniHost HCA latencies are superb. The initial results over MPI show excellent results across the board. Zero byte latencies at 7.0 usec and once the message size increases beyond 16KB, InfiniHost's 10Gb/sec transfer really takes hold and clearly greatly improves on latencies for the market place. These latencies will continue to be improved with firmware and driver optimizations and as future HCAs can connect to and exploit new high speed local I/O interfaces (like PCI-X 2.0 & PCI-Express) on new server chipsets.

Figure 17. .MPI Latency vs Message Size



10.0 Demonstration Partners and Early Adopters

At SC2002 the following partners and early adopters have assisted Mellanox with HPCC demonstrations:

- Abba Technologies
- Ames Laboratory
- Appro International Inc.
- DivergeNet
- Intel Corporation
- JNI Corporation
- Lane 15 Software
- Los Alamos Laboratories

- MicroWay Inc.
- MPI Software Technology Inc. (MSTI)
- NCSA (National Center for Supercomputing Applications)
- Ohio State University and Ohio Supercomputer Center (OSC) with Sandia Labs
- Rack Saver Inc.
- Topspin Communications, Inc.

For InfiniBand product such as HCA cards, switches, and I/O systems look to Mellanox silicon partners including, InfiniCon Systems, JNI Corp, DivergeNet, First Star Networks, Voltaire, InfiniSwitch, Top Spin, and MicroWay. Also, for commercially available InfiniBand software support look to Lane 15 Software, MPI SoftTech Inc., Scali and others.

References

1. Introduction to InfiniBand, http://www.mellanox.com/technology/shared/IB_Intro_WP_180.pdf
2. InfiniBand Specification V1.0.a, http://www.infinibandta.org/
3. Nitro II InfiniBand Server Blade Reference Design, http://www.mellanox.com/
4. Ohio State University, Parallel Architecture and Communication (PAC) Research Group, Department of Computer and Information Science. http://www.cis.ohio-state.edu/~panda/pac.html