InfiniBand Storage

The Right Solution at the Right Time

Figure 1. InfiniBand vs 2Gb/s Fibre Channel Single Port Storage Throughput to Disk Media

1.0 InfiniBand Storage - Past, Present, and Future

In early 2001 the first InfiniBand adapter cards and switch systems began to appear on the market. As an industry standard interconnect delivering 10Gb/sec performance with transport offload, many believed that InfiniBand technology would swiftly steal market share from Fibre Channel. However, as is typical with new technologies, adoption cycles proved slower than the original predictions - as inevitably silicon and software solutions take time to mature and drive down the path to delivering improved price/performance to drive market success. Furthermore, starting in late 2001, the technology market as a whole dipped into the deepest ever downturn experienced by the industry, and ultimately the storage segment was not spared. During this technology downturn, pundits began to predict the demise of InfiniBand.
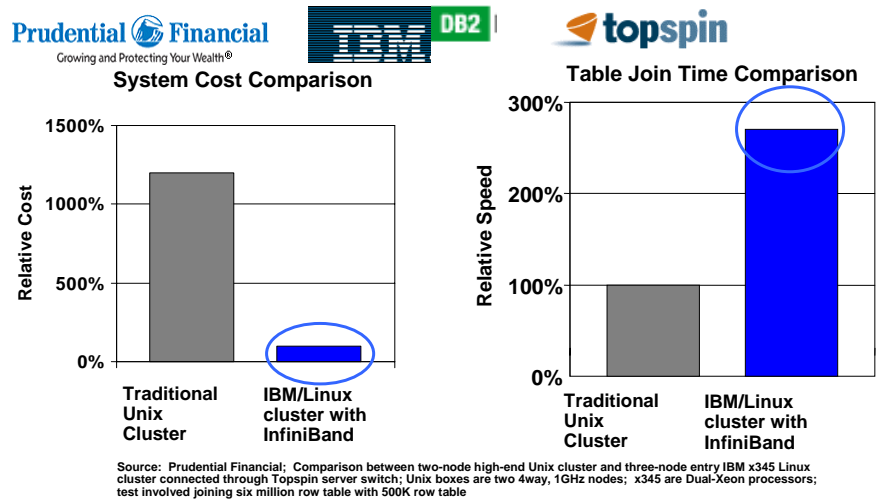
Fortunately the InfiniBand community continued to develop new products with improved features and performance and developed robust and full featured software. Ultimately this led to the introduction of powerful system products, adoption by all of InfiniBand by all of the major server OEMs, customer deployments, and real world price performance proof points. Furthermore, InfiniBand is being widely integrated by server OEMs in "Landed On Motherboard" and InfiniBand Server Blade solutions - further driving down prices and simplifying clustering and I/O aggregation. These successes have driven the widespread acceptance of InfiniBand in the enterprise market for clustered database, financial compute clusters, technical computing, and high performance compute clusters. Success in these markets has in turn enabled the InfiniBand vendors to capture key beach heads in the enterprise data center and to generate significant revenues allowing them to continue to invest in developing new products to drive the price/performance advantages into the future. In light of these market successes it is time once again to consider the place of InfiniBand in the data center specifically with regards to storage.

Mellanox Technologies Inc.   2900 Stender Way, Santa Clara, CA 95054   Tel: 408-970-3400   Fax: 408-970-3403   www.mellanox.com   1

Mellanox Technologies Inc

Document Number
2284WP

Figure 2. All Major Server Vendors Offering InfiniBand Solutions

2.0 The Clustering Beach Head

Ultimately, the success of InfiniBand depended on penetrating the market to cluster industry standard servers together to build very powerful computers at a fraction of the cost of traditional symmetrical multi-processing (SMP) machines. InfiniBand has succeeded exceptionally well to secure this beach head market and customers regularly report achieving 2X performance at 1/10th the price of traditional "big box" computing solutions. For example, an IBM whitepaper (http://www.pc.ibm.com/ww/eserver/xseries/linux_update/pubs/misc/topspin.pdf) demonstrates the price/performance benefits Prudential Financial recognized with InfiniBand clusters. The price improvement was expected, however what surprised Prudential

Source: Prudential Financial; Comparison between two-node high-end Unix cluster and three-node entry IBM x345 Linux cluster connected through Topspin server switch; Unix boxes are two 4way, 1GHz nodes; x345 are Dual-Xeon processors; test involved joining six million row table with 500K row table

1/8th the Cost   2.5X Performance Boost

Figure 3. Cost Performance Benefits

was the significant performance gains that were achieved despite the much lower cost of the InfiniBand cluster.

Delivering these significant price/performance benefits has resulted in major server vendors offering InfiniBand clusters for both performance business computing and technical computing. Essentially all the major server vendors now offer

InfiniBand as part of their industry standard server offering, as well as for connecting larger midrange and highend servers into even more powerful InfiniBand clustered solutions.

This success in the server market in turn is driving the demand for native InfiniBand attached storage and, as shown later in this white paper, results in magnifying the already significant price/performance advantage the technology has over Fibre Channel.

2.1 InfiniBand Clusters Scale Out Performance Business Computing

InfiniBand has proven the ideal interconnect to scale out industry standard servers to tackle business computing problems that demand high levels of performance, reliability, and cost-effectiveness. A "scale out" solution simply refers to connecting multiple industry standard servers with a high performance interconnect fabric and associated software to match the performance and problem solving capabilities of traditional mainframe or super computers. This is opposed to the traditional "scale up" solution which utilizes proprietary processors and interconnects to build monolithic "big box" computing platforms. By leveraging the low latency, 10Gb/s InfiniBand interconnect, clusters of industry standard servers can scale out to match the performance of large SMP machines at a fraction of the cost. This is clearly illustrated by comparing the cost of a typical 32 way SMP system vs. a 32 Processor InfiniBand cluster based on industry standard servers. InfiniBand delivers comparable performance at 1/10th the cost of the SMP platform.

An additional benefit of scaling out is that with an InfiniBand cluster you can grow the performance business computing capacity organically to meet the requirements of the business application as it develops. By contrast, with an SMP box it is required to anticipate the business computing requirements and project these requirement forward for the useful life of the computer - and buy a platform capable of handling these future requirements. Now predicting the future is hard, and in the best case, up-front provisioning means initially over-provisioning and thus higher up front capital expenses. Furthermore, in the worst case, business performance computing requirements can greatly exceed the initial projections. This can result in a downstream loss of business and require a complete fork lift upgrade to increase the compute capacity to meet the unanticipated business needs.

For all of these reasons Infini-Band scale out solutions are being widely adopted across a broad range of performance business applications. For example, clustered databases such as Oracle 10g and IBM DB2 have recognized the price/performance and scalability gains that can be recognized by scale out computing with Infini-Band. A whitepaper developed by Oracle has documented the performance gains achieved with InfiniBand clusters of industry standard servers (http://www.oracle.com/technology//deploy/availability/pdf/oracle_IB.pdf).

Proprietary Scale-Up:
"Buy a Bigger Box"

InfiniBand "Scale-Out":
Add industry standard servers as needed

32-Way SMP Server	32-CPU InfiniBand Cluster
80 - 200 Gflops Peak Rating	195 Gflops Peak Rating
Average Cost = $1,140,000*	Total Cost = < $100,000*

Figure 4. Proprietary and InfiniBand Alternatives

3.0   How does a Server Interconnect Relate to Storage?

Clearly, InfiniBand has established market leadership as the server interconnect of choice for performance business computing and high performance technical computing - but what does this have to do with InfiniBand's use as a storage interconnect? In a word: *EVERYTHING*.

This is because storage and servers go hand in hand. Basically you can do only three things with data:

1. Modify it

2. Move it

3. Store it

The first item, modification, deals with the processing and transformation of data - and changing the data is invariably the responsibility of the compute platform. As explained in the previous section, InfiniBand has emerged as the best way to build truly scalable compute platforms - by scaling out industry standard servers in an InfiniBand cluster.

The second requirement is to move data and this clearly is the role of the I/O interconnect. Data movement takes two useful forms:

- Movement between server <=> server
- Movement between server <=> storage

Again, with features such as 10Gb/sec performance, low latency, remote direct memory access (RDMA), and kernel bypass - InfiniBand has established itself as the standard for server to server communications. Interestingly, the same attributes that have enabled InfiniBand to succeed as the industry standard for server to server communications, also benefit server to storage communications. Even more importantly, *InfiniBand has the bandwidth, features, and scalability to enable* **the same interconnect** *to serve both purposes.*

Finally the third thing you can do is store the data which is the responsibility of the storage platforms. The real value here is inside the storage box - features like striping, fault tolerance, virtualization, backup, and mirroring. Interestingly, these "inside-the-storage-box" features have more to do with modifying and moving the data than just storing it. Thus it is no surprise that first tier storage vendors have adopted InfiniBand as the inside the box interconnect to perform these functions. Obviously for storage platforms that already have InfiniBand inside the box it is relatively straightforward to expose InfiniBand as a native storage interconnect.

3.1 InfiniBand Server Clusters Deployment is Driving Demand for Native InfiniBand Storage

InfiniBand success as a server to server interconnect is one of the key driving forces behind the renewed demand for native InfiniBand storage. To understand why this is so it's useful to look at a new server installation that chooses to deploy an InfiniBand cluster solution rather than a traditional SMP machine. Now clearly, as was shown in Section 2.1 on page 3, the customer will recognize an immediate 10X reduction in cost by choosing InfiniBand. Now let's consider the storage decision.

If the customer already has a Fibre Channel SAN storage solution deployed than there are essentially two options:
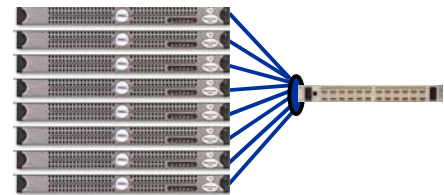
i- Add a Fibre Channel host channel adapter to each of the clusters in the fabric to connect to the FC SAN

ii- Use a transparent InfiniBand gateway to connect the InfiniBand cluster to the FC SAN

Option ii. is much more cost effective since the InfiniBand cluster already includes a switch which provides connectivity to each server node. Using an InfiniBand to Fibre Gateway gateway the same InfiniBand infrastructure can be leveraged to provide storage connectivity. Furthermore it is only necessary to provision as many FC gateway ports as required by the storage platform. Best of all the high speed InfiniBand links provide plenty of performance overhead and can be made transparent to existing Fibre Channel storage management software.

The capital cost benefits get even better if new storage infrastructure is being installed where native InfiniBand storage is an option. Here the InfiniBand server cluster interconnect is leveraged for the storage area network. This dual use of the InfiniBand fabric is referred to as a *system area network* to reflect the greater functionality of the interconnect. In this case however the requirement for gateway ports is completely eliminated. This further reduces the cost and simplicity of the overall system architecture.

16 Node 10Gb/s InfiniBand Server Cluster

Let's look at a real case to make the comparison more concrete. Consider a 16 node InfiniBand cluster which uses 16 InfiniBand host channel adapters (HCA) for each server and a 24 node switch to connect them. Assuming OEM pricing of around $500 per InfiniBand HCA and around $300 per InfiniBand 10Gb/s switch port, the total cost of the InfiniBand system fabric interconnect is then around $15,500.
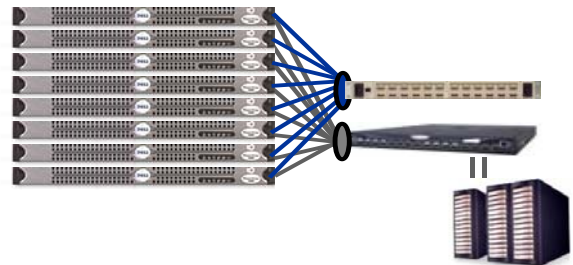
InfiniBand Cluster Interconnect
16 InfiniBand HCA's: $8,000
24 Port InfiniBand Switch: $7500
Total: $15,500

Cluster with Fibre Channel Storage

Now lets add storage to the picture. If we add Fibre Channel storage we need to add a new Fibre Channel switch to provide the storage connectivity. The Fibre Channel server connectivity adds an additional cost $500 per HBA for each server. In addition, assume that the target storage platform requires four Fibre Channel host bus adapters. Despite delivering only 2Gb/s the Fibre Channel switch ports are priced over $700 per port. Thus adding Fibre Channel storage connectivity to the cluster adds over $27,000 to the system cost and results in a total system area network cost of over $43,000.

Total Fibre Channel Cost: $43,044

Unified InfiniBand Cluster

Finally, let's see what happens when we use native InfiniBand storage instead of Fibre Channel. In this case each server already has connectivity via the InfiniBand HCA to the system area network. The incremental cost per server is effectively *zero*. Similarly the same switch that is used for clustering is used for storage and thus incrementally adds *zero* cost. The only additional cost then is the four InfiniBand host channel adapters within the storage targets. Thus the incremental cost of adding InfiniBand storage connectivity is only $2000 vs over $27,000 for Fibre Channel! In fact the cost of the storage target adapters are hidden within the storage platform costs and thus in essence *InfiniBand storage connectivity is free!* The total cost of the unified InfiniBand system area network is then only $17,500 vs over $43,000 vs Fibre Channel.
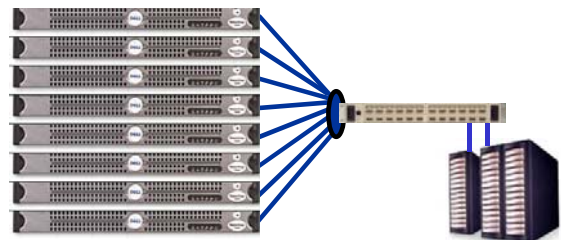
Total InfiniBand Cost: $17,500

This data is summarized in the tables below:

Cluster Interconnect	Item	Qty	Price	COST
Server Cluster Adapter	IB HCA	16	$500	$8,000
Cluster Switch	IB Switch Port	24	$313	$7,500
Total Cluster Cost				$15,500

FC Storage Interconnect	Item	Qty	Price	COST
FC Server Storage Adapter	FC HBA	16	$500	$8,000
FC Storage Target Adapter	FC HBA	4	$500	$2,000
Storage Switch	FC Switch Port	24	$731	$17,544
Total Cluster Cost				$27,544

Cluster Interconnect	Item	Qty	Price	COST
IB Server Storage Adapter	IB HCA	0	$500	$0
IB Storage Target Adapter	IB Switch Port	4	$500	$2,000
Storage Switch	IB Switch Port	0	$300	$0
Total Cluster Cost				$2,000

Total System Area Network Cost	
InfiniBand	**FibreChannel**
$17,500	**$43,044**

Indeed, the deployment of InfiniBand server clusters makes a compelling business case and is driving the demand for native InfiniBand storage.

4.0 But Haven't We Been Here Before?

The potential for InfiniBand as a system area network has long been recognized and the technology was launched with much fanfare as a server *and* storage interconnect. But clearly, despite initial adoption by storage vendors, InfiniBand has been slow to take off in the storage market and, to date, has primarily seen success as in internal storage interconnect. But *past is prologue* and it is important to recognize what has changed in the market that now enables InfiniBand to succeed in the storage market.

First of all it is important to recognize what has *not* changed. InfiniBand is still the only technology with the big four:

- Industry Standard
- 10Gb/sec Performance
- Remote Direct Memory Access (RDMA)
- Hardware Transport

Secondly InfiniBand offers significantly better price/performance value than Fibre Channel. Thirdly the InfiniBand community is executing on a roadmap over the next two years that extends the performance lead from 10Gb/sec to 20Gb/sec in 2005 and to over 100 Gb/s in 2006 - well beyond the 8-10 Gb/s roadmap of Fibre Channel during the same timeframe. Fourthly InfiniBand unifies the data center delivering a converged communications, clustering, and storage interconnect; thereby simplifying management and lowering both capital and operational expenses.

So what has changed in the last six to nine months that has dramatically lifted the prospects of InfiniBand in the enterprise storage market? First and foremost is the growing success of InfiniBand - with all of the major server vendors delivering solutions to a growing customer base demanding InfiniBand scale out clusters. The successful deployment of InfiniBand into the beach head server clustering market has always been a necessary pre-requisite to success in the storage market. Those who failed to recognize this and focused on storage or communications target channel adapters - before the server host channel adapter market was secured - have not survived. However now that there are growing deployments of Infini-Band server clusters - demand for native InfiniBand storage is growing rapidly.

A second change compared to a few years ago is related to the InfiniBand software base which is now mature, widely available and supports a broad range of applications. OpenIB.org (www.openib.org) is an open source development project offering a complete software stack for InfiniBand. OpenIB offers a robust and complete set of drivers including key middleware supporting storage - such as SCSI RDMA Protocol (SRP) for block storage applications, kDAPL for NFS over RDMA, and SDP for iSCSI/iSER and other clustered files systems. The availability of comprehensive, off-the-shelf, and production ready storage drivers is in stark contrast to the relative immaturity of the InfiniBand software stack available even nine months ago. The availability of storage software for InfiniBand means that storage platform OEM's need only focus on integrating their value added features on top of a robust base that uses the same FC, SCSI, and NFS API's thereby making transparent application porting straightforward.

A third key change is that InfiniBand has grown beyond the initial beach head high performance computing market segment to encompass a broad range of performance business computing applications including clustered databases, manufacturing, financial, and other performance data center applications.

A final change is the availability of fully integrated storage development platforms offering 10X better price performance than competitive Fibre Channel. The MTD1000 native InfiniBand storage platform is a good example of such a development platform that delivers 4X better performance and as little as 40% the cost of leading Fibre Channel platforms.

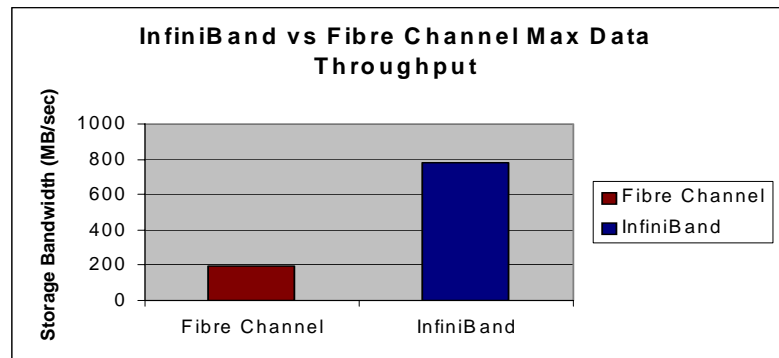# 5.0 Native InfiniBand Storage Platforms

Figure 5. SC933T Native InfiniBand Storage Platform and Measured Performance Advantage vs Fibre Channel

At Storage Networking World 2004 Mellanox introduced the first ever native InfiniBand storage platform. Based on industry standard components, the native InfiniBand storage platform delivers nearly 800MB/sec throughput to the disk. Such storage platforms can greatly accelerate a storage OEM's product development cycle and time to market. Also, these storage platforms are ideal for system integrators, value added resellers, and OEM's looking to quickly gain a strategic foothold in the InfiniBand storage market. For more in depth technical details on the price/performance benefits of InfiniBand

storage platforms see the whit paper: *InfiniBand Storage Platforms - Superior Price/Performance in an Enterprise Ready Platform.*

6.0 Key Application Drivers of InfiniBand Storage

Success in the server clustering market is driving the demand for InfiniBand storage - indeed just as the storage market is beginning to pick up. But there are also many performance sensitive applications independent of clustering, which are driving demand for InfiniBand storage:

* Backup / Diskless Backup

* Mirroring/Snapshot/Checkpointing

* Streaming Video / Graphics

* Clustered Storage for Disaster Recovery

* Data warehousing

These applications benefit directly from the improved price/performance offered by InfiniBand solutions - independent of whether used with an InfiniBand cluster or other non-clustered computing solution.

7.0 Summary

InfiniBand is growing rapidly in the beach head markets of high performance computing and performance business computing in the data center. All of the major server vendors are offering InfiniBand solutions to the market and there is rapid adoption and deployment of the technology across a broad range of applications. This success in turn is driving demand for InfiniBand storage. As an industry standard delivering 10Gb/s performance and transport offload, InfiniBand offers significant price performance advantages over Fibre Channel. These price/performance advantages are magnified in a clustered environment where the InfiniBand cluster network can be utilized for storage as well. Major advancements in InfiniBand software, platform, management, and application support for storage has accelerated the adoption of InfiniBand in the storage market. The demand for native attached InfiniBand storage is being further driven by applications demanding higher performance, lower costs, and better integration. Indeed servers and storage go hand in hand - and with the demand for native InfiniBand storage growing rapidly - InfiniBand is leading the way to a converged data center.