Mellanox Technologies

White Paper

InfiniBand™ in the Internet Data Center

Reliability, Availability, and Serviceability Past, Present, and Future

1.0 Introduction

The Internet Data Center is central to the success of Internet-based companies, whether the business focus is an E-portal, E-Commerce, or Business-to-Business. There are huge productivity gains that can be recognized by moving customer and supplier transactions on-line. As a result, the Internet Data Center has become critical to the profitability of traditional brick and mortar companies. In order to keep up with the unprecedented growth in the Internet, companies have been driven to develop the infrastructure of the Internet Data Center in a rather organic and *ad hoc* fashion (which has unfortunately resulted in inefficiencies and at times chaos). InfiniBand has been developed specifically to address the needs of the modern Internet Data Center and will radically change the nature of systems being deployed in the buildout of the Internet infrastructure and greatly improve RAS (Reliability, Availability, and Scalability).

The InfiniBand Architecture, defined by the InfiniBand Trade Association (IBTA, see http://www.infinibandta.org), is the result of an industry collaboration among most of the major companies in the server, storage, and networking markets. The steering committee includes Compaq, Dell, HP, IBM, Intel, Microsoft, and Sun Microsystems. The sponsoring companies comprise of 3Com, Adaptec, Agilent, Brocade, Cisco, EMC, Hitachi, Lucent, NEC, Nortel, Siemens-Fujitsu. At the most basic level, InfiniBand replaces the PCI bus and provides the next generation of I/O connectivity to PC and server platforms. InfiniBand is, however, a revolutionary, rather than an evolutionary change to the traditional PC bus based architecture, as InfiniBand defines a switch based serial I/O fabric. This means that links can come out of the box, enabling flexible network connections that can scale and provide fault tolerance. Infiniband represents a major architectural shift for the compute server platform with far reaching implications for basic server architecture and system level deployment.

2.0 RAS Capabilities

Proprietary content is the outward face of most E-businesses and is essential to the success of the business. Equally important, however, is the underlying infrastructure supporting the content or services. The key requirement of the Internet Data Center infrastructure is to provide cost effective implementations that support RAS: Reliability, Availability, and Serviceability. Reliability requires that transactions are accurate and complete even in the face of equipment failure. Availability must provide the customer with continuous and uninterrupted service. Serviceability must allow the provider to dynamically service and add capability to the Internet Data Center in order to support an increasing number of users and information. These RAS features must be augmented by system scalability such that the system can grow without requiring the entire system to be brought down in order to bring new capacity on-line.

Reliability and availability go hand in hand and are an economic imperative, because downtime is expensive as evidenced by the costly web outages of E-commerce, auction sites, and web-based financial companies. Traditionally, reliability has meant first assembling systems from higher grade components, and second, implementing fault tolerant systems (since no component can ever be 100% reliable). Fault tolerance is implemented by utilizing redundant components, such that, if any given component fails, the system as a whole continues to operate. The eliminations of single points of failure (SPOF) is key to the design of any highly available system.

2.1 Fault Granularity

As will be shown later InfiniBand effectively changes the way servers are packaged. The first major change in server packaging is well under way with storage having been moved to external NAS or SAN connected RAID arrays. InfiniBand completes the *disaggregation* of the server moving I/O connections out of the box as well. This fundamentally changes fault granularity of a highly available system. InfiniBand breaks the one to one relationship between server and I/O elements, implied by the PCI architecture. In the PCI architecture a given network interface card (NIC) is physically located in a server and associated with this one and only one server. Should the card fail the unit of fault granularity is the entire server-I/O complex. Since the bus is shared dual NIC solutions do not completely eliminate NIC failures as SPOF's. Thus in the conventional fault tolerant clustered architecture when a NIC card goes down within a server, the entire server is taken down and must be failed over to a redundant server-I/O complex. The server hardware itself may be completely functional and only the NIC needs be replaced, however a substantial portion of the cost of maintaining the network is in such management tasks as debugging the failure mechanism. In many cases it is simpler (and cheaper) simply to replace the entire subsystem *in-toto* rather than try to isolate the failed component. Thus it is important from a resource perspective to minimize the fault granularity.

InfiniBand breaks the conventional one-to-one relationship between server and I/O and enables a many-to-many server-to-I/O relationship. This changes the fault granularity of a highly available system based on a disaggregated clustered architecture. When a NIC card fails, the failover process need only migrate the connection to another redundant I/O device. This not only saves resources in terms of the failed element that is taken off-line, it also simplifies the failover process itself. This is so because migrating a *transaction* from one server to another requires a consider-

able amount of state be regularly mirrored to another redundant server. This level of failover capability is considerably more difficult than transitioning to an alternate I/O device while maintaining a connection, since InfiniBand itself supports reliability in hardware.

2.2 RAS Features of InfiniBand

While essential, these RAS functions are not built into the low level protocols of the internet. Traditionally, reliability has been supported by clients operating in client/server configurations, relying on higher level software layers (TCP/IP) to provide limited recoverability from equipment failure. This method is both inadequate and inefficient for a modern Internet Data Center and has led to some interesting *ad hoc* strategies to provide for RAS. As the nature and complexity of the Internet has grown these strategies are being stretched to their limits and are imposing serious cost and performance penalties.

The InfiniBand Architecture has been developed from the ground up to provide a cost effective, high performance solution with RAS support for the Internet Data Center. The architecture supports many RAS features including CRCs (Cyclical Redundancy Check), reliable transport, and failover. The InfiniBand link protocol incorporates multiple CRC fields, providing error detection capabilities on both a per-hop link level and an end-to-end basis. The 16 bit VCRC (Variant CRC) field is recalculated at each hop and checks the data integrity of an individual point-to-point link. The 32 bit ICRC (Invariant CRC) remains constant from source to destination by masking bits which can change hop to hop (such as the VL field), thus providing end to end assurances of data integrity. A second RAS feature is the InfiniBand transport support of reliability for both connection and datagram services. A large packet sequence number space (24 bits) and sophisticated acknowledgment protocol insure the hardware delivers robust, error-free transport services in hardware. The InfiniBand architecture defines a failover mechanism which allows a network to heal itself as individual links fail. InfiniBand subnets, which incorporate redundant connections, can detect link errors and migrate traffic from a failed link to a redundant link that is still functional. InfiniBand supports both managed and un-managed failover. Un-managed failover is essentially a hardware mechanism whereby traffic is switched to a pre-allocated redundant channel automatically as errors are detected. Managed failover is a function of the subnet manager detecting errors in the fabric and reconfiguring the forwarding tables of individual switches.

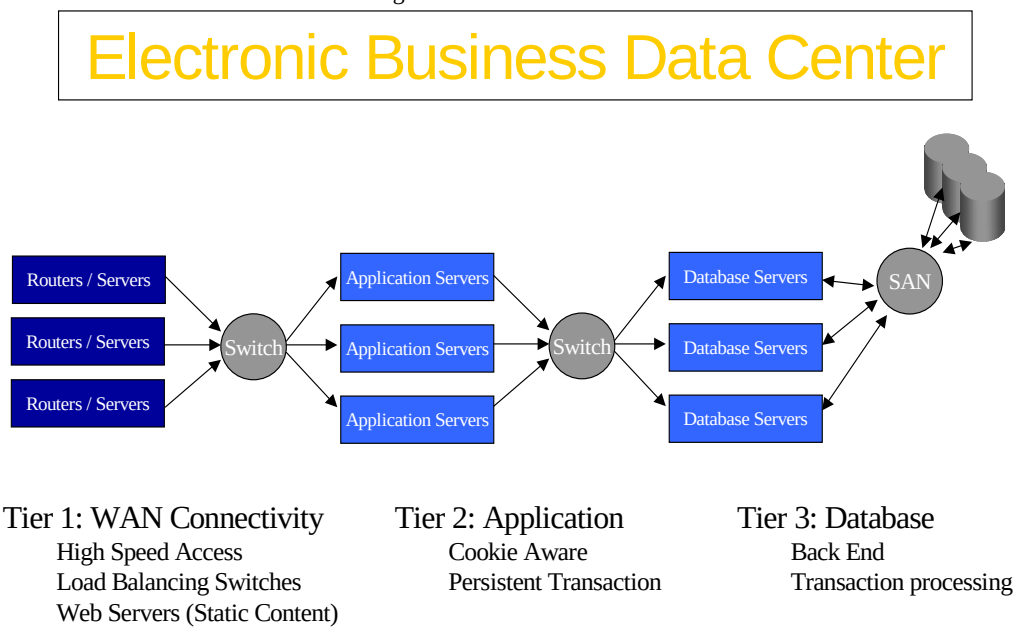
Despite InfiniBand’s advanced RAS features, the technology world often belies the adage: *if you build a better mouse trap, they will beat a path to your door.* The information highway is littered with the carcasses of manifestly superior technologies that simply did not succeed in the market for one reason or the other. Thus beyond demonstrating the technical benefits of InfiniBand, it is vital to demonstrate its economic benefits in order to insure InfiniBand’s success in the Internet Data Center.

3.0 Three-Tiered Implementation

Today’s state of the art implementations of Internet Data Centers employ what is known as a three-tiered infrastructure. There are several important points to note about this three-tiered implementation. First, each tier is implemented with redundant components, such that, failure of any one component does not necessarily cause the system to fail. Second, each tier provides spe-

cialized services as the components within each are optimized to provide these services. The first tier provides the WAN connectivity, load balancing, and static web content. The second tier provides application services that are typically specific to the user. The third tier (or backend) is typically a heavyweight implementation supporting the complete enterprise database and transaction processing capabilities. One advantage of having a tiered infrastructure is that each portion of the infrastructure can be maintained and upgraded independently.

Figure 1. Electronic Business Data Center



In the early days of the web, a single tier implementation was sufficient as all web content was static and web based applications were primitive. Today sophisticated web based applications are supported by dedicated servers. The predominant communication protocol between tiers is TCP/IP, typically running over some form of Ethernet based local area network. One notable exception to this is the SAN (Storage Area Network) in the third tier which is typically implemented using Fibre Channel based technology.

In addition to providing WAN connectivity, the first tier evolved to provide load balancing switches. A basic load balancing switch simply monitors the utilization of the various servers and funnels client requests to those with the lowest workloads. This requires the load balancing switch to monitor the server's workload and distribute client requests to static web content servers accordingly, but does not require the switch to parse the client request.

The second tier application servers typically utilize user information (in the form of cookies stored on the users machine or sent as part of the http header) to determine web content. Furthermore, in e-commerce or other applications where user transactions occur, the server must maintain a persistent connection with a specific user. This requirement of a persistent connection between the client and server makes the task of load balancing much more difficult, since the switch must parse the data received by the client and determine the exact server supporting that client. This requires the load balancing element to look deeper into the data to determine the content.

As previously noted, three-tier networks utilize TCP as the communications protocol between tiers. This method becomes problematic as persistent connections are required. When serving static web content, a load balancing switch may distribute each request from a client independently to the appropriate server based on work load. However, a cookie aware application server provides content based on the context of the user and therefore a TCP/IP session must be persistent. Once a server has been allocated to a user and has generated user specific information, this client/server connection needs to persist throughout the life of the transaction. However, the complicated structure behind the scenes must be hidden from the user. The user only sees a single URL, and multiple servers with various IP addresses serving this location must be hidden from the client. This requires the load balancing elements to perform some port and sequence number manipulations to maintain a persistent connection and simple client interface. Essentially the entire TCP/IP protocol stack must be duplicated at each node, requiring the data center to implement significant hardware and software overhead.

These persistent connections are maintained as sockets between the ultimate endpoints. A sockets is the standard layer 5 interface above TCP used to maintain a connection. A socket connection relies on an underlying transport mechanism offering reliable, in-order, connected service. TCP offers such a service however since it is implemented in software requires powerful hardware in order to support many TCP connections (general purpose CPU, network processor, or dedicated ASIC, plus memory subsystem). The TCP protocol stack is necessarily complex since the underlying link and network layer protocols offer only best effort, out-of-order, datagram services. In the heterogeneous environment of the Internet at large the TCP has evolved to include mechanisms to cope with the unpredictable and dynamic nature of the network. In the controlled environment of a data center these features of TCP are overly complex and a simpler protocol is preferable.

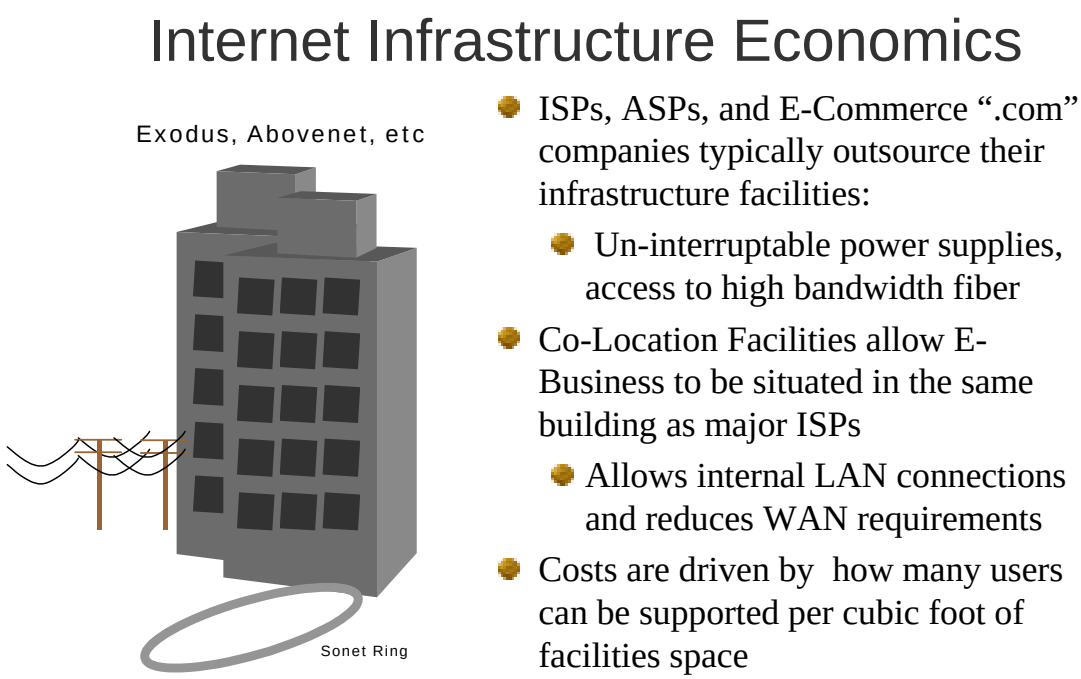
InfiniBand offers a Reliable, In-order, Connection (RIC) transport service, implemented *entirely in hardware*. The challenge is to be able to preserve the legacy application software which interfaces at the socket level, and still be able to take advantage of the performance and scaling benefits offered by InfiniBand's hardware implementation of RIC transport service. Several "Sockets Direct" implementations make this possible. These preserve the standard sockets API's yet do not mandate that TCP is used as the underlying transport. Applications utilizing sockets direct interfaces can use *any* underlying RIC transport service including InfiniBand. This allows applications originally written assuming TCP transport to run unmodified over InfiniBand. Furthermore applications optimized to take advantage of InfiniBand hardware will not require CPU cycles to support TCP and thus run substantially faster.

Thus an alternative data center architecture is suggested. The TCP connection may be terminated at the edge of the data center and a direct socket connection maintained to the ultimate connection endpoint within the data center. This architecture localizes the TCP processing specifically to where it is required at the edge of the network. Highly optimized InfiniBand connections can then be maintained within the more controlled environment of the data center. The transparency to higher level applications offered by Sockets Direct implementations is key to this new architecture.

4.0 Internet Data Center Economics

Technical aspects aside, much can be said about the economics of the modern three tier Internet Data Center. The “always-up” requirement of a modern Internet Data Center has serious implications in the attempt to maintain a functional infrastructure in the face of power failure, back hoe cuts, and even sabotage. The cost for individual companies to develop facilities resilient to such failure modes is prohibitively expensive.

Figure 2. Internet Infrastructure Economics



Thus, a viable business model has developed where these building and infrastructure facilities are out-sourced to providers such as Abovenet, Exodus, and Qwest. These facilities providers can amortize the cost of redundant high bandwidth wide area connections, un-interruptible power supplies, and even armed guards, over all of their customers co-located in their building. These so called “CoLo” facilities (short for co-location) provide only the facilities infrastructure and the actual equipment deployed within the leased cage is up to the individual Internet Data Center system manager.

Such co-location provides further benefits, since the clients from an Internet Service Provider are frequently browsing content or performing transactions connected to any other E-Commerce site or service provider situated within the same structure. Thus a local high bandwidth connection can be made within the facility thereby reducing the requirement for costly wide area network bandwidth.

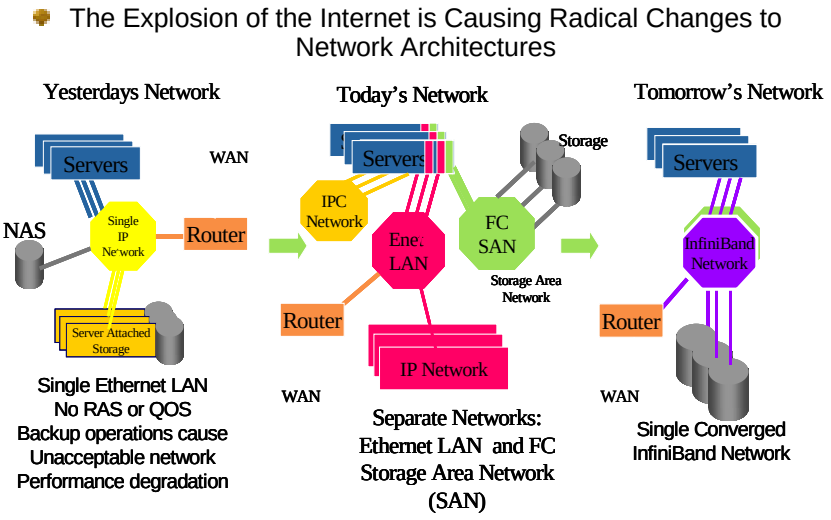
The benefits of co-location are not without costs. Rent in these facilities is the dominant expense for many E-businesses. Rates can range from $60 to more than $100 a square foot for space in such a co-location facility with limits placed on power consumption. Thus, building a highly reliable Internet Data Center boils down to a calculation of how many users can be supported per

cubic foot of space, producing the demand for smaller servers capable of processing more and more user transactions.

5.0 Storage and System Area Networks

The SAN acronym originally referred to System Area Networks but has more recently been used to describe Storage Area Networks and in this section the term “SAN” will be used to refer to the latter. Fibre Channel SAN deployments have only recently been accepted as a mainstream storage technology. Segregating storage on its own network provides several benefits including: isolation of network and backup traffic, scalability, fault tolerance, and server-less backup. These same benefits can also be provided by Network Attached Storage (NAS) solutions. The fundamental difference is that NAS uses files as the unit of storage while SAN’s use block based I/O (SCSI).

Figure 3. Evolution of the Data Center Architecture



Traditionally storage has shared the same local area network as compute servers. In this environment adding storage to the enterprise Internet Data Center means adding additional hard drives to expensive servers with large disk storage capacities. Upgrading a server to add more storage typically requires scheduled down-time, and server attached storage offers only limited scalability. Adding additional servers with attached hard disk drives is an expensive mechanism to scale storage. Furthermore normal enterprise practices require periodic scheduled backup of data to tape drive units. During backups, the resulting network traffic affects compute server traffic, resulting in noticeably degraded performance. This degraded performance impacts users who may experience unacceptably long delays. In the case of a web based company this can result in lost revenues as fickle clients jump to a competitive site with better performance. Furthermore SANs enable remote mirroring of data, thereby supporting disaster recovery.

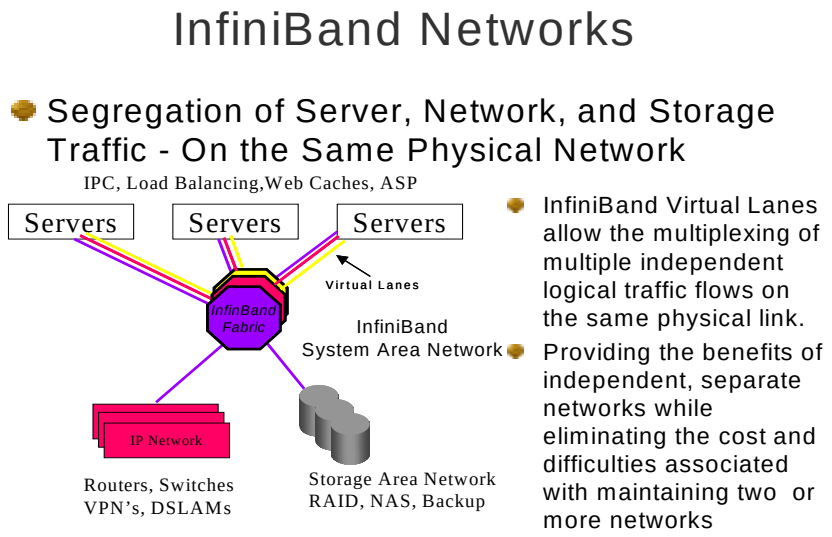
For these and many other reasons, SANs have become attractive alternatives to server attached storage. A Fibre Channel SAN creates a completely independent network from the compute local area network. This segregation allows low priority backup or mirroring operations to occur without affecting the latency-sensitive compute traffic. Thus, network compute performance does not suffer when scheduled backups occur.

Having multiple independent networks does not however, come for free. Connecting to multiple networks requires separate Network Interface Cards (NICs) for each server. These NICs are called Host Bus Adapters (HBAs) which must be installed in each server to provide connectivity to the

Fibre Channel network. These separate cards impose cost, power, and area expense to each server attached to the separate networks. Furthermore separate networks require multiple management functions with possible complex interactions between these functions. In addition, the IT manager must now contend with multiple technology roadmaps, thereby complicating the hardware and software upgrades required to keep pace with technological advancements.

There is a strong drive to converge these disparate networks back into a single network, where it began. Essentially there are three candidates that one can offer as the unified fabric for the Internet Data Center: InfiniBand, Fibre Channel, and Ethernet. It is our assertion that InfiniBand, over time, will prevail as the unifying system area network at the heart of the compute server infrastructure. The basic reasons that InfiniBand will replace other technologies include cost, power, features, performance, and RAS.

Figure 4. Infiniband Networks



The motivation for converging the networks is clear but begs the question as to how this can be done without experiencing the same issues that drove network administrators to separate networks to begin with, namely congestion. Answering this question requires some insight into the link level architecture of the InfiniBand protocol itself. The InfiniBand layer two link protocol includes the notion of Virtual Lanes (VLs), which allow multiple independent logical data flows to be multiplexed onto the same physical network. InfiniBand assigns separate buffering to these independent data flows and creates a truly segregated network environment. InfiniBand defines an arbitration architecture between these flows which combines both a priority and weighted round robin mechanism plus a starvation prevention mechanism to insure fairness. The net result is, tape backup traffic (or other less mission critical traffic) can be given a lower priority than compute intensive network traffic. Traffic on one Virtual Lane can be experiencing congestion and flow control (either link level or end to end) without affecting traffic on a higher priority Virtual Lane even though both traffic flows on the same physical wire. The use of Virtual Lanes to enable a system area network is shown in Figure 4, “Infiniband Networks,” on page 8. InfiniBand’s ability to provide Quality of Service (QoS) is central to the design of a converged system area network for the Internet Data Center.

6.0 Virtual Interface Architecture in the IDC

In the traditional I/O Architecture model, communication between elements is implemented by treating I/O as memory mapped devices. Now that CPUs have surpassed the GHz performance

threshold (greatly exceeding traditional I/O performance), they must poll remote devices in a tight loop waiting for I/O operations to complete. Infiniband, supporting the Virtual Interface Architecture (VI), decouples CPU and I/O operation and allows processors to perform useful work rather than waiting very fast.

Hardware support for the VI is a key component of the InfiniBand architecture. The Virtual Interface Architecture is a channel based software model that enables devices to access the memory of remote devices and similarly allows remote devices to have controlled access to local memory. InfiniBand offers hardware support for VI resulting in higher performance than other interconnect architectures which implement these functions in software. The primary benefits of VI derive from:

- The use of work, completion, and event queues to decouple CPU and I/O operations
- The use of a message passing channel based architecture
- Hardware support for protected remote DMA (RDMA) operations
- Enables efficient interprocessor communication IPC or clustering

With VI, the CPU uses work, completion, and event queues to manage I/O operations. The use of these queues decouples the CPU operation from remote I/O devices, thereby increasing efficiency and reducing CPU overhead. The CPU posts work requests to a queue, and then monitors their completion via either the completion or event queues. Work requests create Work Queue Elements or WQE's (pronounced "Wookies") on the work queues. Pipelined operation is enabled since WQE's may be outstanding on multiple work queues. This provides several benefits. First the CPU no longer performs I/O operations through load and store commands to a memory mapped I/O device. Today's super-scalar processors with Ghz performance levels can perform millions of operations in a millisecond and it is important that these operations not be spent in a useless spin loop waiting for an I/O device to signal that it has data or space available. With the VI architecture the I/O devices themselves schedule activity as resources are available. This results in better CPU utilization as the processor no longer needs to poll the I/O device waiting for results.

7.0 Why InfiniBand in the Internet Data Center?

InfiniBand is architected as a low overhead protocol implementation from a software perspective and does not attempt to solve issues associated with a wide area network. Instead, it assumes that the typical Internet Data Center provides a secure environment of trusted servers and a relatively controlled bandwidth environment. As a result InfiniBand is able to support the CPU intensive portion of I/O operations in hardware, decreasing the load on the CPU and thereby allowing the number of user processes supported per server to increase.

In contrast, the Internet is characterized by heterogeneous link bandwidths, unreliable connections, dynamic routing, various protocols, and variable latency. The TCP/IP protocol was designed with these constraints in mind and provides robust capabilities designed to operate in variable connections typical of the Internet. Such robustness does not come without cost and TCP/IP is not a lightweight protocol, requiring significant CPU processing overhead. TCP/IP requires that the CPU attempt to determine the link bandwidth with the so called "slow start" windowing

algorithm, allowing efficient utilization of the underlying network. However, it is expensive in terms of transport complexity. TCP/IP must maintain various software counters and information about the state of each connection. The complexity of the TCP/IP transport fundamentally limits how many sessions a processor can support simultaneously. Despite its complexity, TCP/IP does not provide the basic RAS features required of the Internet Data Center, nor does it provide any native support for QoS.

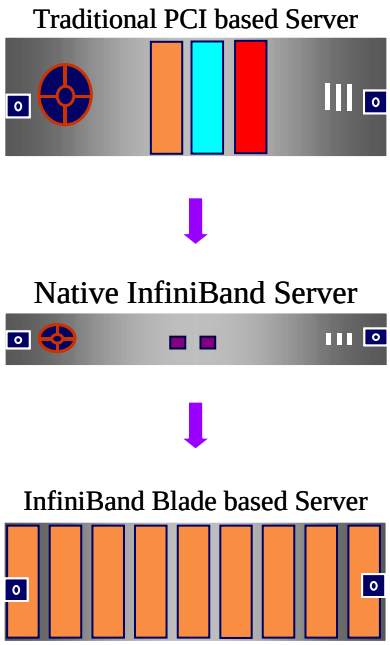
This complexity is essentially overkill in the more controlled network environment of the Internet Data Center. The InfiniBand architecture takes advantage of the more controlled environment and utilizes hardware to provide reliable underlying connections, de-multiplexing of process data flows, and zero copy remote DMA. The net effect is that InfiniBand transport substantially reduces the CPU load, thereby allowing a given compute server to support considerably more simultaneous communications channels.

Furthermore, InfiniBand provides native QoS support even at the layer two link level, rather than trying to add this functionality as an afterthought. The ability to multiplex multiple logical connections on the same physical connection is critical to the re-integration of the System Area Network. Providing networks with independent traffic congestion characteristics was the original rationale for the segregation of System Area Networks into Local Area Networks and Storage Area Networks. InfiniBand offers Virtual Lanes, which offer independent logical networks, on the same physical network, providing the benefits of logical segregation with the cost and management advantages of an integrated network.

Thus TCP/IP will remain the dominant protocol of the Internet. However, InfiniBand will become the preferred transport to provide connectivity between servers, storage, and I/O in the more controlled environment of the Internet Data Center. Initially conventional device drivers will provide InfiniBand support for the major operating systems. These device drivers preserve the higher level “socket” interface allowing applications, originally written assuming TCP/IP transport, to run transparently over InfiniBand. As independent operating system vendors develop kernel level support for InfiniBand, these vendor supplied device drivers will become unnecessary, and standard applications programming interfaces will emerge. This smooth transition path with application level compatibility, allows clustering, storage, and management software to take advantage of the RAS and QoS features of InfiniBand, while maintaining transparent support for legacy applications.

8.0 InfiniBand and the Evolution of the Server

Figure 5. IB and the Evolution of the Server



- Scales by adding I/O cards
- CPU must demux and copy data for each I/O process
- I/O cards add power and increase size
- Upgrades - Open the box

- Enables high density 1U stackable servers and then to even higher density blades servers
- InfiniBand pushes I/O to the edge of the network (Disaggregation)
  - Eliminates I/O cards
  - Amortizes the cost of I/O
- Virtual Interface Architecture
  - Major performance improvement

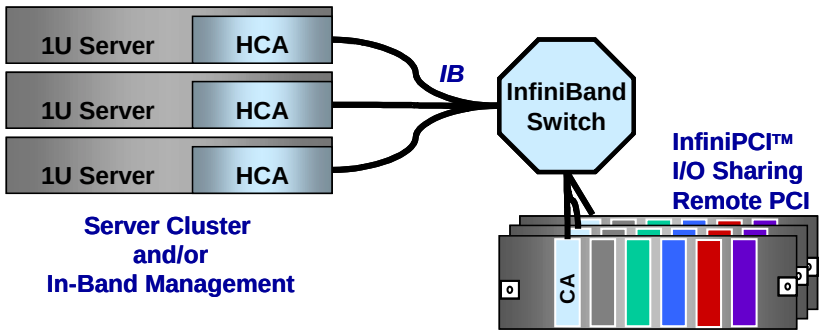
InfiniBand allows the “densification” of the Internet Data Center by providing servers with native networking capabilities which not only optimize performance but allow small form factor 1U platforms and eventually server blades. Using InfiniBand, the server is reduced to effectively three basic components: CPU, Memory, and I/O (InfiniBand). This reduction in components will enable the development of Server Blades. These 4x6 inch blades will contain a CPU plus memory and InfiniBand as the I/O interconnect that communicates with the System Area Network (SAN). These blades will be mounted vertically in 3U server chasis to a backplane, greatly increasing the density of CPUs in the server rack. Communications functions are disaggregated or pushed to the edge of the network and are amortized over many servers. This is simply an extension of the specialization occurring in the Three Tier Internet Data Center.

9.0 Migration of InfiniBand Solutions into System Area Networks

InfiniBand growth will be created by the continued growth of three-tiered data centers, expansion from IPC (or clustering) and the replacement of other technologies. This migration will happen in a few steps: Infiniband products better enable IPC, replace older IPC interconnects, relocate PCI and replace Fibre Channel. The migration of InfiniBand into System Area Networks and the evolution of ever-denser severs can be categorized in three distinct phases.

In the first phase, Figure 6, "IB System Area Network: Phase 1," on page 12 this migration begins with InfiniBand Host Channel Adapters (HCA) that utilize a single existing PCI slot within a server quickly enabling IPC support. Clustering will provide improved compute performance as multiple processors can be utilized efficiently solve complex tasks.

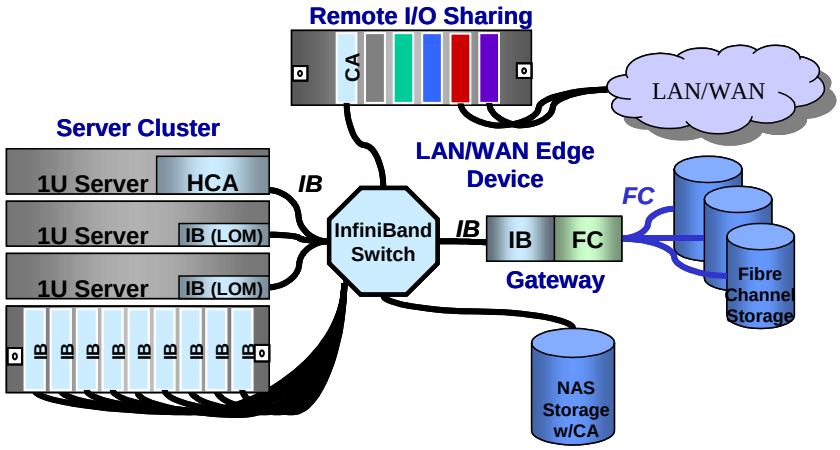
Figure 6. IB System Area Network: Phase 1



HCAs will also provide, with the aid of Mellanox's unique InfiniPCI™ Technology, the ability to transparently create standard PCI to PCI bridges over InfiniBand that will allow most PCI cards to be removed from the server and located in remote PCI or CompactPCI chassis. Mellanox's InfiniPCI Technology enabled in Mellanox's InfiniBridge™ product family has the ability to be recognized by Windows 2000 or Unix as a standard PCI to PCI bridge with no PCI hardware, BIOS, operating system, or device driver changes (see Mellanox for more info). This technology allows for greater density within a server rack as 1U servers, enabled with InfiniBridge, can support multiple PCI cards via a remote chassis over InfiniBand. The disaggregation of the sever has begun. Where Fibre Channel SANs are implemented the HCA would most likely remain in the server, preventing 1U utilization. (We'll show how this barrier is removed in the next phase.)

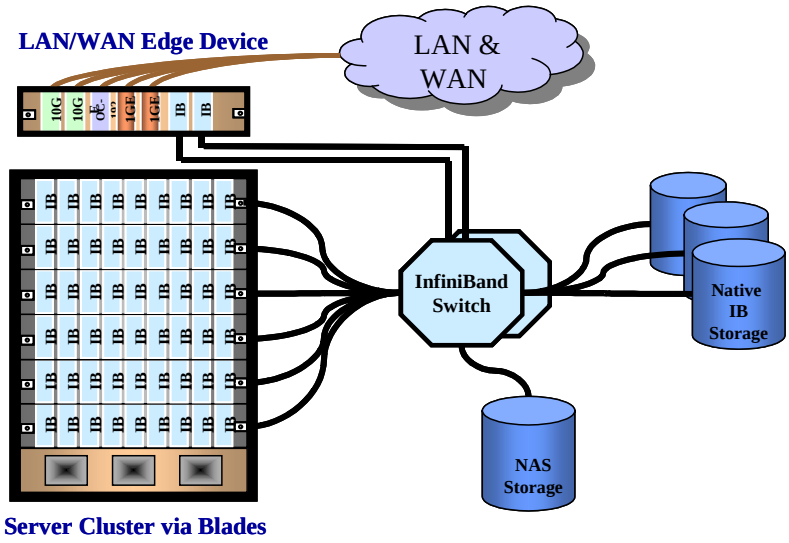
The second phase, Figure 7, "IB System Area Network : Phase 2," on page 12 of the rollout of InfiniBand, has two significant changes. First, HCAs will begin to be landed on the motherboard (or LOM) or be implemented as InfiniBand Server Blades. This will lower cost and should eliminate PCI from the motherboard. Secondly, the migration incorporates gateways between InfiniBand and Fibre Channel.

Figure 7. IB System Area Network : Phase 2



This can be implemented as either a gateway device or through the use of Target Channel Adopters (TCAs) within a RAID chassis backplane. We show in Figure 7 that RAID storage connects as part of the Fibre Channel SAN to InfiniBand with the gateway. It's important to note that where Fibre Channel HBAs existed (although not shown in Figure 6) can now be removed from the server reducing cost, power consumption and simplifying connectivity. Fault tolerant fabric connections are now being leveraged in the integrated InfiniBand connectivity of the area networks that communicate with WAN via gateways.

Figure 8. IB System Area Network: Phase 3



In the third and final stage, Figure 8, “IB System Area Network: Phase 3,” on page 13 a native InfiniBand system area network or single integrated fabric is achieved within the data center. Clustered servers can now be created from either existing servers with HCAs or LOM InfiniBand or from multiple InfiniBand blades or a combination of both (as shown). Native InfiniBand RAID arrays provide a direct connection to the InfiniBand system area network. In addition, connectivity to the Ethernet LAN is moved to the edge of the network through an InfiniBand to multi-purpose blade based WAN gateway. This gets all of the I/O out of the server box and completes the disaggregation of the network.

The consolidation of the system area network to InfiniBand allows low power, low profile 1U server or blade implementations. Servers become essentially CPU + Memory + InfiniBand, an extension of the decentralization of the system area network. All I/O is moved to the edge of the SAN. This allows each segment of the SAN, servers, storage, and I/O the flexibility of being independently upgraded, changed or serviced.

10.0 Summary

InfiniBand provides a powerful low overhead layered protocol designed to serve the needs of the Internet Data Center. InfiniBand offers a reliable, in-order, connection oriented transport service *implemented in hardware*. Further InfiniBand offers much higher performance and greater RAS than other networks as features such as quality of service, fault tolerance, channel de-multiplexing, and memory protection checks are implemented in hardware. Thus InfiniBand enables the implementation of an integrated, high performance, low latency, reliable, fault tolerant system area network supporting native quality of service. In addition to offering higher performance than existing network architectures, InfiniBand adds the RAS features required to implement mission critical computer/storage infrastructure. The critical change from a bus to a switch fabric allows InfiniBand to come “out of the box” enabling servers to support a native networked architecture. This native networking capability provides significant cost and power savings by allowing dense server clusters. This simple architectural change fundamentally changes the way servers can be packaged. A server becomes simply CPU(S), memory, and InfiniBand ports with I/O pushed to the edge of the network. Moving I/O to the edge of the network and allowing it to be shared by multiple servers completes the *disaggregation* of servers, storage, and I/O. This disaggregation offers performance, power, density, and fault granularity benefits to the data center.

A clear migration path to InfiniBand has been outlined, no forklift upgrade will be needed here. With over 250 members in the IBTA, industry momentum is clearly behind InfiniBand. This includes all of the major server and storage vendors having announced support for native InfiniBand connectivity. The cost, power, density, and RAS advantages fueled with industry momentum will result in a rapid adoption rate into the Internet Data Center enabling the re-integration of the System Area Network into a single, fast, and reliable fabric, namely InfiniBand.

11.0 About Mellanox

Mellanox is the leading supplier of InfiniBand semiconductors, providing Switches, Host Channel Adapters, and Target Channel Adapters to the server, communications, and data storage markets. In January 2001, Mellanox Technologies delivered the InfiniBridge™ MT21108, the first 1X/4X InfiniBand device to market, and is now shipping second generation InfiniScale silicon. The company has raised more than $33 million to date and has strong corporate and venture backing from Intel Capital, Raza Venture Management, Sequoia Capital, and US Venture Partners.

In May 2001, Mellanox was selected by the Red Herring Magazine as one of the 50 most important private companies in the world and to Computerworld Magazine Top 100 Emerging Companies for 2002. Mellanox currently has more than 200 employees in multiple sites worldwide. The company's business operations, sales, marketing, and customer support are headquartered in Santa Clara, CA; with the design, engineering, software, system validation, and quality and reliability operations based in Israel. For more information on Mellanox, visit www.mellanox.com.