High-speed Remote Direct Memory Access (RDMA) Networking for HPC:

Comparative Review of 10GbE iWARP and InfiniBand

February, 2010

A Margalla Communications Special Report

About Margalla Communications

Margalla Communications provides the following strategic and technical marketing consulting services:

Custom Consulting
Margalla Communications houses vendor-independent storage and server networking domain experts who provide counsel on product planning and strategy, product positioning, market validation, target market selection, business development, custom market research, competitive positioning, demand creation, and other issues of concern to anyone in the networking industry. Margalla has consulted for vendors, end-users, and venture capitalists, and can be contracted on a retainer, daily or project basis. Contact info@margallacomm.com for further details.

Technology White Papers
Margalla Communications can provide technology and marketing white papers to suit a variety of enterprise needs. Contact info@margallacomm.com for further details.

About the Author

Saqib Jang is Founder and Principal at Margalla Communications, a Woodside, CA-based firm providing strategic and technical marketing consulting services to storage and server networking markets. He has more than ten years of experience as a marketing consultant and industry analyst for the networking market. Prior to independent consulting, Saqib held senior management and marketing roles with several public and private high-technology companies, including holding the positions of CEO/Co-founder of a startup developing provider-grade systems for IP videoconferencing services delivery, and Director of Marketing at NEC Systems (the system integration and software arm of NEC Corporation). Previously, Saqib was responsible for software product management and marketing at Auspex Systems, including Auspex's award-winning entry into the CIFS market, and spent over 7 years at Sun Microsystems/SunSoft in a range of executive marketing roles developing and implementing product marketing strategies for industry-leading networking and storage networking products.

Mr. Jang holds a B.S. degree in Electrical Engineering from Massachusetts Institute of Technology (MIT) and an M.B.A. in Marketing from the Wharton School, University of Pennsylvania. He can be reached at saqibj@margallacomm.com

Executive Summary

Cluster computing systems, separate compute nodes built from standard component technologies, have caused disruptive changes in the HPC market. While algorithm and application tuning is often required to obtain the performance benefits of HPC clusters, so often are cost, bandwidth, message rate, and latency of cluster interconnects.

The leading interconnects in HPC are Gigabit Ethernet (GbE) and InfiniBand (delivering upwards of 10X performance vs. GbE). The deployment of 10 Gigabit Ethernet (10GbE) cluster networking is emerging at this point. The price of this interconnect has been falling as the volume of its shipments grow. This growth is based on a combination of its 10X performance over GbE along with the ease of deployment due to its Ethernet heritage.

As cluster systems have grown, so has the total amount of data in play in the average parallel HPC application and HPC storage systems need to have the best possible bandwidth and latency characteristics. In this context, the demand for interconnect solutions that supports a converged storage and cluster interconnect fabric is expected to grow significantly.

The iWARP or RDMA over TCP/IP standard implements a number of mechanisms to provide a low-latency means of passing RDMA over Ethernet for applications such as cluster inter-process communications (IPC). 10GbE iWARP NICs (or R-NICs) provide hardware support for iWARP extensions.

InfiniBand is a point-to-point switched I/O fabric architecture designed to increase the communication speed between CPUs, devices within servers and subsystems located throughout a network. A single InfiniBand link supports 2.5 Gbps in each direction per connection. InfiniBand supports double (DDR) and quad data rate (QDR) speeds, for 5 Gbps or 10 Gbps respectively, at the same data-clock rate.

Because it is layered on top of TCP, iWARP is fully compatible with existing Ethernet switching equipment that is able to process iWARP traffic out-of-the-box. In comparison, deploying InfiniBand requires installing and managing two separate network infrastructures as well as specialized InfiniBand to Ethernet gateways for bridging between the two infrastructures.

10GbE infrastructure is available from a range of incumbent and startup vendors. Intel, Broadcom, and Chelsio provide 10GbW iWARP adapters, while 10GbE switches are available from a broad range of vendors including Cisco, HP, IBM, Extreme, Force10, Arista, and Voltaire. InfiniBand host channel adapter and switch silicon is only available from two vendors (Mellanox and QLogic), who in turn have signed up a number of OEMs to carry adapter and switching systems.

Both 10GbE iWARP and InfiniBand interconnects offer equivalent capabilities from an operating system support standpoint. The OpenFabrics software stack that is fully integrated into the flavors of Linux distributed by Novell and Red Hat fully supports both 10GbE iWARP and InfiniBand.

10GbE iWARP leverages its Ethernet heritage to also support acceleration of emerging Ethernet-based storage protocols, including file storage (NFS-RDMA). In addition, 10GbE iWARP adapters can also provide concurrent, native support for standard Ethernet storage protocols such as NFS, CIFS, and iSCSI. In comparison, InfiniBand has had minimal deployments for server-to-storage communications, whether for file or block storage.

Regarding pricing, major server vendors are starting to add 10 GbE controllers to the motherboard – known as LAN-on-Motherboard (LOM) – and NIC prices will continue to drop as LOM technology lets NIC vendors reach the high volumes they need to keep costs down, which in turn will drive switch port prices down as well. InfiniBand, on the other hand, has reached a mature market position and, consequently, reductions in the pricing of InfiniBand products will be relatively gradual.

Large-scale clusters built using 10GbE iWARP technology and high port-count 10GbE switches are gaining ground, and cluster scalability is no longer viewed as inhibiting 10Gbe deployment. InfiniBand technology has gained an established position for building large node-count clusters.

From a roadmap standpoint, the Ethernet market is moving forward aggressively to develop and implement 40G and 100G-based standards. The standard for these Ethernet versions is expected to be ratified during 2010 and initial implementations based on these standards will be shipping from a range of vendors in the blade server and Ethernet networking switch markets in the next 2-3 years. The roadmap initiatives in the InfiniBand space consist of QDR, EDR, and RDMA over CEE. However, these roadmap initiatives suffer from the same limitations that have been a traditional challenge for InfiniBand, namely, limited vendor support.

The emerging RDMA over Converged Enhanced Ethernet (RoCEE) protocol is designed to allow the deployment of RDMA semantics on Converged Enhanced Ethernet (CEE) fabric by running the IB transport protocol using Ethernet frames. CEE refers to a set of standards being developed by IEEE with a goal of enabling an Ethernet link to be split into multiple “virtual links” that operate independently and provide “lossless” communication.

RoCEE is unproven and its deployment faces significant hurdles including standardization and application and upper layer adoption. In addition, RoCEE is dependent on the deployment of 10GbE CEE infrastructure; currently only one vendor (Cisco) offers CEE switches, which are at relatively high price points.

Table of Contents

1.0 Introduction: The Rise of HPC Cluster Computing

While the HPC market is expected to experience a revenue dip in 2009, growth is expected to resume in 2010 and remain a bright spot in the overall IT market. The most important feature of the HPC growth trend is that it will continue to be fueled primarily by purchases of Linux cluster systems priced under $250,000. Entry-level clusters designed for the technical workgroup (systems selling for under $100,000) at smaller firms and in back-office locations are expected to show particularly strong growth in the coming years. Cluster computing systems, separate compute nodes built from standard component technologies (x86 processors, commodity motherboards, standards-based networking technology, and primarily the Linux OS), have caused disruptive changes in the HPC market.

As the component technologies of cluster systems have improved and buyers have become more confident running cluster systems, they have inevitably redirected capital once earmarked for large custom systems to larger cluster systems. These much larger clusters, often with thousands of processors, present opportunities for huge performance gains through improved parallel performance resulting in an overall higher order of magnitude return-on-investment (ROI). While algorithm and application tuning is often required to obtain these benefits, so often are cost, bandwidth, message rate, and latency of cluster interconnects.

One consequence of the range of requirements for cluster networking is that the leading interconnects in HPC are Gigabit Ethernet (which is based on Ethernet networking standard) and InfiniBand (delivering upwards of 10X performance vs. GbE). Both show significant deployment in HPC. The latest Top500 list of HPC systems has 259 Gigabit Ethernet-based deployments compared to 181 InfiniBand-connected systems[1]. The deployment of 10 Gigabit Ethernet (10GbE) cluster networking is emerging at this point. The price of this interconnect has been falling as the volume of its shipments grow. This growth is based on a combination of its 10X performance over GbE along with the ease of deployment due to its Ethernet heritage positions it for a bright future as a cluster interconnect.

As cluster systems have grown, so has the total amount of data in play in the average parallel HPC application. It is much easier to scale jobs with large or increased data sets (weak scaling) than those whose data sets are fixed even as the number of processors applied to the problem grows (strong scaling). This has significant implications for HPC storage systems. Storage systems need to have the best possible bandwidth and latency characteristics. HPC storage systems have themselves become increasingly clustered and parallel as well as network-attached and accessible from all nodes on the cluster through the interconnect.

[1] http://www.top500.org/stats/list/34/connfam

In this context, the demand for interconnect solutions that supports a converged storage and cluster interconnect fabric is expected to grow significantly.

2.0 10GbE iWARP Overview and Value Proposition

For years, Ethernet has been the *de facto* standard LAN for connecting users to each other and to network resources. Ethernet sales volumes make it unquestionably the most cost-effective data center fabric to deploy and maintain. The latest generation of Ethernet, 10 Gigabit Ethernet (10GbE), offers a 10 Gbps data rate, which simplifies growth for existing data networking applications while removing the bandwidth barriers to deployment for highest-performance HPC clustering and storage networking.

- 10 GbE end-to-end performance now compares very favorably with that of more specialized data center interconnects, which eliminates performance as a drawback to the adoption of an Ethernet unified data center fabric.
- Off-loading cluster and storage protocol processing from the central CPU to intelligent 10GbE NIC can also improve the power efficiency of end stations because off-load ASIC processors are generally considerably more power-efficient in executing protocol workloads.

Achieving 10GbE performance for latency-sensitive HPC communications has required solving Ethernet's long-standing overhead problems; problems that, in slower Ethernet generations, were adequately overcome by steadily increasing CPU clock speeds.

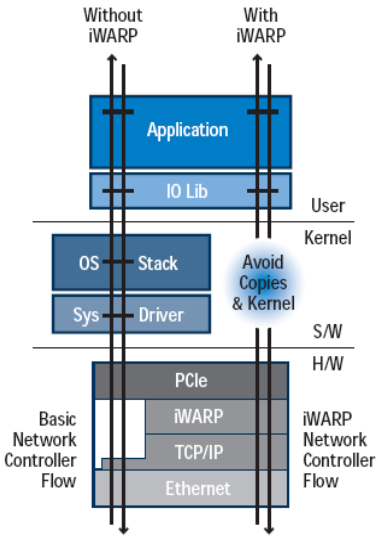
- Traditionally, the end systems' CPU has performed TCP/IP protocol processing in software. The load on the CPU increases linearly as a function of packets processed, with the usual rule of thumb being that each bit-per-second of bandwidth consumes about one Hz of CPU clock (e.g., 1 Gbps of network traffic consumes about 1 GHz of CPU). As more of the host CPU is consumed by the network load, both CPU utilization and host send/receive latency become significant issues.
- The value of implementing TCP/IP protocol processing in silicon at 10 Gbps data rates is clear. Effectively, such approaches have the potential of reducing the relative bandwidth and latency overhead effect of TCP/IP protocol processing to zero.

Enter 10GbE iWARP

The RDMA Consortium, which was founded to solve the long-standing overhead problems associated with Ethernet, released the iWARP extensions to TCP/IP in October 2002. Together, these extensions eliminate the three major sources of networking overhead – transport (TCP/IP) processing, intermediate buffer copies, and application context switches – that collectively account for nearly 100% of CPU overhead related to networking. Specifically, iWARP implements a number of mechanisms to provide a low-latency means of passing RDMA over Ethernet.

- Delivering a kernel-bypass solution. Placing data directly in user space avoids kernel-to-user context switches, reducing latency and processor load.
- Eliminating intermediate buffer copies. Data is placed directly in application buffers rather than being copied multiple times to driver and network stack buffers, reducing latency as well as memory and processor usage.
- Accelerated TCP/IP (Transport) Processing. TCP/IP processing is done in hardware instead of the operating system network stack software, enabling reliable connection processing that speed and scale.

The iWARP extensions utilize advanced techniques to reduce CPU overhead, memory bandwidth utilization, and latency by a combination of offloading TCP/IP processing from the CPU, eliminating unnecessary buffering, and dramatically reducing expensive OS calls and context switches – moving data management and network protocol processing to an accelerated RDMA over TCP/IP NIC (or R-NIC) 10 Gigabit Ethernet adapter.



(Source: Intel)

Figure 1. iWARP improves throughput and latency by reducing the overhead associated with kernel-to- user context switches, intermediate buffer copies, and TCP/IP processing.

An iWARP or RDMA over TCP/IP NIC (or R-NIC) provides hardware support for iWARP extensions. R-NICs allows a server to read/write data directly between its user memory space and the user memory space of another R-NIC-enabled host on the network, without any involvement of the host operating systems.

- R-NICs eliminate the delay and overhead associated with copy operations among multiple buffer locations, kernel transitions and application context switches.

- R-NICs can reduce CPU utilization for 10 Gbps transfers to less than 10 percent and can reduce the host component of end-to-end latency to as little as 5–10 microseconds. High port-count 10GbE switches are available, which delivers HPC-class latency performance within 100's of nanoseconds.

10GbE iWARP Market Adoption

iWARP has a number of inherent advantages due to its Ethernet legacy. These include compatibility with existing networking infrastructure, support by a range of established vendors, enabling high-speed clustering and storage fabric convergence, and integrated support within major operating systems.

Compatibility with Existing Data Center Infrastructure:

- iWARP was developed to perform within an Ethernet infrastructure, and thus does not require any modifications of existing Ethernet networks or equipment.
- iWARP's Ethernet compatibility enables IT organizations to take advantage of enhancements to Ethernet, such as Data Center Bridging, low-latency switches, and IP security.
- Standard Ethernet switches and routers carry iWARP traffic over existing TCP/IP protocols. Because iWARP is layered over TCP, network equipment does not need to process the iWARP layer, nor does it require any special-purpose functionality.
- An iWARP infrastructure is routable, which eliminates the need for gateways to connect to the LAN or WAN.

Vendor Support:

- A range of credible vendors such as Intel, Chelsio, and Broadcom, all provide stable, reliable 3rd generation 10GbE R-NIC products, including 10GbE iWARP server adapters and controllers. These products fully offload iWARP and TCP/IP protocol layers in ASICs, which enables high-throughput, low-latency operation for HPC cluster networking.
- 10GbE switches are offered by a range of providers including Cisco, HP, IBM, Extreme, Arista, Blade Networks, and Voltaire.

Operating Systems Support:

- The Open Fabrics Alliance (*www.openfabrics.org*) provides an open source RDMA software stack that is hardware-agnostic and application-agnostic for iWARP. These characteristics have allowed the easy integration of iWARP into existing environments while meeting stringent cost and performance requirements.
- The OpenFabrics software stack, known as the OpenFabrics Enterprise Distribution (OFED), is fully integrated into Linux distributions from Novell and Red Hat. The Linux kernel has included portions of OFED since mid-2005.

Converged Networking Support:

- For clustering, OFED enables out-of-the box iWARP acceleration of HPC cluster applications written for the Message Passing Interface (MPI) API.
- Various implementations of MPI middleware are written for OpenFabrics, including Open MPI, MVAPICH, MVAPICH2, HP MPI, and Intel MPI.
- iWARP also supports acceleration of Linux sockets applications, including legacy sockets applications, via Sockets Direct Protocol (SDP).
- iWARP also supports acceleration of popular network storage protocols, including file storage (NFS-RDMA) and Lustre networking (LNET). NFS-RDMA allows an NFS client and server to communicate using an RDMA-capable network transport, such as iWARP. Benefits of using iWARP for NFS communications include significantly lower CPU utilization and higher throughput. The Linux NFS-RDMA implementation runs on the OFED stack. This stack provides the software infrastructure and device driver support for RDMA capable devices.
- The Lustre shared file system is used for many different kinds of clusters. It is best known for powering over 40% of the 100 largest high-performance computing (HPC) clusters in the world, with tens of thousands of client systems, Petabytes (PB) of storage and hundreds of Gigabytes-per-second (GB/sec) of I/O throughput. It is currently available for Linux and provides a POSIX-compliant UNIX® file system interface. Lustre networking (LNET) provides support for major RDMA transports including iWARP. LNET uses the Linux OFED stack to provide support for iWARP.

Beyond support of emerging RDMA-based Ethernet storage protocols, 10GbE iWARP adapters can also provide concurrent, native support for standard Ethernet protocols such as NFS, CIFS, and iSCSI. The result is that 10GbE iWARP enabled servers can efficiently access network storage without the need for multiple, disparate adapters and/or inefficient server-based storage protocol processing.

With all of its inherent advantages, until recently, broad-based 10GE iWARP adoption within HPC environments was inhibited because of a few - but significant - problems involving 10GbE pricing and large cluster support. However, those problems are nearly overcome now, and 10GE iWARP is in the process of crossing the chasm into volume deployment in HPC.

NIC Pricing:

- Until recently, the only (including R-NICs) available for 10GE applications cost about $800 with most users preferring to use two of these per server. Standalone 10GbE NIC prices are now as low as $500.
- Instead of using a separate board, major server vendors are starting to add a 10 Gigabit Ethernet chip to the motherboard known as a LAN-on-Motherboard (LOM).

- This advance will drop the cost to well under $100 and removes the NIC price obstacle from 10GE, while NIC prices will continue to drop as LOM technology lets NIC vendors reach the high volumes they need to keep costs down.

Switch Port Pricing:

- Like R-NICs, initial 10GE switch prices inhibited early adoption of the technology. The original 10GE switches cost as much as $20,000 per port, which was more than the price of a server.
- Now list prices for 10GE switches available from a range of major and startup vendors (including Cisco, IBM, Extreme, HP, Blade Networks, Arista, and Voltaire) are on an average lower than $500 per port, and street prices are even lower.
- Equivalent pricing is available for embedded blade switches as well as for top-of-rack products available from a range of vendors. As discussed earlier, 10GbE is an initial point in its lifecycle and, thus, 10GbE prices will decline significantly as the Ethernet to 10Gbe transition ramps up.

Large Cluster Support:

- Another market inhibitor for 10GbE for large clusters was how to connect switches together to create large non-blocking clusters. Most clusters are small enough that this is not an issue. For larger clusters, the combination of high port-count switches and CLOS (or fat tree) technology for scaling Ethernet switches provides a solution, and is starting to become established in the market. 48-256 port 10GbE switches are available from a range of vendors including Cisco, HP. IBM, Extreme, Voltaire and Arista that support the highest density clusters.

As an example, Purdue University[2] recently built a 1200 node HPC cluster using 10GbE infrastructure Ethernet high-performance compute cluster. Specifically, Purdue's "Coates" cluster uses Cisco Nexus 7000 and 5000 series 10Gbps Ethernet Switches and Chelsio S310E-CR 10Gbps Unified Wire adapters with RDMA offload using iWARP for delivering message passing functionality and network performance for this cluster. Each node in the cluster was a dual-processor HP utilizing 2.5 GHz quad-Core AMD 2380 "Shanghai" processor running Red Hat Enterprise Linux 5 (RHEL5).

Ethernet Roadmap: 40G and 100G Ethernet on the Horizon

In July 2007, the 802.3ba study group was named, and it is the first standard to include 2 different Ethernet speeds – the 40 Gbps speed for server applications and 100 Gbps for the Internet backbone – to server both market needs. In December 2007, the official 802.3ba task force was formed to begin work on the new standard that is expected to be ratified in 2010.

[2] http://chelsio.com/pr_072109.html

40G Ethernet and 100G Ethernet will play an important role in the on-going evolution of Ethernet. While speeds higher than 10G can be achieved using link aggregation, aggregating links is complex to configure, and can easily get out of balance and carry less than the promised bandwidth. Thus, 10GbE link aggregation is considered a stop-gap measure and not ideal as a long-term solution to the high-performance needs for tomorrow's networking requirements.

- Used in a blade server, 40G Ethernet can be run without optics, providing economical yet high-speed connections for HPC applications.
- Even today, 4 lanes at 10Gbps each (utilizing the 10GBASE-KR standard) is being integrated into the mid-plane of emerging blade server models, so 40G Ethernet should be a strong fit for blade servers.
- The availability of 40G or 100G inter-switch links will also benefit servers. When used as an aggregation uplink for many servers – each connecting into the switch fabric with 10GbE interfaces – will prevent traffic from the servers from being caught in a 10G bottleneck.

3.0 InfiniBand Overview and Value Proposition

InfiniBand is an I/O architecture designed to increase the communication speed between CPUs, devices within servers and subsystems located throughout a network. Unlike the PCI-based I/O architecture, InfiniBand extends its feature set outside the server to devices on the network. The original goal behind the release of the InfiniBand specification by the InfiniBand Trade Association (www.ibta.org) in 2000 was to address the mismatch between the speed of CPUs and the PCI I/O bus, as well as other deficiencies of the PCI bus, including bus sharing, scalability, and fault tolerance.

- InfiniBand is a point-to-point, switched I/O fabric architecture. Both devices at each end of a link have full access to the communication path. To go beyond a point and traverse the network, switches come into play. By adding switches, multiple points can be interconnected to create a fabric. As more switches are added to a network, aggregated bandwidth of the fabric increases. By adding multiple paths between devices, switches also provide a greater level of redundancy.
- A single InfiniBand link supports 2.5 Gbps in each direction per connection. InfiniBand supports double (DDR) and quad data rate (QDR) speeds, for 5 Gbps or 10 Gbps respectively, at the same data-clock rate. InfiniBand links use 8B/10B encoding — every 10 bits sent carry 8bits of data — making the useful data transmission rate four-fifths the raw rate. Thus single, double, and quad data rates carry 2, 4, or 8 Gbps respectively.
- A quad-rate 12X link therefore carries 120 Gbps raw, or 96 Gbps of useful data. At present, most systems use 4X 10 Gbps (SDR), 20 Gbps (DDR) or 40 Gbps (QDR) connections. However, InfiniBand QDR performance is bounded by the 26Gbps PCIe Gen2 throughput limitation.

- Latency performance of InfiniBand SDR ad DDR switch chips is around 200 nanoseconds. InfiniBand Host Channel Adapters (HCAs) are rated 1-3 microseconds (though effective application-level performance is a different matter and is discussed later).

High-end clustering architectures have provided the main opportunity for InfiniBand deployment. Using the InfiniBand fabric versus Gigabit Ethernet as the cluster inter-process communications (IPC) interconnect typically boosts cluster performance and scalability while improving application response times. InfiniBand also provides exceptional scalability and failover in comparison to Gigabit Ethernet. In short, compared to Gigabit Ethernet, InfiniBand stands out in providing the mechanisms necessary to support the demanding requirements of high-end clustering.

InfiniBand Market Adoption

While InfiniBand has gained significant traction in the high-end HPC space primarily at the expense of Gigabit Ethernet, there are a number of challenges standing in the way of its expanding beyond that market.

Compatibility with Existing Data Infrastructure:

- High-performance computing clusters that use an InfiniBand interconnect for server clustering also use Ethernet. Ethernet networking is the standard for user and storage connectivity, and for the management network that orchestrates the cluster.
- Thus, deploying InfiniBand requires environments where two separate network infrastructures are installed and managed as well as specialized InfiniBand to Ethernet gateways for bridging between the two infrastructures.

Vendor Support: The vendors offering InfiniBand server and switch silicon, as well as adapters, are Mellanox and Qlogic, who have signed-up a number of HPC-focused server and switch OEMs to offer InfiniBand-based infrastructure.

Operating Systems Support:

- InfiniBand is also fully supported by the OpenFabrics software stack (OFED). OFED is supported by Mellanox and InfiniBand systems vendors to enable OEMs and System Integrators to meet the needs of HPC cluster applications.
- For traditional TCP/IP and sockets-based applications, OFED includes IP-over-IB enabling IP-based applications to work over InfiniBand. OFED also includes the Sockets Direct Protocol (SDP) enabling traditional TCP/IP sockets-based applications to leverage RDMA transports, including InfiniBand.

Converged Networking Support: The main application of InfiniBand infrastructure is supporting low-latency inter-process communications for HPC clusters.

InfiniBand has had minimal deployments for server to storage communications, whether for file or block storage.

NIC and Switch Port Pricing: InfiniBand DDR switch port street pricing is in the range of $300-400, whereas per-port pricing for DDR server adapters is in the range of $300-400.

Large Cluster Support:

- InfiniBand has a major problem which has become obvious as it has been more broadly deployed within HPCC data centers: InfiniBand switches cannot adjust to congestion as 10GbE iWARP devices can. As a result, switch buffers can fill up, block upstream switches and even block flows that are not contending for the congested link.
- For small-scale application environments with a predictable load level, this is not a problem. However, for large-scale deployments spanning hundreds to thousands of servers supporting a range of applications, this lack of congestion control has become a major challenge.

InfiniBand Roadmap:

- InfiniBand vendors are starting to ship QDR InfiniBand nominally rated at 40 Gbps. Because, as discussed earlier, InfiniBand uses 8b/10b encoding, 40 Gbps InfiniBand is effectively 32 Gbps.
- Infiniband EDR with a nominal per-port rating of 80 Gbps will start to ship in 2011.
- The real limitation for InfiniBand server networking is the PCIe bus inside the server, which is typically capable of only 13 Gbps except newer servers use PCIe "Gen 2" to get to 26 Gbps. So, net-net, the effectiveness of InfiniBand QDR is limited by PCIe Gen2 throughput limitation.

4.0 RoCEE Overview and Value Proposition

Mellanox, the leader in the InfiniBand market, is behind the emerging RDMA over Converged Enhanced Ethernet (RoCEE) protocol proposal. RoCEE is designed to allow the deployment of RDMA semantics on Converged Enhanced Ethernet fabric by running the IB transport protocol using Ethernet frames.

The IEEE has been developing standards collectively referred to as "Data Center Bridging" (DCB) or "Converged Enhanced Ethernet" (CEE) This refers to high speed Ethernet (currently 10 Gb/sec, with a clear path to 40 Gb/sec and 100 Gb/sec), plus a number of new features. The main new features are:

- *Priority-Based Flow Control (802.1Qbb), sometimes called "per-priority pause"*
- *Enhanced Transmission Selection (802.1Qaz)*
- *Congestion Notification (802.1Qau)*

The first two features allow splitting an Ethernet link into multiple “virtual links” that operate independently — bandwidth can be reserved for a given virtual link, and by having per-virtual-link flow control, CEE can ensure that certain traffic classes do not overrun their buffers thus avoiding dropping packets. This congestion notification capability means that we can tell senders to slow down to avoid congestion spreading caused by that flow control.

CEE was developed primarily for use in Fibre Channel over Ethernet (FCoE). FC requires a very reliable network — it simply does not work if packets are dropped because of congestion — and, so, CEE provides the ability to segregate FCoE traffic on top of a “no drop” virtual link.

Mellanox’s RoCEE proposal was motivated in order to create a protocol analogous to FCoE for Ethernet-based cluster networking. In other words, to take the InfiniBand transport layer and package it into Ethernet frames, instead of using the iWARP protocol for Ethernet-based high-performance cluster networking. But there are a number of challenges associated with this proposal:

- First, one of the major motivations behind the RoCEE proposal is that it is the fastest path forward for an Ethernet-based alternative to InfiniBand. However, this ignores the fact that iWARP adapters are already shipping from multiple vendors, including Intel, Chelsio, and Broadcom. In addition, iWARP will automatically leverage the performance benefits of CEE as support for it will be ubiquitous in all 10 GbE server adapter and LOM implementations, iWARP and non-iWARP alike.
- Second, the idea that an InfiniBand over Ethernet (IBoE) specification will be quick or easy to develop flies in the face of the experience with FCoE; while FCoE sounded simple in concept, it turns out that the standards work took at least three years. In comparison, IBoE is more complicated to specify, and fewer resources are available for it, so a realistic view is that a true standard is very far away.
- Last, RoCEE proponents point to the performance overhead challenges related to iWARP based on the TCP/IP protocol. However, this does not take into account the efficiency of silicon-based implementations of 10 Gbps TCP/IP. In addition, iWARP is also positioned to automatically take advantage of CEE as that protocol gains ubiquity in 10GbE server LOM and adapters.

In summary, RoCEE is unproven and its deployment faces significant hurdles including standardization and application and upper layer adoption. In addition, RoCEE is dependent on the deployment of 10GbE CEE infrastructure; currently only one vendor (Cisco) offers CEE switches, which are at relatively high price points.

5.0 10GbE and InfiniBand Performance: An Application Perspective

Large-scale HPC implementations require an interconnect that provides sufficient application performance. While there is a range of micro-benchmark data available from vendors, what is most meaningful from the end-user standpoint is comparative performance data for the two interconnects for volume applications, specifically, the advantage in terms of applications execution time that InfiniBand or iWARP may offer.

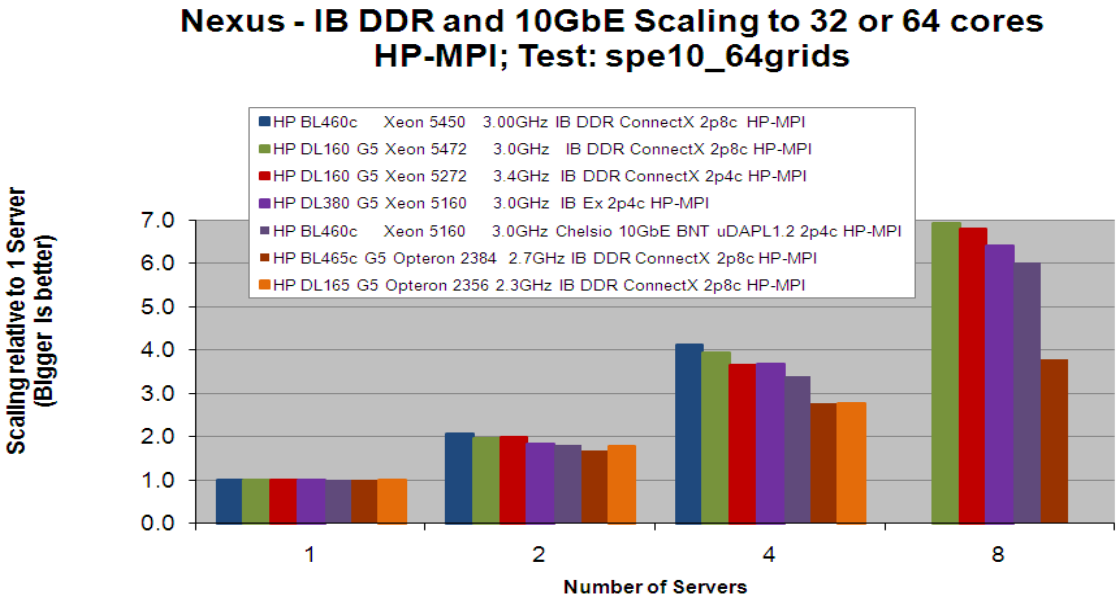
Examining the question of application performance reveals that some HPC applications that are loosely coupled (or don't demand excessive low latency) can run perfectly well over either interconnect. In fact, some TCP/IP applications actually run faster and with lower latency over 10GE iWARP than over InfiniBand.

For more performance-hungry and latency-sensitive applications, the performance potential of 10GE iWARP for HPC is comparable to current developments in InfiniBand technology, especially as compared to Gigabit Ethernet.

An HP benchmarking study[3] conducted with Landmark Nexus® reservoir simulation software to determine its performance characteristics on a variety of HP ProLiant® x86_64 server cluster environments found that both 10GbE iWARP and IB DDR showed significant improvements in Nexus parallel performance when compared to GbE. In addition, the 10 Gigabit Ethernet performance is comparable, though slightly slower, than InfiniBand performance.

Because the two high-speed interconnects perform in a similar manner vis-a-vis performance for Nexus software, HP suggests that non-performance characteristics may be a deciding factor for a customer desiring the most appropriate technology for his site, including the choice of the high-speed interconnect in other parts of the customer network and current and future price estimates for the interconnect.

[3] http://www.bladenetwork.net/userfiles/file/PDFs/09-LM61-04NexusConfigurationWhitePaper.pdf



6.0 Comparative Review Summary

This section provides a comparative review of 10GbE and InfiniBand technologies using the market adoption criteria discussed earlier, as well as a summary comparison of 10Gbe iWARP and the emerging RDMA over CEE (RoCEE) protocol.

iWARP and InfiniBand

- As far as its compatibility with existing data center infrastructure, because it is layered on top of TCP, iWARP is fully compatible with existing Ethernet switching equipment that is able to process iWARP traffic out-of-the-box.
- In comparison, deploying InfiniBand requires environments where two separate network infrastructures are installed and managed as well as specialized InfiniBand to Ethernet gateways for bridging between the two infrastructures.
- 10GbE infrastructure is available from a range of incumbent and startup vendors. Intel, Broadcom, and Chelsio provide 10GbW iWARP adapters, while 10GbE switches are available from a broad range of vendors including Cisco, HP, IBM, Extreme, Force10, Arista, and Voltaire.
- InfiniBand host channel adapter and switch silicon is only available from two vendors (Mellanox and QLogic), who in turn have signed up a number of OEMs to carry adapter and switching systems.
- Both interconnects offer equivalent capabilities for supporting operating systems (OSs). The OpenFabrics software stack that is fully integrated

into the flavors of Linux distributed by Novell and Red Hat fully supports both 10Gbe iWARP and InfiniBand.

- 10GbE iWARP leverages its heritage to also support acceleration of emerging Ethernet-based storage protocols, including file storage (NFS-RDMA), which is fully supported by the Linux OFED stack. In addition, the Linux OFED stack also enables 10GbE iWARP to out-of-the-box support Lustre networking (LNET). In addition, 10GbE iWARP adapters can also provide concurrent, native support for standard Ethernet protocols such as NFS, CIFS, and iSCSI.
- In comparison, InfiniBand has had minimal deployments for server-to-storage communications, whether for file or block storage.
- Regarding pricing, major server vendors are starting to add a 10 Gigabit Ethernet chip to the motherboard-known as LAN-on-Motherboard (LOM). NIC prices will continue to drop as LOM technology lets NIC vendors reach the high volumes they need to keep costs down, which in turn will drive switch port prices down as well.
- InfiniBand, on the other hand, has reached a mature market position and, consequently, reductions in the pricing of InfiniBand products will be relatively gradual.
- Large-scale clusters built using 10Gbe iWARP technology and high port-count 10Gbe switches are gaining ground, and cluster scalability is no longer viewed as inhibiting 10Gbe deployment. InfiniBand technology is an established interconnect for building large node-count clusters.
- From a roadmap standpoint, the Ethernet market is moving forward aggressively to develop and implement 40G and 100G-based standards. It is expected that the standard for these versions of Ethernet will be ratified during 2010 and initial implementations based on these standards will be shipping from a range of vendors in the blade server and Ethernet networking switch markets within the next 2 to 3 years.
- The roadmap initiatives in the InfiniBand space consist of QDR and RDMA over CEE. However, these roadmap initiatives suffer from the same limitations that have been a traditional challenge for InfiniBand, namely, limited vendor support.

Attribute	10GbE iWARP	Infiniband	RDMA over CEE
Overview	RDMA over TCP/IP Ethernet (can also use CEE))	Point-to-point switched fabric for RDMA	RDMA over CEE
Maturity	Established - production deployment in large clusters	Established - 2nd position as interconnect in Top500 list	Pre-standard
Requirements	Runs on standard Ethernet infrastructure	Requires IB to Ethernet gateways	Requires upgrade to CEE infrastructure
Standardized	IETF standard	IBTA standard	No
Routes (i.e. does not require gateways)	Yes	No	No
Generation	Intel and Chelsio shipping 3rd generation silicon	Mature	First generation
Converged Networking Support	10GbE R-NICs support iSCSI, FCoE, as well as integrated TCP/IP processing for NAS	Limited IB adoption in storage	Mellanox bundles 10GbE NIC and open-FCoE
Cost	10GbE switch port pricing is in midst of major declines	Low switch port pricing	CEE switch port pricing is relatively high

Table 1: Comparative Review of iWARP, InfiniBand, and RoCEE Networking