White Paper

Mellanox and HPC Clustering

Enabling TeraFlop Computing at a Fraction of the Cost

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 12 times or more than that of Ethernet and up to 6 times the bandwidth of the current HPC clustering technologies. Furthermore, Mellanox switch silicon provides the highest level of integration and scalability - over 480Gbps of non-blocking switching bandwidth in a single integrated device.

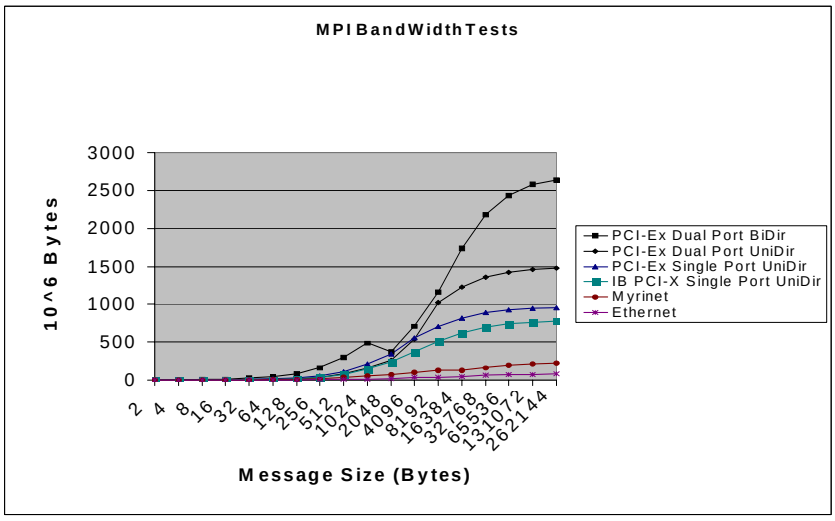

Figure 1. InfiniBand MPI Bandwidth (PCI-Express HCA) Courtesy DK Panda, Ohio State and NCSA

2.0 Why InfiniBand: "Bandwidth Out of the Box"

There are many compelling 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

Updated March 2004

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 all levels, InfiniBand offers 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.
With the introduction of the Mellanox InfiniHostIIIex PCI express HCA, the InfiniBand architecture now enables a fully balanced architecture that is able to keep pace with new CPU and memory performance. Combined with ultra-low latency and multi-protocol features like QoS and in-band management, InfiniBand provides the most scalable and efficient interconnect for now and in the forseeable future.
Figure 2 illustrates this balanced architecture with the **"Bandwidth Out of the Box"** concept. The chart shows how Mellanox and InfiniBand are today delivering full duplex bandwidth communication of up to 20Gb/s all the way from the processor to the edge of the data center. Furthermore, new versions of the InfiniBand Specification are working on InfiniBand Double Data Rate (DDR) and Quad Data Rate (QDR) to maintain this balance into the future. This guarantees a return on investment in InfiniBand technology and is a key motivation of InfiniBand adopt in today HPC and enterprise market..

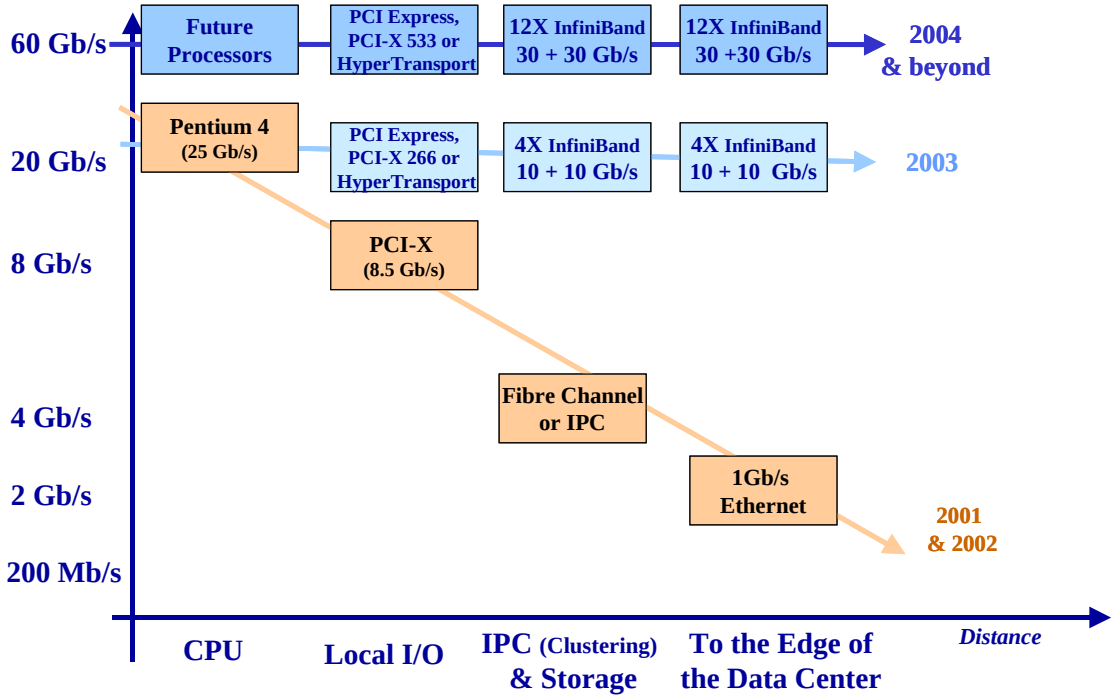

Figure 2. InfiniBand provides a balanced architecture by providing bandwidth to the edge of the datacenter

Mellanox is committed with an aggressive silicon roadmap to achieve this vision. The modular design of Mellanox InfiniHost HCAs enables rapid adaptation to match future server chip sets, whether PCI-X 2.0, PCI-Express, or other interfaces. Our Switch Silicon, embedded SerDes and other critical elements of our solution are architected with the future speed upgrades in mind. This strong roadmap gives both the HPC and Enterprise markets confidence to invest in InfiniBand technology for today's systems.

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 hardware transport and kernel bypass mechanisms that enable 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.

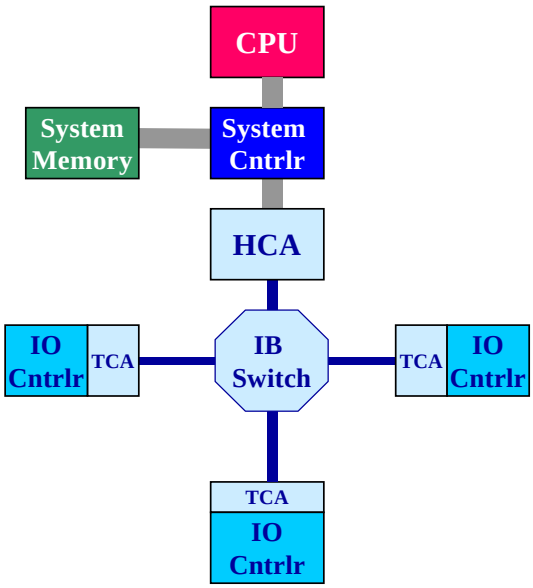

Figure 3. Three Basic Building Blocks of InfiniBand Fabric

4.0 TOTS: TeraFlops Off-The-Shelf

TOTS or TeraFlops Off-The-Shelf describes the significant advantage of having an Industry Standard for a High Performance interconnect. TOTS demonstrates how simple and cost-effective it is to deploy a cluster of 128 or more InfiniBand server nodes with over 1 TeraFlop of computing performance. TOTS is a play on the term: COTS for Commercial-Off-The-Shelf products. This term is widely used in the embedded market to identify industry standard products that enable customer investments in NON-proprietary solutions. InfiniBand is a COTS technology that is enabling the clustering of industry standard servers into powerful compute systems.

The TOTS demonstration at SC 2003, the Virginia Tech 1105-node cluster, the Sandia 128-node Intel® Xeon® cluster and other clusters illustrate a huge step forward in computation power. The InfiniBand clusters clearly demonstrates the ability to deploy a huge computation power, simply, easily and affordably.
At 1 trillion mathematical operations per second, a TeraFlop was once the Holy Grail of supercomputing performance. Just two years ago there were only 12 computers in the whole world that ranked over 1 Teraflop and on average these systems cost more than $20M per teraflop. Thanks to the InfiniBand architecture and fast processors, it is now possible for almost any university, lab, or commercial research facility in the world to deploy Teraflop levels of computing power to solve truly complex issues. It is estimated that a One TeraFlop cluster can be deployed today for well less than $1M. The most recent example is at Virginia Tech which recently deployed a 10+ Teraflop InfiniBand cluster from industry standard servers for only $5.2M, at only $600K per Teraflop.
InfiniBand and extremely fast processors in today's standard servers create an inflection point in computing. The high bandwidth, low latency 10 Gb/sec InfiniBand interconnect enables today's industry standard servers to be clustered together, with open software, to create massive amounts of compute power efficiently and inexpensively. Not only that, these powerful clusters can be depolyed quickly, in just a matter of weeks, as all the components are readily available in the market.
TOTS capability is critical to the market today as greater computation power is needed by many industry mission critical applications such as automobile engine design, crash simulation, oil and gas exploration, drug discovery, genome

analysis, computational fluid dynamics, chip design verification, movie renderings and others. In addition, many labs & universities also have an insatiable appetite for power needed in research projects such as weather simulations, geophysics, nanoscale electronics, medical research, molecular modeling, applied physics and many others.
TOTS clusters run the Linux operating system and the widely supported application layer: MPI (Message Passing Interface). There are both open source and commercial MPI solutions available for these many applications. Many clustering tools are also available that make it easy to create one master software image and replicate it across the servers in the cluster. Plus, many management tools for the cluster management.
With standard servers and InfiniBand, it is now possible for many more researchers to build and obtain greater computing resources. What was nearly unimaginable compute power just two years ago is now within a budget that most research organizations can afford. For well funded, commercial and government research programs it is now even feasible to deploy systems achieving more than 10 TeraFlops performance. With more than a ten-fold reduction in the cost per TeraFlop, more researchers can apply more compute power to the critical research issues that the world needs addressed today.

5.0 Mellanox Products

Mellanox offers a complete line of InfiniBand silicon devices, software, and enabling products to support InfiniBand clusters. Our OEM partners use this technology to produce complete HCAs, routers, switches, and software end customer solutions..

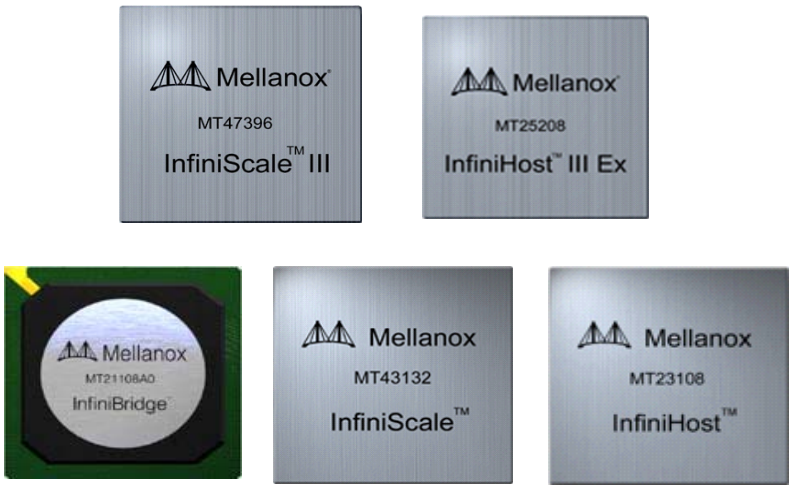

Figure 4. Mellanox Silicon Product Family

- InfiniHost III Ex: Third Generation Dual Port 10Gb/sec 8X PCI-Express HCA
- InfiniScale™ III: Third Generation 24-Port 10Gb/sec and the first 30 Gb/sec 8-Port Switch
- InfiniHost™: Second Generation Dual Port 10Gb/sec PCI-X 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 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 product design cards, switches and other InfiniBand solutions to improve time-to-market for OEM partners. These designs are ideal to develop InfiniBand

clusters or fabrics for testing, evaluation or general availability through our OEM partners.

6.0 Mellanox Designs

6.1 HCA Cards

6.1.1 MTLP25208

Mellanox offers the MTLP25208: an industry-leading, low profile PCI-Express HCA card. The HCA is based on the InfiniHost III Ex chip and has 8X PCI-express, Dual 10Gb/s ports and 128 MB (or more) of HCA memory.

Figure 5. InfiniHostIIIex MTLP25208 Low Profile PCI-Express HCA

6.1.2 MTLP23208

Mellanox also offers the InfiniHost MTPB23108 and low profile MTLP23108 dual port HCA PCI-X cards. This PCI-X card offers dual 4X ports, hardware transport, low latencies and 128 MB (or more) of HCA memory. It is available in both the standard and low profile form factors.

Figure 6. InfiniHost MTLP23108 Low Profile PCI-X HCA Card

6.2 Switches

Mellanox offers multiple switches for building HPC clusters:

6.2.1 MTEK43132-C08-S

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

Figure 7. MTEK43132-C08-S Switch

6.2.2 MTS2400

This 24-Port InfiniScale III Switch features 24-Ports at 10 Gb/sec or 8-Ports at 30 Gb/sec or a combination of the two.

The switch utilizes only one InfiniScale III device to achieve 24-ports of wire speed non-blocking communicaiton. The 1U design includes dual hot swap power supplies (second supply is optional), and a hot swap fan tray. Mellanox also offers a 8-Port 30 Gb/sec version to our OEM partners.

Figure 8. MTS2400 Switch Shown are the 24-Port 10 Gb/sec and 4-Port 12X & 12-Port 10 Gb/sec configurations

6.2.3 MTS9600 InfiniBand 96-Port HPCC Switch Design

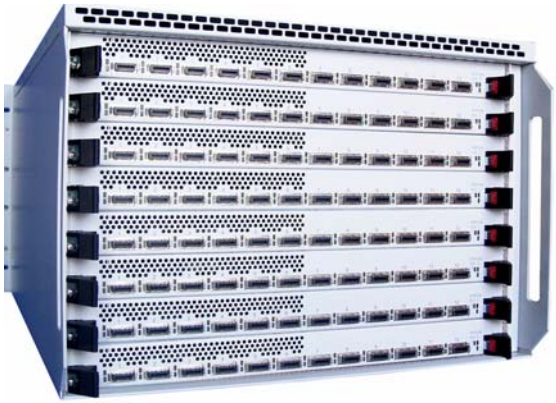

Figure 9. Mellanox 96-Port InfiniBand Switch

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.

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

"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. This server blade concept 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.

Figure 10. InfiniBand Server Blades

The concept 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 the 24 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.

8.0 InfiniBand MPI Support

Multiple software sources have announced InfiniBand MPI support for Mellanox InfiniHost:

- 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: The Nowlab at OSU is providing a version of MPI called MVAPICH for InfiniBand .
- 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 and proven to run over Mellanox InfiniHost HCA.

9.0 HPC Environments

Using the previously described InfiniBand 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.

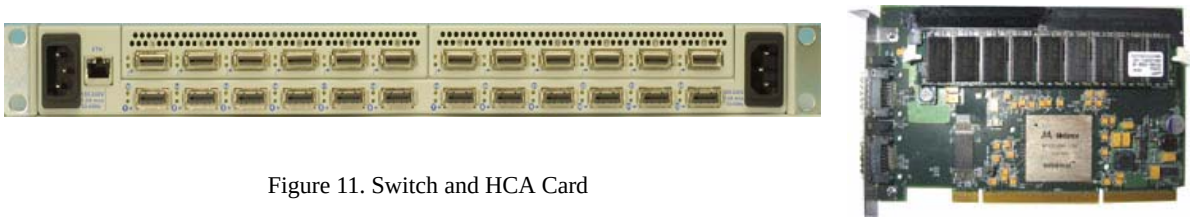

Figure 11. Switch and HCA Card

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

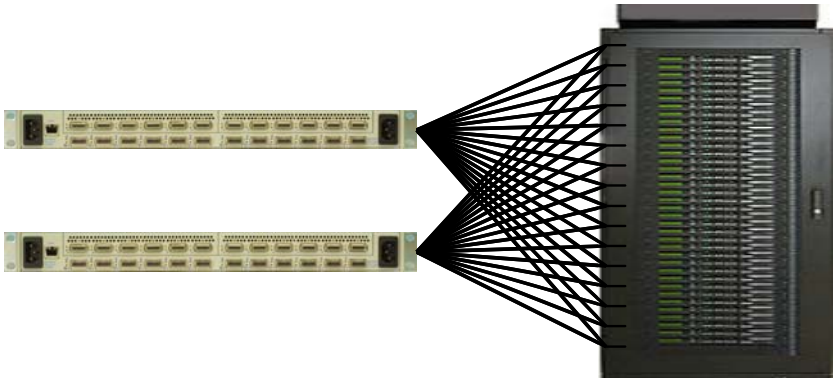

Figure 12. 24-Node Fully Redundant IB Cluster using only two InfiniScaleIII switches

3. Create 96 to thousand node (or MORE) HPC clusters with any combination of HCAs and switches.

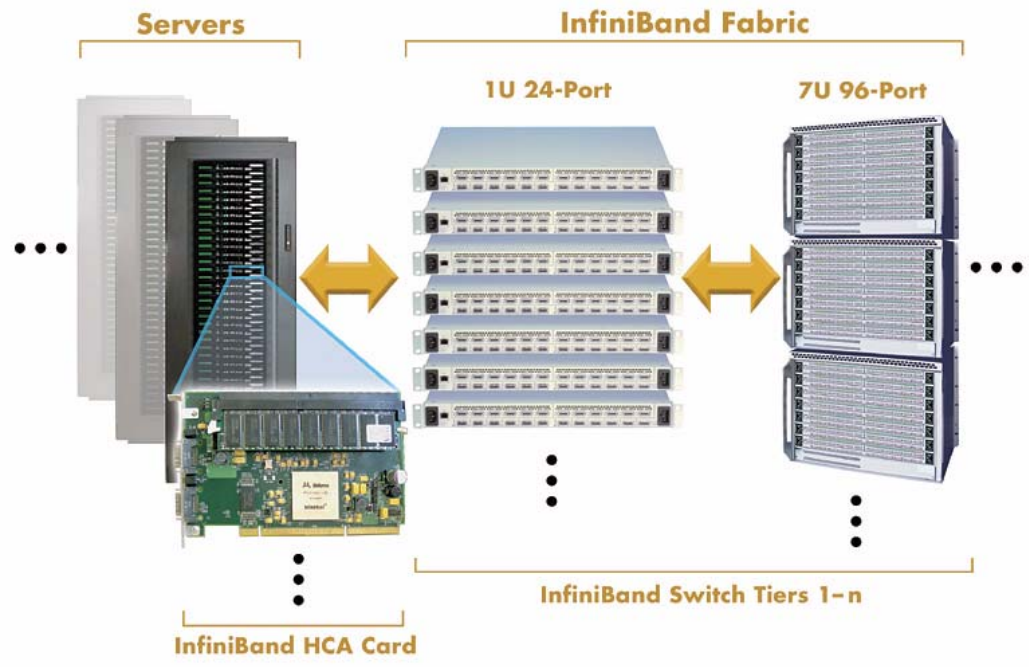

Figure 13. InfiniBand Cluster

10.0 Hardware Performance Results

Mellanox has achieved outstanding HCA and switch performance. Mellanox recently demonstrated greater than 2.5 GBytes/sec of aggregate data bandwidth in a Verbs level performance test. Best of all, this level of bandwidth can be achieved with processor utilization of less than 4%.

10.1 Application and MPI results over InfiniBand

HPC clustering has achieved a new level of performance. Mellanox and our partners and customers are experiencing impressive application level results. Figure 14, “SpecEnv Benchmark Performance Comparisons (www.spec.org),” shows the SpecEnv results from spec.org.It can be seen that InfiniBand delivers greater than 29% improvement in run times compared to competing interconnects.

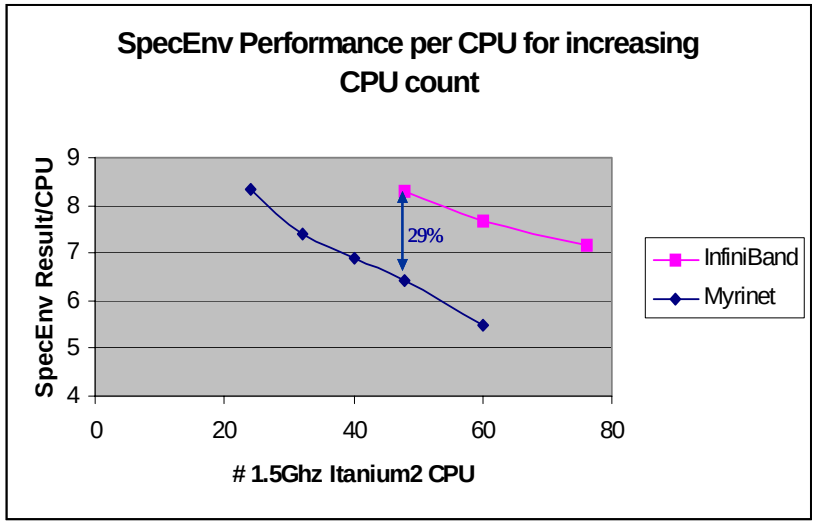

Figure 14. SpecEnv Benchmark Performance Comparisons (www.spec.org)

Furthermore, InfiniBand shows a trend of increasing advantage and efficiency at higher CPU counts. MPI bandwidth results shown in Figure 15, “Bandwidth vs. Message Size,” have topped 2.5 GB/sec.

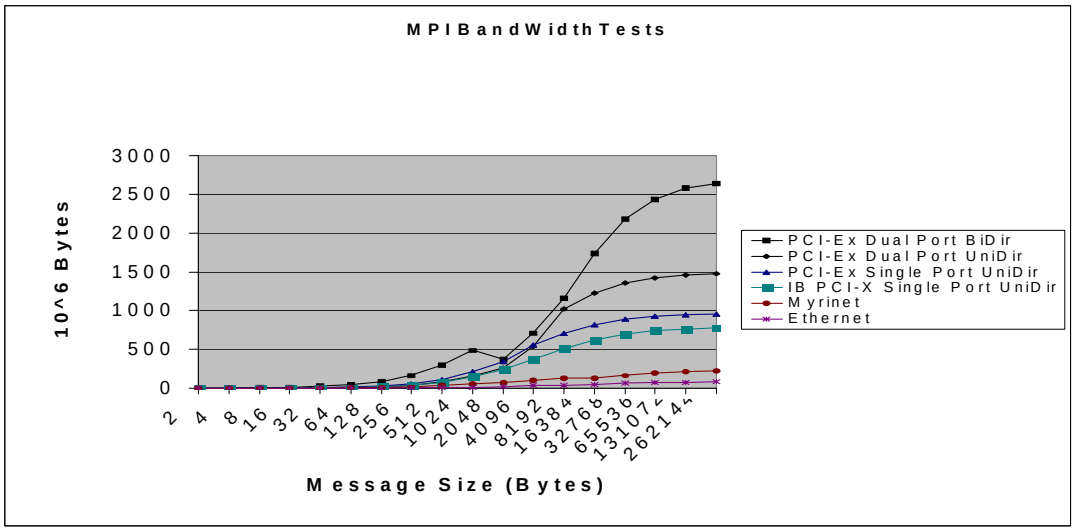

Figure 15. Bandwidth vs. Message Size

InfiniHostIIIex HCA MPI results are superb across the board. User latencies for small messages at 3.5 usec and MPI latency at 4.6usec. Even for smaller message sizes the bandwidth is clearly superior and as the message size increases beyond 16KB, InfiniHostIII Ex’s 10Gb/sec transfer really takes hold. These latencies and bandwidth will continue to be improved with firmware and driver optimizations and as future HCAs can exploit new high speed local I/O interfaces on new server chipsets.

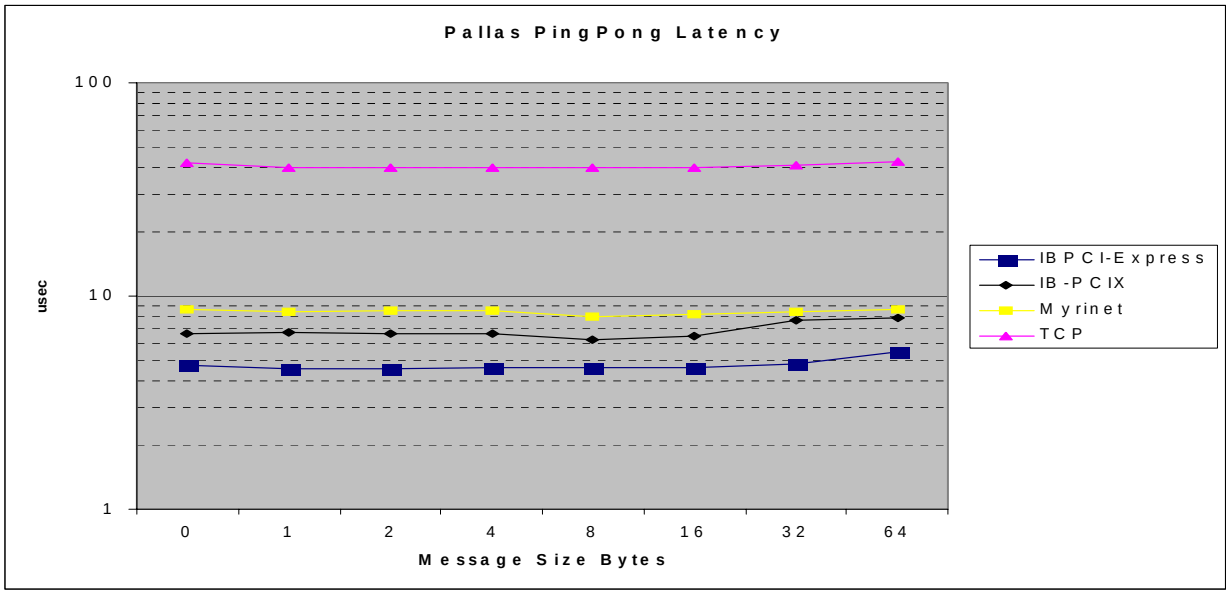

Figure 16. MPI Latency vs Message Size. Courtesy DK Panda, Ohio State University

10.2 Summary

Mellanox is the key InfiniBand technology provider to the data center and high performance computing markets. Mellanox provides silicon, board, systems and software technology to OEM and VAR partners, who offer fully integrated solutions to end customers. InfiniBand technology delivers unmatched bandwidth and latency on an industry standard, high performance interconnect for the high performance computing market. Only InfiniBand delivers the

big four: open standard, 10Gb/sec performance, transport offload, and remote DMA (RDMA) capabilities. This combination makes high performance computing easy and affordable to deploy with InfiniBand today.

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. 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