

June 2005

Contents

InfiniBand Performance and Price Advantages	3
Higher Performance.....	4
Performance Benchmarks	5
How Well Do Benchmarks Predict Performance?	9
InfiniBand Scalability	9
Flow Control and Virtual Lanes.....	9
Trunk Links Simplify Multi-Tier Networks	10
Flexible Routing Schemes	10
InfiniBand Increases Cluster Reliability.....	11
Isolation and Partitioning.....	12
Supporting Existing Codes	12
Message Passing Interface (MPI)	12
IP Network Emulation and Socket API	12
High Performance Storage over InfiniBand	13
Consolidating Server I/O.....	13
Voltaire InfiniBand Software Stack.....	14
Voltaire HPC Fabric Solutions	15
Voltaire HCA400/EX Adapter...	15
Voltaire ISR 9288.....	15
Voltaire ISR 9096.....	16
Voltaire ISR 9024.....	16
Voltaire ISR 6000 and router modules	16
Comprehensive Management	16
Conclusion	17

Closing the Gap in High Performance Clusters and Grids

Voltaire InfiniBand Interconnect Solutions for High Performance Computing

Abstract: Clusters are an increasingly attractive alternative to large monolithic commercial computers in many high performance computing applications. Their popularity has been driven by advancements in the parallelization of applications and the rapid price/performance gains in industry standard computer platforms. Major computer manufacturers see InfiniBand technology as the backbone for integrated Grid and Blade Server computing.

Memory speed and I/O bandwidth have not kept pace with CPU advances and these bottlenecks have limited cluster scalability. Now, Remote Direct Memory Access (RDMA) and InfiniBand (IB), the latest I/O interconnect architecture, provide standards-based technologies and are poised to deliver a new leap in cluster performance. This combination of standards and performance is driving the creation of new Grid computing solutions, Clusters, and Blade Computing. This technology allows for virtual servers to be “custom tuned” to applications providing the highest level of versatility, price, and performance.

The InfiniBand standard enables much greater scalability at lower costs and higher performance than preceding cluster interconnects. It provides a high speed, low latency, point-to-point or switched fabric solution optimized for server clusters. InfiniBand is versatile. It can be deployed in high performance computing clusters for scientific and engineering applications as well as for commercial customers who want highly available and scalable application clusters.

The benefits of clusters using an InfiniBand computing infrastructure include:

- Lower cost of computing by optimizing computing assets and resources within a single, high bandwidth, fabric architecture.
- Improved efficiency and utilization by matching jobs to virtual resources.
- Integrated direct attached storage and network I/O into a unified easily managed system.

InfiniBand opens the door for new applications to take advantage of the power of clusters. Its scalability makes more compute power available for a significantly lower price. It brings low cost compute power to organizations that can benefit from parallel computational techniques. The ability to support NAS and SAN storage subsystems directly gives InfiniBand clusters better access to application data. Cluster file systems such as Terrascale, PVFS and Lustre enable InfiniBand clusters to compete for traditional mainframe applications, such as real-time signal processing, computer aided engineering, and financial modeling.

Your feedback is valuable. Voltaire welcomes any questions or feedback. Please send your comments to: Info@Voltaire.com

Notice

Reproduction of this publication in any form without prior written permission is not allowed. The information in this publication is subject to change without notice and is provided "AS IS" WITHOUT WARRANTY OF ANY KIND. THE ENTIRE RISK ARISING OUT OF THE USE OR INTERPRETATIONS OF THIS INFORMATION REMAINS WITH RECIPIENT. IN NO EVENT SHALL VOLTAIRE BE LIABLE FOR ANY DIRECT, SPECIAL, PUNITIVE OR OTHER DAMAGES.

Performance results will vary based upon a number of system factors. Some of these include: server configuration of the processor, chip set, memory size, firmware and driver release versions, MPI version and OS kernel version. The configuration or configurations tested or described may or may not be the only available solution. These tests are not a determination of product quality or correctness, nor does it ensure compliance with any federal state or local requirements.

Product names mentioned herein may be trademarks and/or registered trademarks of their respective companies.

InfiniBand Performance and Price Advantages

InfiniBand is an open industry standard developed by the world's leading computer manufacturers and supported by over 100 member companies in the InfiniBand Trade Association (www.infinibandta.org). This wide industry support benefits customers by offering a strong standard and enables an array of products from which to choose, and at competitive prices.

InfiniBand's scalability makes more compute power available for a significantly lower price

The InfiniBand architecture defines a communication network for interconnecting processing systems, storage I/O (SAN and NAS), and Network I/O (IP) into a single easy to manage environment. In an InfiniBand network, processing nodes connect to the communication fabric via host channel adapters (HCAs). I/O such as Direct Attached storage, NAS storage systems and network (IP) nodes can attach directly to InfiniBand switches transparently using their native connectivity environment, as opposed to using compute nodes dedicated to these functions. This allows InfiniBand clusters to easily be deployed in existing data centers with little hardware impact.

InfiniBand Cluster Applications benefit from:

- A high performance switched 10Gbps interconnect with optional support for 30Gbps.
- Bidirectional bandwidth that delivers over 2,700MB/second depending on the architecture of the system bus.
- Less than 3.0 microsecond latency.
- Remote Direct Memory Access (RDMA) with CPU and OS bypass technologies that greatly reduce memory copy overheads and associated CPU utilization.
- Internet Protocol over InfiniBand (IPoIB) and Socket Direct (SDP) support, which allows IP and/or TCP applications to operate seamlessly over an InfiniBand network.
- Native File and Block protocols enabling much faster storage solutions.
- DAPL (Direct Access Provider Library) support, which provides a standard API to reduce porting efforts and costs.

InfiniBand standardization promotes economies of scale as more vendors provide products and services. Similar to Ethernet, the manufacturers of InfiniBand systems use standard chips, connectors and cables, increasing volumes and decreasing prices. Major systems vendors have made significant investments. Silicon Graphics, IBM, Hewlett Packard, Apple, NEC, Dell, Sun Microsystems, Hitachi, Fujitsu, and recently Cisco all offer InfiniBand for cluster customers.

Current generation InfiniBand products support copper cabling, as well as fiber optics. At their first introduction, InfiniBand products are price competitive with proprietary system area networks that have reached the bottom of their cost curve. As InfiniBand products and enhanced capabilities such as storage connectivity continue to mature and volumes grow, an InfiniBand cost advantage will become consistent and more significant on both pricing and performance.

InfiniBand leverages years of research and development accumulated in the production of "channel-based" I/O systems for mainframe computers. This lineage ensures that InfiniBand scales reliably in high performance computing applications.

Higher Performance in Every Dimension

InfiniBand features RDMA, which places data directly into application buffers and reduces latency

InfiniBand is a highly scalable, switched I/O technology. It features bandwidth up to 60 Gbps and supports up to 48,000 nodes on a single sub network. In addition, it incorporates message passing as well as memory mapping via RDMA. It offloads most of the I/O communication from the host CPU with very low CPU utilization, making it an excellent solution for very high performance environments.

In today's high performance computing applications, latency is generally the most critical performance characteristic, followed closely by throughput and CPU utilization. InfiniBand achieves superior performance in all three dimensions. Early clusters were deployed on low-cost interconnect solutions such as Ethernet. As the number of CPUs and compute nodes increase however, bottlenecks occur and high CPU overhead diminishes the overall performance. InfiniBand standards provides for low CPU utilization freeing CPU cycles for application use.

InfiniBand takes advantage of RDMA, which enables a remote node to place data directly into the buffer of an application program without any involvement from the operating system. This powerful hardware mechanism enables InfiniBand to achieve the lowest latency and CPU utilization specifications available for a standard commercial cluster interconnect. The more CPU cycles available to the application, the faster the job gets done or more complex simulations can be accomplished in the same time.

InfiniBand end-to-end latency between software processes on two nodes is lower than 4.0 microseconds for a short message. Latency will be further reduced with the introduction of new Hardware this year to 1-2 microseconds. The below table is designed to show a general comparison as measured by the National Center for Supercomputing Applications between the various interconnects in regards to performance latency using the Pallas MPI Benchmark. InfiniBand continues to offer lower latency compared to other interconnect solutions. It shows even more dramatic improvements as message size increases.

Latency (uSec) vs. Message size (Bytes)*

	<i>1B</i>	<i>512B</i>	<i>1KB</i>	<i>4KB</i>	<i>8KB</i>	<i>32KB</i>	<i>524KB</i>
<i>InfiniBand</i>	5.42	8.3	10.36	16.21	24.17	60.23	628
<i>Myrinet</i>	8.29	11.56	15.31	32.51	53.15	166.3	2155
<i>GbE</i>	58	72.96	86.89	165.2	281.27	774.78	4631

Source: National Center for Supercomputing Applications (NCSA)
<http://vmi.ncsa.uiuc.edu/>

The InfiniBand 4x data rate is 8 Gbps, or four times faster than Myrinet

The InfiniBand data rate, using the 4X links that are available today, is 8 Gb/s (10 Gb/s signaling rate) or 16 Gb/s with double data rates (DDR). This data rate is 8-16 times faster than Gigabit Ethernet and 4-8 times faster than Myricom's Myrinet. Moreover, InfiniBand 12X can be used for trunking today and will be also supported on adapters such as the IBM Galaxy adapter, support for double and quad data rates will quadruple the base bandwidths to unlimited speeds, multiplying this performance advantage and achieving up to 96Gbps per single link.

InfiniBand host channel adapter cards using RDMA, can exploit the full bandwidth available on the InfiniBand link. Benchmarks conducted by Ohio State University show InfiniBand HCAs can nearly saturate an 8 Gbps link on both ends and deliver 1841 MB/s, while Myrinet adapters have reached the theoretical limit of their 2 Gbps

links and deliver only 450 MB/s. Most HCA adaptors also contain dual-ports (4X+4X), which if used improves the bandwidth. Dual-ports eliminate any bottlenecks between multiple CPU's and in cluster benchmarks show bandwidth as high as 2724 MB/s on a standard server when using both ports.

Lately, newer technology has been implemented with double data rates (DDR) which provide a 16 Gbps per link (or 48 Gbps per a 12X link), this technology drives more than 2500 MB/s on a single InfiniBand 4X link.

Industry Standard Performance Benchmarks

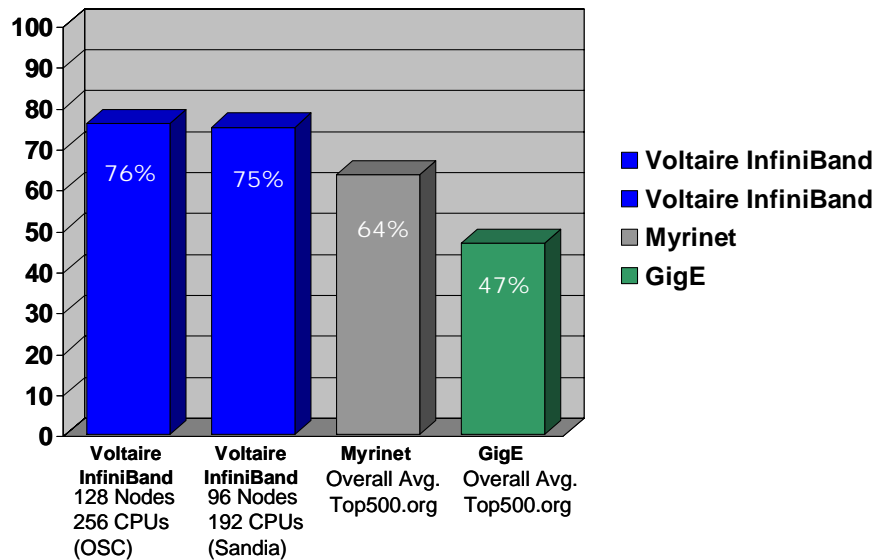
Application level performance benchmarks reinforce this advantage. Among those industry standard benchmarks relevant to a wide variety of industries and disciplines, InfiniBand leads by a substantial margin in many of the primary benchmarks.

LINPACK

The Linpack benchmark, which measures floating point performance and memory bandwidth, solves a dense system of linear equations. It allows the user to scale the size of the problem and to optimize the software in order to achieve the best performance for a given machine.

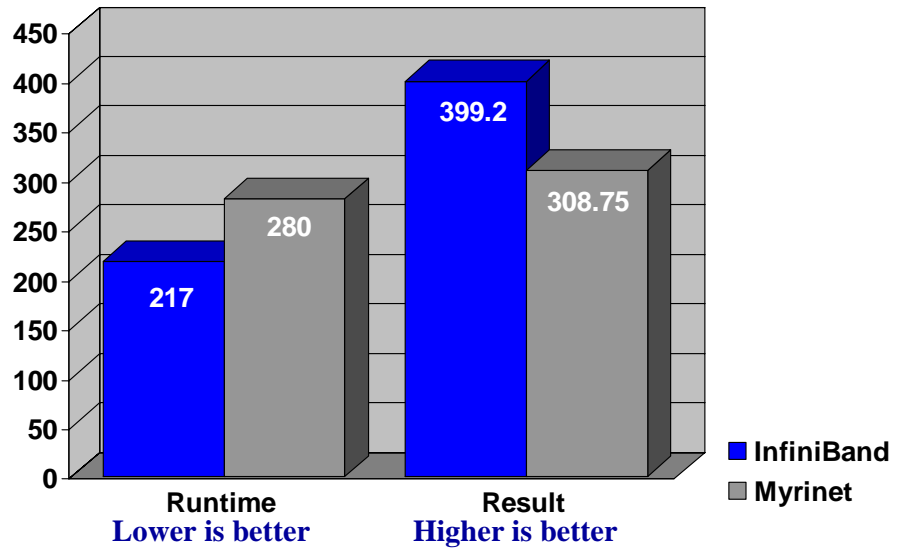
Linpack performance does not reflect the overall performance of a given system, as no single number ever can. It does, however, reflect the performance of a dedicated system for solving a dense system of linear equations. Since the problem is very regular, the performance achieved is quite high, and the performance numbers give a good indication of peak performance.

Actual deployed High-Performance clusters from Voltaire show extremely high Linpack (HPL) efficiency measurements.



As the above chart shows, Voltaire provides a high level of cluster efficiency in actual deployed solutions when compared to the average of the supercomputers listed on the TOP500.ORG running either Myrinet or GigE. This is true even in clusters configured for less than full cross-sectional bandwidth such as the 128 node cluster at Ohio Supercomputer Center.

On Identical Platforms the SPECenv results reveal InfiniBand performance over Myrinet



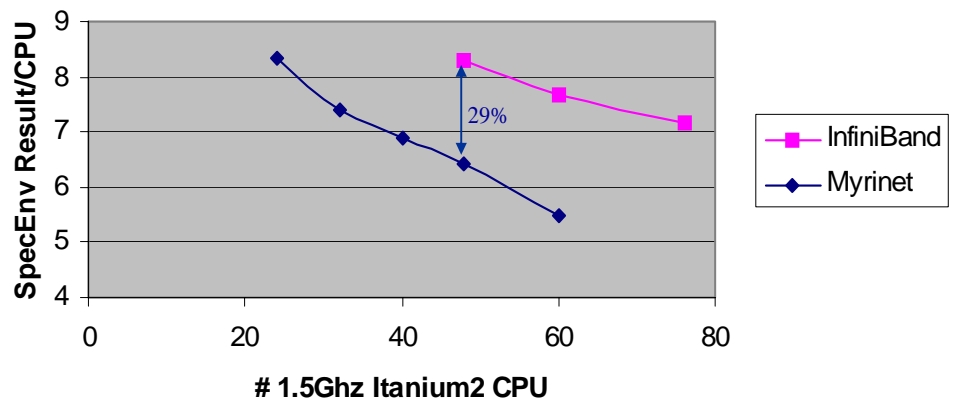
SPECENV2002

The SPEC ENV2002 benchmark is based on the Weather Research and Forecasting (WRF) model, a state-of-the-art, non-hydrostatic mesoscale weather model. SPEC/HPG integrated version 1.2.1 of the WRF model into the SPEC tools for building, running and verifying results.

The SPECenvM2002 chart above shows the performance of a computing system in simulating the weather over the continental United States for a 24 hour period at a 22km resolution using the WRF Model.

This performance benchmark of real MPI-based HPC applications runs on identical platforms and shows performance with both Myrinet and InfiniBand interconnects. It provides precise apples to apples interconnect comparison and demonstrates the performance benefits of InfiniBand.

The chart below from the same benchmark shows the scalability advantage with a 29% improvement for 48 CPUs; the gap increases even more to 40% for 60 CPUs. As the cluster grows and more CPUs are added to the system, the gap continues to widen.



The SpecEnv performance per CPU reveals increasing advantage and efficiency as the cluster CPU count increases.

Technical Computing

This broad sector includes scientific and applied research, finance, design, and manufacturing, covering a wide range of tasks from CAD modeling, computation, and programming to fluid dynamics, image recognition, and graphics.

Benchmarks abound here, and as InfiniBand systems are deployed in commercial solutions, the list of published benchmarks is growing. In many cases, InfiniBand systems posted improved results in applications demanding high bandwidth communications, beginning with the SPEC and other industry benchmarks noted earlier and continuing across a range of HPC application benchmarks, which include:

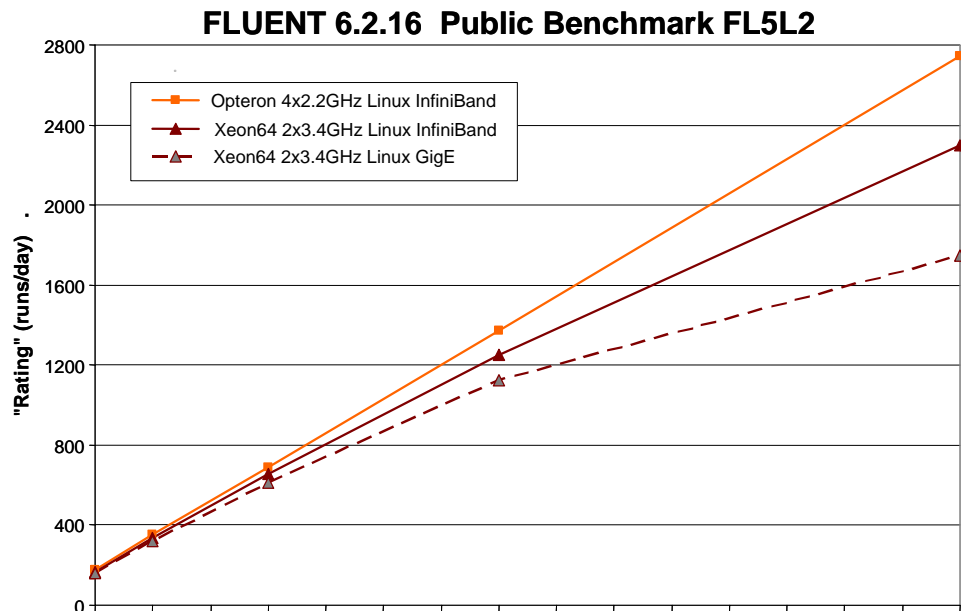
FLUENT

FLUENT CFD tools are used extensively in the Automotive, Aerospace, and Defense Industries to solve a diverse array of design challenges. Their strengths are their ease of use and versatility. With equal success they can be applied to compressible high speed flows and low speed, natural convection flows.

Physical data models allow for the rigorous analysis of the complex phenomena that occur in environmental control systems, anti-icing systems, aerodynamics, rotor-airframe interactions, fuel sloshing, fire suppression, propulsion systems, combustion, and moving-body problems.

The chart below shows the results of a recent test and shows that InfiniBand clusters have runtime improvements of 1.5 times that of fast Ethernet clusters and continues this performance trend as the number of cluster CPUs increase. It further shows that overhead associated with the CPU having to deal with the communications can affect overall performance.

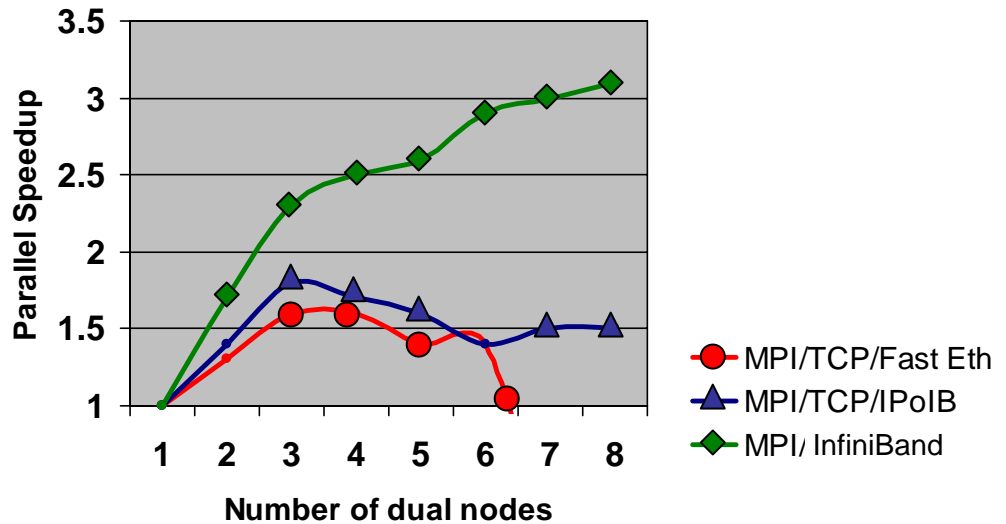
Real World application performance show the benefits of InfiniBand improve even more as the number of nodes increase.



LS-DYNA

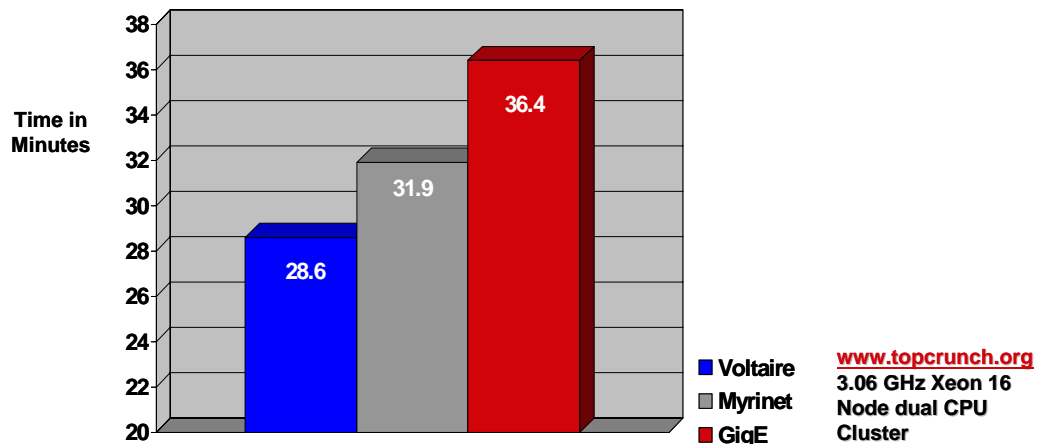
LS-DYNA is a popular finite element analysis application developed by LSTC. It helps car designers simulate crashes and structure modeling. Other product designers use various data sets to simulate dropping the product for impact analysis.

Small-car rigid pole (WPI rp_Isd93)



On identical platforms, Voltaire InfiniBand substantially outperforms Myrinet and GigE on HPC Commercial Applications

Being faster means that the product may be brought to market sooner than competition. The same test by Intel using LS-Dyna shows over twice the time performance using InfiniBand than fast Ethernet. As more nodes are added to the cluster, performance continues to improve. Again, CPU overhead affects performance using Ethernet.



As shown above, another example of the performance benefit InfiniBand solutions bring to the actual application performance is demonstrated when comparing identical systems with various interconnect solutions. A popular LS-Dyna benchmark site, Topcrunch.org, lists various hardware architecture solutions all running the same identical benchmark suite. The benchmark centers on crash analysis of a Dodge Neon collision and the time it takes to complete the simulation.

As the 32 bit Intel Xeon comparison shows, the InfiniBand solution is 23% faster than GigE, and 11% faster than Myrinet.

How Well Do Benchmarks Predict Performance?

In today's high performance computing arena, customers get better application performance from their InfiniBand systems than most of the standard benchmarks imply. Since InfiniBand offers better performance in many areas, applications and their data sets can be fine tuned for greater performance once deployed. Benchmarks and standard measurements help to see what systems should be considered, so Voltaire continues to be committed to running all the important tests for business and technical computing. As more application benchmarks become available, their results are showing InfiniBand leadership over other standard interconnect solutions.

InfiniBand continues to improve as manufacturers engineer better performance.

The family of InfiniBand solutions currently dominates a wide range of benchmarks, and has done so over a considerable period of time. This indicates a solid track record of expected performance. Now, as an emerging standard, the future growth path of InfiniBand is again widening the gap. In the age of Grid computing and virtual services, this means that low cost, high performance computers can keep up with the demands of the dynamic and the unpredictable; it also means tasks are done faster, costs are lower, and most importantly customers are more satisfied.

In general benchmarks are only indicators. Experience in matching the application parameters to take full advantage of the benefits of InfiniBand clusters is most important. Benchmarks for applications if managed correctly can give an excellent indication of comparative price/performance. Real applications such as Weather modeling (WRF) provide better benchmarks. In actual WRF benchmarks InfiniBand outperforms Myrinet by 30% with an increasing advantage as the cluster scales. In fact, InfiniBand users have reported that with the affordability and price/performance of InfiniBand systems, they are now able to afford to tackle problems that could not possibly have been solved before.

InfiniBand Scales to Thousands of Nodes

InfiniBand is designed to scale to hundreds or even thousands of nodes without the performance degradation that often occurs in other data center clusters. It incorporates several features that ensure consistent performance as fabrics increase in size, and a flexible connection routing scheme that allows fabric bandwidth to be allocated to meet the needs of the application.

As InfiniBand matures, more and more benchmarks point to the benefits of InfiniBand solutions.

Flow Control and Virtual Lanes Eliminates Congestion

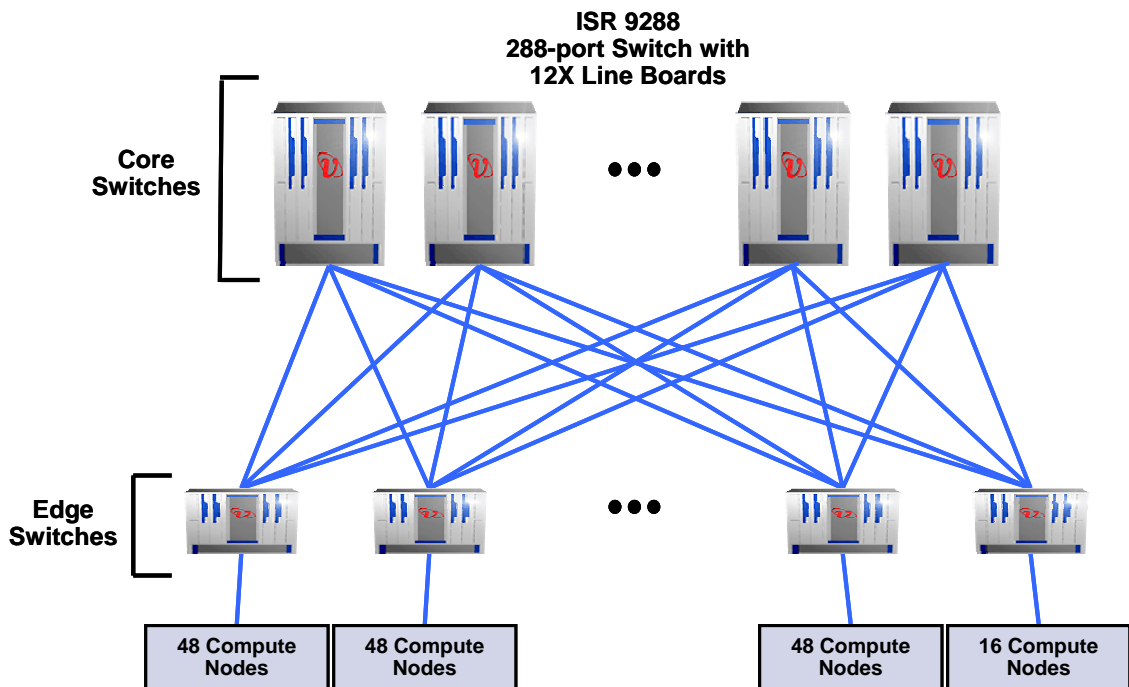
Flow control is critical to ensure efficient fabric bandwidth utilization and minimize congestion and latency. The InfiniBand specification includes a powerful credit-based end-to-end flow control mechanism that optimizes bandwidth utilization. End nodes only send as much data as the receiver has memory allocated to store it. This also eliminates wasted bandwidth due to retransmissions, which occurs when a TCP/IP/Ethernet network becomes congested and loses a packet.

In InfiniBand, multiple switching layers (up to 16) can be formed over the same physical link, each layer is isolated from the others and provides completely

independent flow-control mechanisms. These mechanisms called Virtual Lanes (VL) enable different classes of traffic such as IPC, storage I/O, or network (IP) I/O on separate isolated channels (planes). In addition, the availability of multiple traffic planes enables the spreading of real-time traffic across multiple layers to reduce congestion and interference (layered routing).

Trunk Links Simplify Multi-Tier Networks

InfiniBand defines a hierarchy of three link speeds (1X, 4X, 12X) that facilitate the design of non-blocking multi-tier fabrics. For example, 12X (30 Gbps) trunk links can be used to create a high performance two-tier fabric, with edge switches connected to core switches (see figure). Using trunks and a variety of switch platforms, users can create a range of fabrics to satisfy their unique bandwidth requirements, from fat tree topologies to simple cascaded networks.



InfiniBand fully specifies the switch-switch protocol so that these fabrics can be built using products from multiple vendors. The 12X trunk speed saves cabling and simplifies the management and operation of the fabric.

Flexible Routing Schemes

InfiniBand is designed to give network managers the flexibility to decide how to route connections across the fabric. The InfiniBand Subnet Manager (SM) function centrally configures the connection routing tables in each fabric switch. The SM can use a range of routing algorithms that can maximize bandwidth and/or minimize latency. The algorithms can be custom tuned to different application scenarios, can be fully automatic, or be controlled by the clients. In large fabric topologies, the SM fully exploits the available bandwidth and spreads the traffic across all available routes, as well as maintains a deadlock free network.

Scalability Features	InfiniBand	Myrinet	Ethernet
Flexible Routing	✓	✓	
Credit-based Flow Control	✓		
Virtual Lanes (VL)	✓		
Hierarchical Link Speeds	✓		✓
Source Load Balancing	✓	✓	✓

Voltaire brings high-availability engineering and high performance for InfiniBand cluster solutions.

InfiniBand’s routing flexibility contrasts sharply with Ethernet’s spanning tree algorithm, which provides only one path between nodes and uses only one algorithm (shortest path) to define that path. This is why Ethernet fabrics do not make full use of available fabric bandwidth, nor ensure congestion-free operation.

The standard also gives end nodes the ability to request multiple paths between source and destination. This source load balancing feature enables nodes to control how the traffic is distributed in the fabric and optimizes it by addressing specific application needs.

InfiniBand Increases Cluster Reliability

Voltaire’s extensive experience in high performance computing simplifies the adoption and reliability of InfiniBand for system development. This experience combined with InfiniBand standards ensures solid reliable solutions. The basic InfiniBand specification defines a multipurpose high-performance interconnect designed to deliver enterprise-class reliability. As such, it incorporates built-in mechanisms to ensure high availability. Comparable mechanisms are not available in any other system area network.

Several mechanisms guard against network failures. Automatic path migration enables an end-node to establish a standby path between source and destination. It will use this standby if it detects an interruption in service over the primary path. Separately, InfiniBand specifies a mechanism to detect switch/link failures by the subnet manager and immediately reconfigure the fabric (switches forwarding tables) to bypass the problem within milliseconds. Multiple subnet managers can coexist where one functions as a master and others as standby slaves. A standby SM can take the master SM’s place in case of a failure without any data loss or service degradation.

InfiniBand incorporates robust, hardware-enforced security mechanisms. It allows network managers to establish partitions to isolate groups of users/nodes from each other. It also features memory protection mechanisms that restrict RDMA access to an application’s memory space, and uses special keys for management access.

InfiniBand Allows Isolation and Partitioning

In many cases users want to utilize the same infrastructure for multiple applications simultaneously, those applications may require isolation from one another, whether it's for security purposes like in a multi-user on-demand center, or to guarantee fault or resource isolation.

InfiniBand provides strong hardware enforced partitioning, where each node may belong to one or more dedicated partitions. In addition, InfiniBand allows unique partial partition membership that allows access only to shared resources (not exposing the other clients in that partition using these resources.)

InfiniBand Supports Existing Codes

While InfiniBand may be a new interconnect technology, it provides complete transparency to existing software applications. The specification defines a flexible verbs interface that Voltaire has used to develop a comprehensive suite of upper layer protocols, including support for MPI and the Berkeley Sockets API.

Multiple MPI implementations provide transparent support for applications and access to InfiniBand's superior performance

Message Passing Interface (MPI)

MPI is the most widespread HPC API for interprocessor communications (IPC), and its usage dominates in the technical parallel computing field. For the runtime environment, MPI implementations may be either coded down to the driver level interface or may be layered on top of another API and runtime environments such as DAPL. The widely ported MPI-CH code base may map down to a TCP-based API, but it is more efficient to map to an API that comprehends Remote Direct Memory Access (RDMA), such as DAPL or directly to the InfiniBand access layer.

Message passing interface support is available from commercial and open source suppliers, including Ohio State University (MPI-CH). Voltaire supports additional commercial MPI protocol suites as well. These MPI implementations work transparently with many existing MPI applications and provide access to the superior performance and low latency characteristics of the InfiniBand fabric.

IP Network Emulation (IPoIB) and Socket API (SDP)

InfiniBand supports any standard/legacy network application by providing an Ethernet driver emulation over InfiniBand (IPoIB). IPoIB can also be used to communicate with external Ethernet networks through Ethernet router blades that fit into the Voltaire switches.

In addition, Voltaire is leading in the support of the sockets direct protocol (SDP), which provides a standard BSD Socket API interface enabling standard TCP/IP-based applications to interoperate without modification. Applications with native TCP socket level support will integrate seamlessly and be able to take advantage of the price/performance benefits InfiniBand offers. By moving these applications from an Ethernet network to an InfiniBand fabric, users can realize up to 5X bandwidth improvement with lower latencies and less CPU overhead.

High Performance Storage over InfiniBand

InfiniBand is the highest performing system fabric to date. This enable InfiniBand networks to run storage traffic between file servers or block devices and servers at unprecedented speeds. The Voltaire protocol suite supports access to block storage by using IETF's iSER protocol (iSCSI RDMA). iSER enables the use of block storage devices directly attached to InfiniBand (native) ports or externally in FC or iSCSI fabrics through router blade modules. These modules blades fit into the switches.

In addition it is possible to connect to high-speed file or object storage device that uses native RDMA protocols such as NFS/RDMA or Lustre.

All Voltaire protocols, include the MPI, DAPL and iSCSI RDMA (iSER), and are integrated and tested. Also included is Voltaire's TCP/IP/Ethernet LAN emulation capability, which creates a familiar easy to use Ethernet/IP look and feel to the solution.

Consolidating Server I/O

In InfiniBand, a single card and fabric are used for clustering, networking, and storage, through partitioning and Virtual Lanes (VL). The fabric resources can be segmented and prioritized/preserved dynamically. In addition the traffic on the InfiniBand fabric can be bridged transparently to existing external IP or storage networks while coexisting in heterogeneous environments.

Those capabilities of InfiniBand make it ideal to build large grids that are provisioned and virtualized dynamically through software. This significantly improves resource utilization, simplifying the deployment and maintenance of the overall infrastructure.

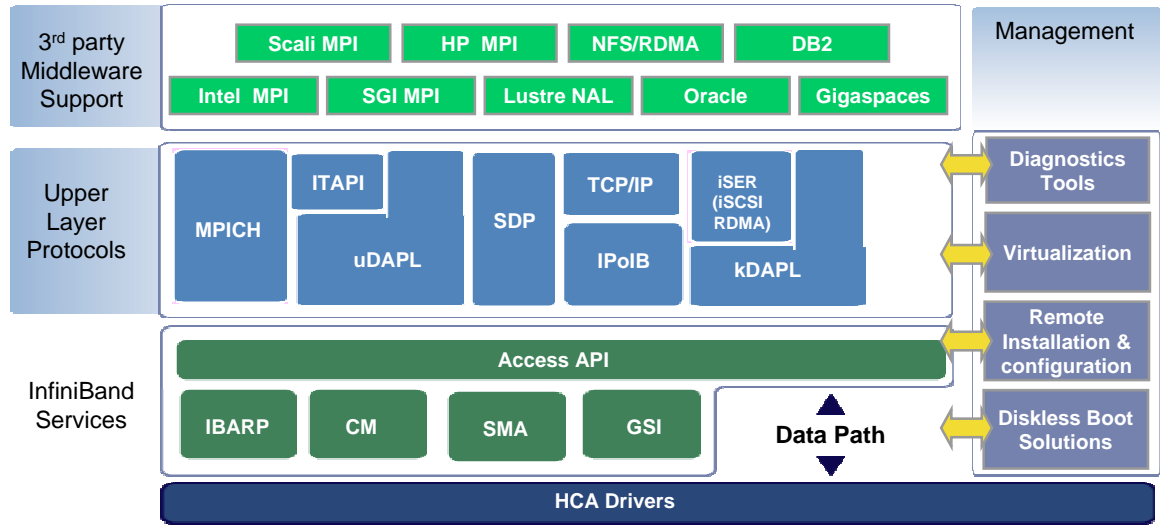
The underlying architecture is completely transparent to the host or target resources. When consolidating the server I/O, the result is lower capital and operations overhead. This includes physical costs associated with additional cards and wiring and time spent reconfiguring the cluster or grid. The management benefit is significant allowing a single point of managing multiple fabric technologies easily, and effectively.

The InfiniBand architecture integrates fiber channel storage, Ethernet, and RDMA resources into a single, easy to manage environment.

Voltaire InfiniBand Software Stack

UPPER LAYER PROTOCOLS (ULPs) PROVIDED WITH THE HCA

The host software implementation includes protocols and APIs for Networking (IP), Storage (SCSI), and IPC that will enable application and device transparency and make optimal use of InfiniBand and RDMA operations.



Voltaire Makes InfiniBand Transparent through unique protocol integration. The Voltaire InfiniBand technology suite enables companies to implement a high performance InfiniBand I/O system while maintaining the familiar TCP/IP/Ethernet management and administration paradigm used in legacy servers.

INTEGRATION AND APPLICATION TRANSPARENCY

This unique Voltaire capability means InfiniBand communications as well as Storage devices for clusters including File I/O, Block I/O and IP can be run, and managed, together over one fabric. They can continue to use familiar management tools and administrative interfaces while leveraging the full performance of a native InfiniBand server. Voltaire embeds unique software into each I/O system element that translates InfiniBand management constructs into traditional TCP/IP constructs. For example, unique IP addresses and IP naming services are translated into InfiniBand LIDs and services delivered by InfiniBand ULPs. Ethernet VLANs are translated into InfiniBand partitions.

When combined with upper layer protocols like SDP, the Voltaire InfiniBand technology suite creates a server system that is fully compatible with an existing server environment. The ULPs provide complete compatibility with network application programs based on TCP sockets and SCSI. The Voltaire management software emulates a TCP/IP LAN and enables existing third party management systems to configure and trouble-shoot the I/O system as if it were a standard LAN.

Voltaire experience helps customers to engineer clusters from entry level to thousands of nodes.

Voltaire HPC Fabric Solutions

Voltaire offers a complete family of end-to-end InfiniBand fabric solutions to meet the most demanding HPC applications. The family includes high performance PCI-X and PCI-Express host channel adapter cards and software, modular switch chassis supporting up to 288 ports at 4X, and an InfiniBand-to-TCP/IP router that provides complete TCP termination with layer 4 features for server virtualization and provisioning.

The Voltaire family of InfiniBand switch routers provides a comprehensive management suite, including performance management tools and utilities that monitor performance bottlenecks and provide statistics.

Voltaire MPI software support on PCI-Express slots shows the HCA400 achieving user-level unidirectional data rates of ~960MB/s, and is limited only by the bus. The user-level MPI latency is ~4μs using MVAPICH MPI supplied by Voltaire, or less depending on the server chipset.

This performance is significant with the new PCI-Express over similar interconnects with both better latency, and increased bandwidth. These results are representative of user level message-passing performance in application programs. The performance measurements are between user processes, with full protection using RDMA writes, and with end-to-end data-integrity checking.

InfiniBand and PCI Express: Voltaire HCA400/EX Adapter

The Voltaire HCA 400/EX series is a single chip, dual-port, 10Gb/s InfiniBand host channel adapter card that enables PCI-X and PCI-Express based servers to access the full performance of a high-speed InfiniBand fabric. It includes complete software support for a range of demanding enterprise and high performance computing applications. This features:

- **Kernel By-Pass:** This feature enables applications to eliminate kernel calls to the O/S, which greatly reduces CPU overhead, and latencies.
- **Transport Off-load Hardware:** Enables wire speed capabilities as well as lower latencies.
- **RDMA (Remote Direct Memory Access):** Direct access to remote memory eliminates the need to copy data, which saves host memory bandwidth.

The HCA 400/EX features Voltaire's TCP/IP emulation software, which makes the InfiniBand fabric appear to the server operating system and application programs as a standard Ethernet network. This HCA card with an 8X PCI Express interface enables an aggregate of 20 Gb/s of full duplex bandwidth. The PCI Express version offers two to four times peak bandwidth improvement over PCI-X slots coupled with lower latencies.

Voltaire ISR 9288

As the industry's largest production Multi-Service switching solution, the Voltaire ISR 9288 enables high performance applications to run on a distributed server with storage and network fabric solutions fully controlled by a single management resource. It is elegantly designed with 10 GB/s of full bisectional bandwidth for up to 288 ports in a single chassis. The 9288 allows any combination of Ethernet, FC, or

InfiniBand ports and advanced layer 2-7 switching capabilities. Multiple Voltaire ISR 9288s can also be interconnected to form very large clusters. Possible configurations range from dozens to thousands of nodes with "pay as you grow" scalability in any direction; storage, network I/O, or IPC.

Voltaire ISR 9096

Similar to the ISR 9288 in every way except port count, the Voltaire ISR 9096 is designed with 10 GB/s of full bisectional bandwidth for up to 96 ports in a single 6U chassis. Voltaire ISR 9096s can also be interconnected to storage targets and external networks through integrated layer 2-7 router modules. This makes the ISR9096 perfect for commercial grade grid solutions. With many components interchangeable with the ISR9288, configurations and ranges are flexible with "pay as you grow" scalability in any direction; storage, network I/O, or IPC.

Voltaire ISR 9024

The Voltaire ISR 9024 switch family is a high performance, low latency, fully non-blocking switch with a throughput of 480 Gb/s. Designed for high availability and easy maintenance, the Voltaire ISR 9024 is a cost effective alternative to proprietary interconnect technologies and is totally IBTA 1.1 compliant. With twenty-four, 10 Gb/s ports in a 1U chassis, the standards-based Voltaire ISR 9024 delivers three times the port density of proprietary offerings. It is engineered to run without the need for integrated cooling fans allowing for reliable, quiet performance and can be mounted in any orientation. Hundreds of ISR9024 switches can be centrally managed through the Voltaire Fabric Manager providing simple management and configuration eliminating the need for on board element management.

Using the Voltaire ISR 9024, end users can build InfiniBand fabrics that scale from several to thousands of nodes. It is available with fully integrated management or, in a unique configuration without management. This allows even better price performance for larger clusters. A single 9024 switch equipped with the management module can manage a single cluster of thousands of nodes. Other switches can be the lower cost units without the integrated management application.

Voltaire ISR 6000 and router modules

The Voltaire ISR 6000 is a flexible 1U platform for creating high performance InfiniBand clusters and interconnecting them with fiber channel SAN or TCP/IP/Ethernet networks. The Voltaire ISR 6000 is designed to enable data centers to leverage the performance of InfiniBand-based clusters by making the fabric accessible to a TCP/IP infrastructure. A complete, three-module system provides high performance connectivity between InfiniBand-enabled servers and storage and server systems resident on a data center TCP/IP network.

Comprehensive Management

Voltaire provides a powerful and comprehensive management suite that simplifies the management and proactively maximizes the performance and availability of

InfiniBand enabled servers, networks, and storage grid environments. The VoltaireVision suit manages the fabric and networking resources in the grid, including the ability to automatically diagnose, optimally configure, provision, monitor traffic, partition, and fail-over. VoltaireVision is embedded directly into Voltaire InfiniBand Switch Routers. The VoltaireVision software suite enables the scaling of standards-based grid environments to thousands of nodes with minimal user intervention making clusters and grids an easy-to-use application work environment.

VoltaireVision is based on industry standard APIs and can manage 3rd party InfiniBand switches, as well as be integrated with a variety of existing management platforms. Configuration and maintenance of the fabric are simplified with a full-featured graphical user interface (GUI) and command line interface (CLI) via serial console, Telnet, SSH or HTTP. It enables dynamic provisioning, remote monitoring, upgrades, and troubleshooting.

At a high level, VoltaireVision enables administrators to manage Voltaire's entire product family of switch routers and InfiniBand adapters and provides the essential subnet management (SM) functionality required for any InfiniBand fabric. Inspection mechanisms are aimed at detecting faults, miss-configuration, performance bottlenecks, and security violations, and to act by notifying the administrator and/or by automatically performing corrective measures. Moreover, VoltaireVision facilitates the integration of InfiniBand fabrics to storage and IP networks by providing virtualization and advanced routing capabilities within the fabric and for I/O connectivity and data storage devices.

Conclusion

InfiniBand technology shows promise by bringing performance and scalability to the high performance computing industry at attractive costs. It is the core technology systems integrators and manufacturers see as the interconnect of choice. Grid computing solutions and blade server technology depend on InfiniBand to offer the speed, and lower cost necessary to build high performance solutions. With InfiniBand solutions, organizations can create cluster fabrics with terabits of bandwidth and build the next generation high performance computing systems.

InfiniBand's scalable bandwidth, low latency, and high availability characteristics make it much more cost effective than third-party and proprietary interconnects. With its native RDMA support, InfiniBand eliminates I/O and memory bandwidth bottlenecks increasing CPU availability to applications. Ethernet solutions cannot compare to the performance and scalability.

The InfiniBand technology is proven and as a standards based solution, more and more companies are providing support for InfiniBand solutions in HPC and enterprise applications. The experience in high performance computing and certified solutions Voltaire brings to InfiniBand gives ideal interconnect solution for performance clusters.

This scalability and performance is offered at the attractive prices of an open industry standard technology. Strong future development by leading systems and Blade computing manufacturers to integrate InfiniBand technology shows strong acceptance toward this interconnect standard. Voltaire products achieve price-performance ratios several times better than competing technologies. Combined with Voltaire's innovative Fibre Channel routers, TCP/IP/Ethernet routers and VoltaireVision technology, organizations can dramatically improve total cost of ownership for their cluster infrastructure and a faster return on IT investment.