



InfiniBand in the Enterprise Data Center

InfiniBand offers a compelling value proposition to IT managers who value data center agility and lowest total cost of ownership

Introduction

InfiniBand is a high-speed server-interconnect technology that is specified and maintained by the InfiniBand Trade Association (IBTA). Through use of Host Channel Adapters (HCA) and Switches, InfiniBand Technology is used to connect servers with remote storage and networking devices, and other servers. It can also be used inside servers for inter-processor communication (IPC) in parallel clusters. SDR and DDR versions of HCAs and Switches available in the market today can reach 10Gb/s and 20Gb/s data rates with end-to-end latency of lower than 3 microseconds.

Today's data centers need an agile infrastructure that incorporates ongoing improvements in computer, storage, networking, and application technologies, and empowers IT to support changing business processes. InfiniBand fabric solutions enable IT organizations to turn computing and storage resources from monolithic systems to service-centric shared pool of resources consisting of standardized components that can be dynamically provisioned and accessed through an intelligent network.

InfiniBand in n-Tier Data Center Architectures

Server farms in the data center spans across resources that fall in the Aggregation Layer, Access Layer and Storage Layer (see figure 1). The Access Layer architecture has progressed from client-server models to multi-tier or n-tier service models for improved scalability and maintainability. Multi-tier service models force separation of user interface software in the Front End layer from business, file-system and database logic or middleware in the Application layer; and Application layer from the back-end storage (database or file-system) servers in the Back End layer. Finally, connecting to the Back End layer is the Storage layer comprising of storage systems like storage switches, NAS targets etc. In this model, Front End, Application, Back End and Storage layers could sit in their own islands, connected by different networks.

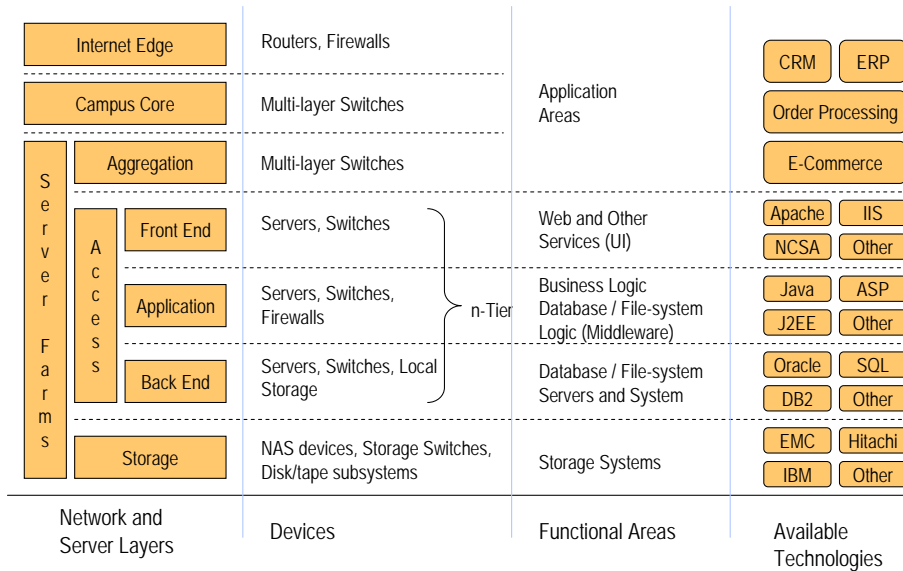


Figure 1: Data Center Architecture Layer and Storage Layer (source: Mellanox)

Different Networking Requirements

Network interface, bandwidth and latency requirements for connectivity between the n-tier components differ. For example, Front End to Application may require a lower bandwidth network



with IP or sockets based network interfacing, while Application to Back End typically requires a TCP streams based high bandwidth, low latency network. Similarly, IPC performance requirements within the Application and Back End components are critical for meeting end user service level agreements (SLA). As such, various networking and interconnect technologies have been deployed in different sections of the n-Tiered architecture.

Role of InfiniBand

Where performance and SLA cannot be compromised, InfiniBand is ideally suited as the interconnect technology for Access Layer and Storage components – specifically for Application and Back End IPC applications, for connectivity between Application and Back End layers, and from Back End to Storage Layers. There are five distinct areas where InfiniBand shines with respect to other interconnect technologies:

1. Deterministic performance through use of a loss-less fabric, Constant Bisectonal Bandwidth (CBB, also called fat-tree) architecture for scalability, and hardware based end-to-end congestion control.
2. Highest available bandwidth with facilities for granular bandwidth allocation and guarantees
3. Support for high-speed and zero-copy RDMA protocols and required interfaces for IPC and TCP sockets based communication. Sockets Direct Protocol (SDP) implementations over InfiniBand offers compelling bandwidth and CPU utilization benefits compared to other interconnect or networking technologies (see figure below for bandwidth advantages).

NetPIPE TCP BW (Uni-directional)

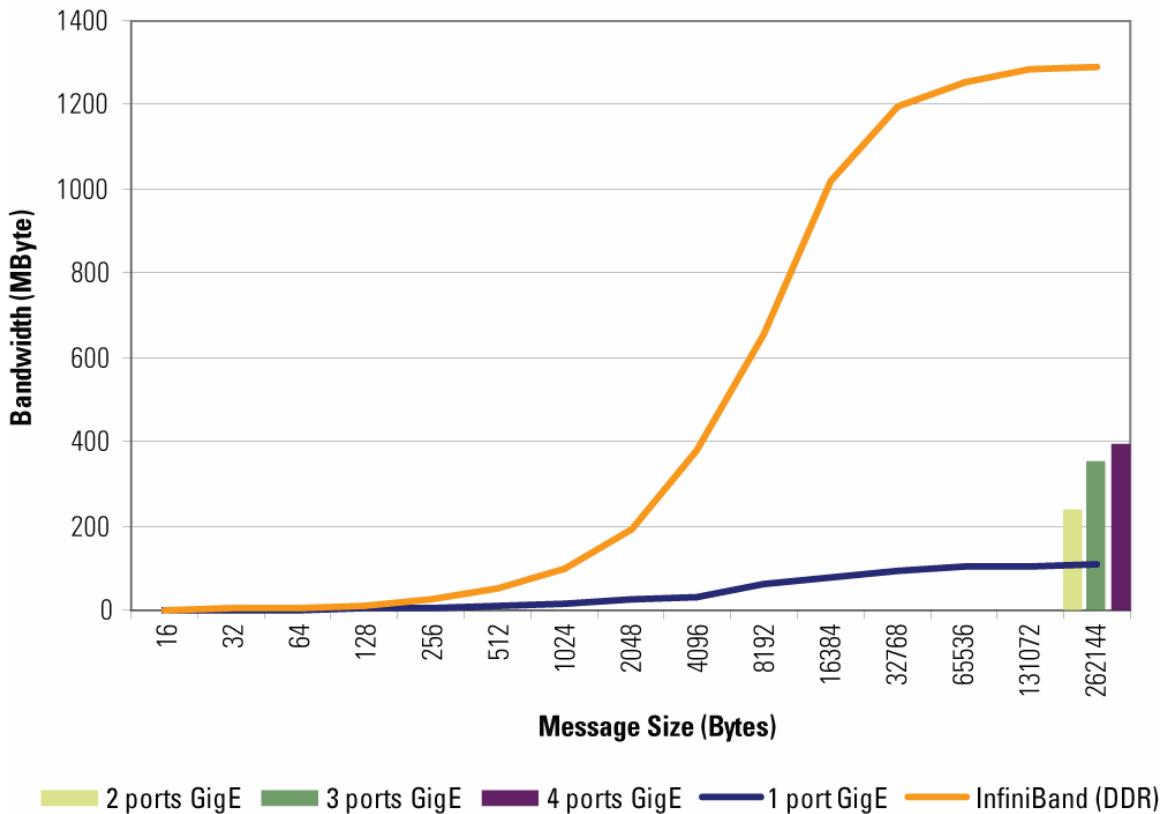


Figure 2: Sockets Direct Protocol over InfiniBand Bandwidth (source: Mellanox)

4. End user and per-port cost for InfiniBand compared to other comparable technologies is at least ten times lower
5. Growing ecosystem of suppliers that meet the networking needs of the n-Tiered service model:
 - o Access Layer and Aggregation Layer server switches are available from Cisco Systems, Flextronics, SilverStorm and Voltaire.
 - o High performance native InfiniBand solutions for connectivity between Back End and Storage layers are available from Data Direct Networks, Isilon, Engenio/LSI, SGI and others with companies like FalconStor and PolyServ providing management and storage services software.
 - o Gateway solutions for connectivity between storage servers and legacy Fibre Channel based NAS solutions from Cisco, SilverStorm and Voltaire.
 - o High performance database storage solutions from IBM DB2 and Oracle

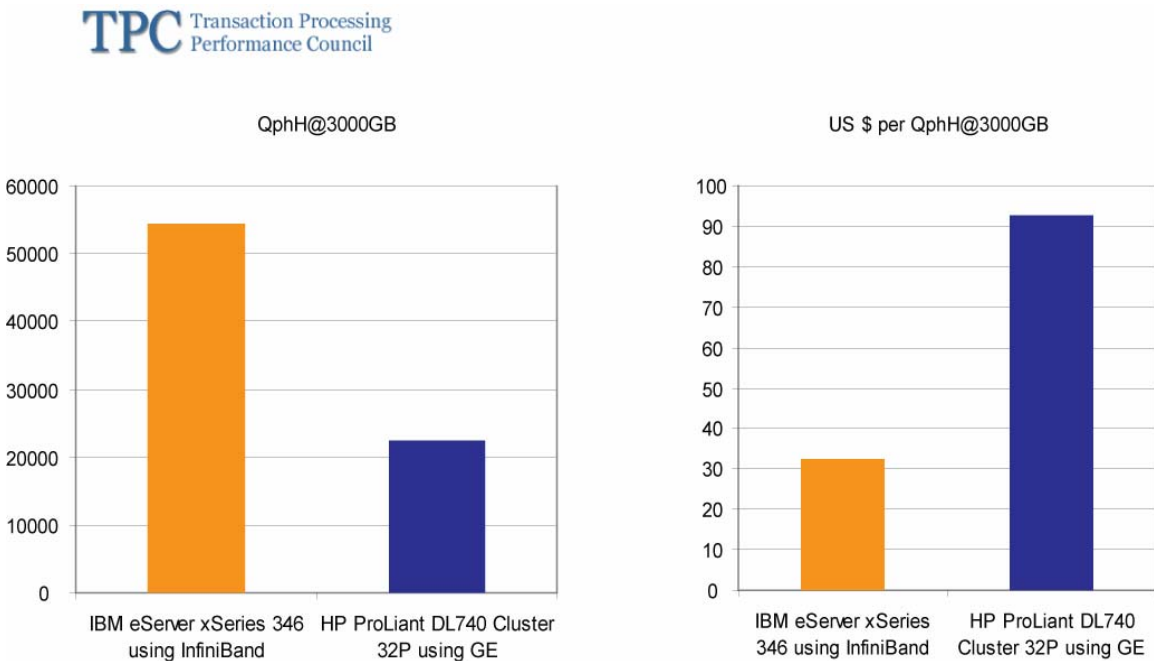


Figure 3 (a) IBM DB2 with InfiniBand benefits (source http://www.tpc.org/tpch/results/tpch_price_perf_results.asp?resulttype=cluster)

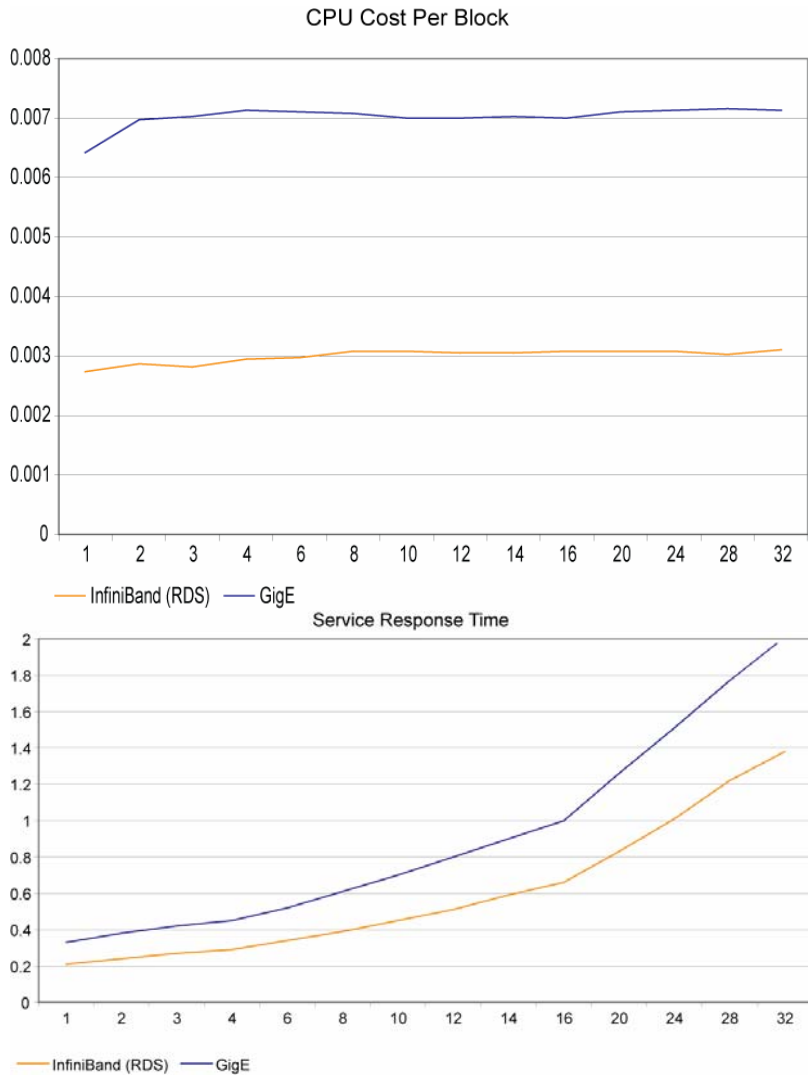


Figure 3 (b) Oracle 10G RAC with InfiniBand benefits (source: SilverStorm / Oracle)

- High performance file-system middleware and back-end server solutions from Cluster File Systems Inc. (Lustre™), [HP StorageWorks Scalable File Share \(HP SFS\)](#), Terrascale Technologies (TerraGrid software and Storage Brick hardware platform, Panasas (ActiveScale Storage Cluster™) and others)
- High performance middleware business logic solutions will be available in the future.

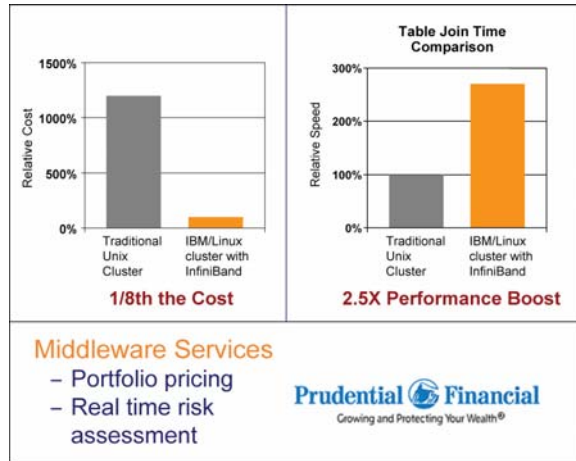
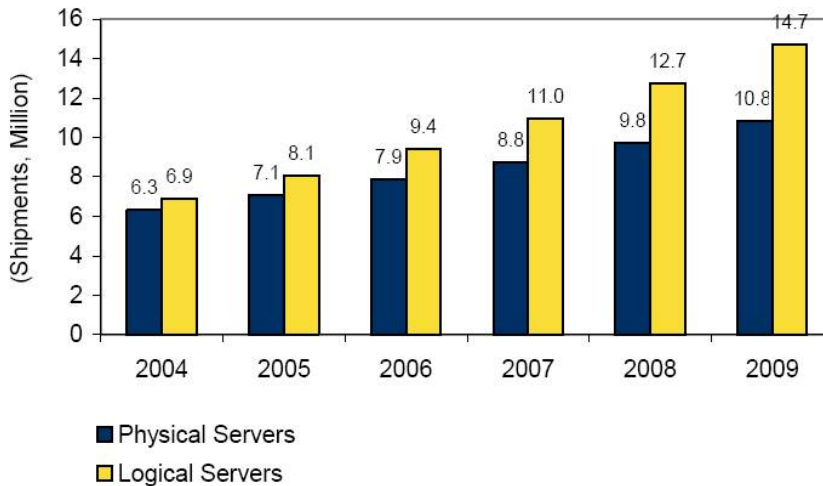


Figure 4: Prudential Financial Portfolio Analysis / Risk Assessment (source: <http://www-03.ibm.com/servers/eserver/linux/xseries/pdf/topspin.pdf>)

InfiniBand in Grid Computing Architectures – Enabling the Service-centric Model

Operational inefficiencies and low resource utilization with existing architectures (including n-tier) are causing IT managers to look for consolidation and virtualization technologies. The trend is toward using virtual or logical servers (see figure 5) for creating a common shared pool of resources, rather than dedicated pools for web servers, middleware or application servers and storage servers. IT managers are realizing that only such a shared model can result in an effective service-centric service model.

Worldwide Physical and Logical Server Shipments by Year



Source: IDC, 2005

Logical servers are created through the use of virtualization technologies where a single server node is partitioned into two or more virtual servers. (Source: IDC)

Figure 5: Worldwide Physical and Logical Server Shipments by Year

Fading Network Boundaries

Use of virtualized servers is a key enabler of the shared resource pooling and service-centric service model. With server resources in the data center being shared dynamically across the components of the traditional n-tiered model, the boundaries between the n-tiers cease to exist.

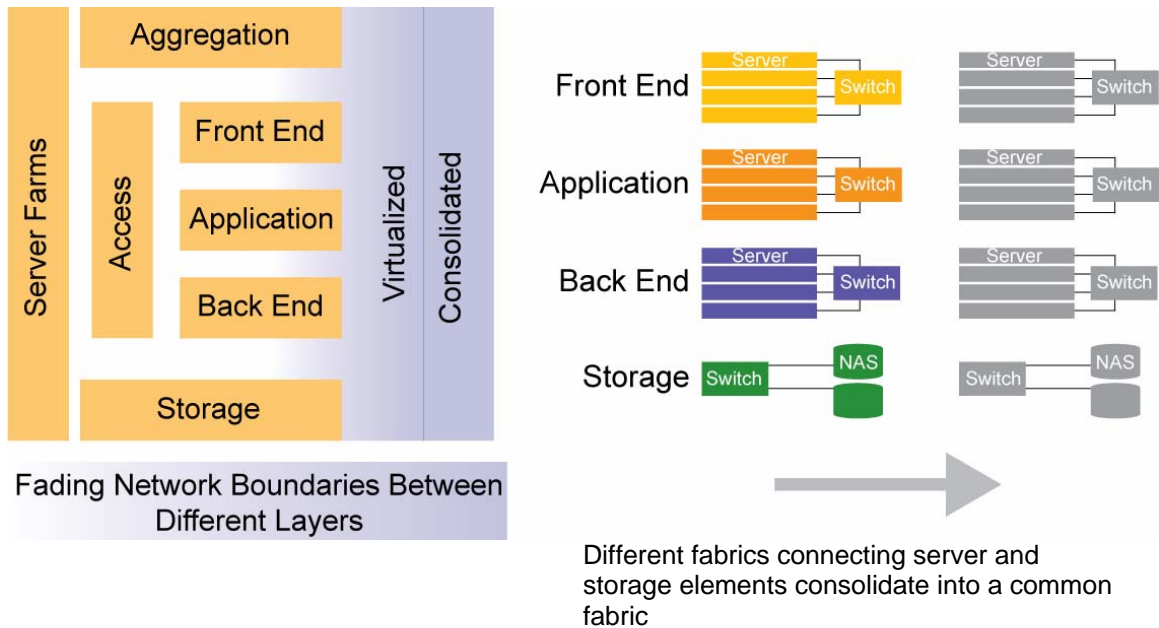


Figure 6: Fading network boundaries leading to consolidation of network elements

With the network boundaries gone, the viability and need for having different networks and bandwidth requirements connecting those boundaries go away too. Each physical server now needs to be equipped with adequate networking and storage I/O resources so that it can take any role in the n-tier model. And the common physical I/O resources-set required across all servers in the n-tier data center becomes the one that provides the SLA and performance requirements for the most demanding applications in the data center. For example, if three Gigabit Ethernet cards, one Fast Ethernet card, and two Fibre Channel HBAs are required, then all servers need to be populated with all of the above I/O components, increasing cost and complexity significantly. Consolidation over one fabric that meets the most demanding I/O requirements becomes a must!

InfiniBand solves the Consolidation Challenges

InfiniBand is the right architecture for connecting all elements of the virtualized and shared resource pool, thereby enabling the service-centric service model in the datacenter better than any other interconnect available today. This is why:

- InfiniBand supports channels-based IO (see figure 7 below) enabling consolidation of all data center connectivity functions for IPC, networking and storage functions.

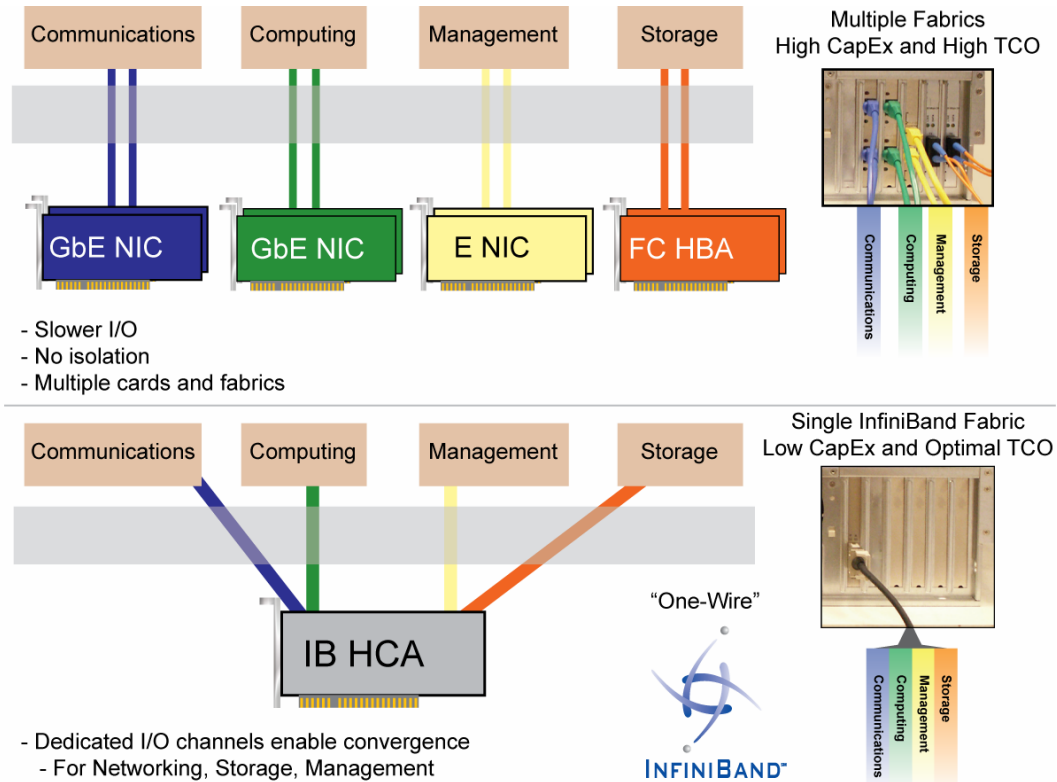


Figure 7: Use of InfiniBand I/O channels to consolidate over one wire

- Available OpenIB.org based software stacks from InfiniBand and OS vendor distributions provide the desired application interfaces to enable the consolidation
- InfiniBand architecture supports bandwidth allocation and guarantees per I/O channel that can be linked individually to all data center applications, middleware, storage and networking services. This further enables effective consolidation of IO in the data center.
- The channels-based IO architecture lends itself very well to virtual server environments where multiple virtual machines on a single physical server require to be served by multiple end points-based physical I/O devices. Since the end-points (i.e., IO channels) in an InfiniBand HCA can support networking, storage, IPC and other functions, they also serve as intelligent I/O consolidation points. Guest OS-transparent (see figure 8 below) solutions can enable off-the-shelf OS, management software and applications to leverage higher bandwidth InfiniBand I/O and consolidation without knowing that they are actually running over InfiniBand.

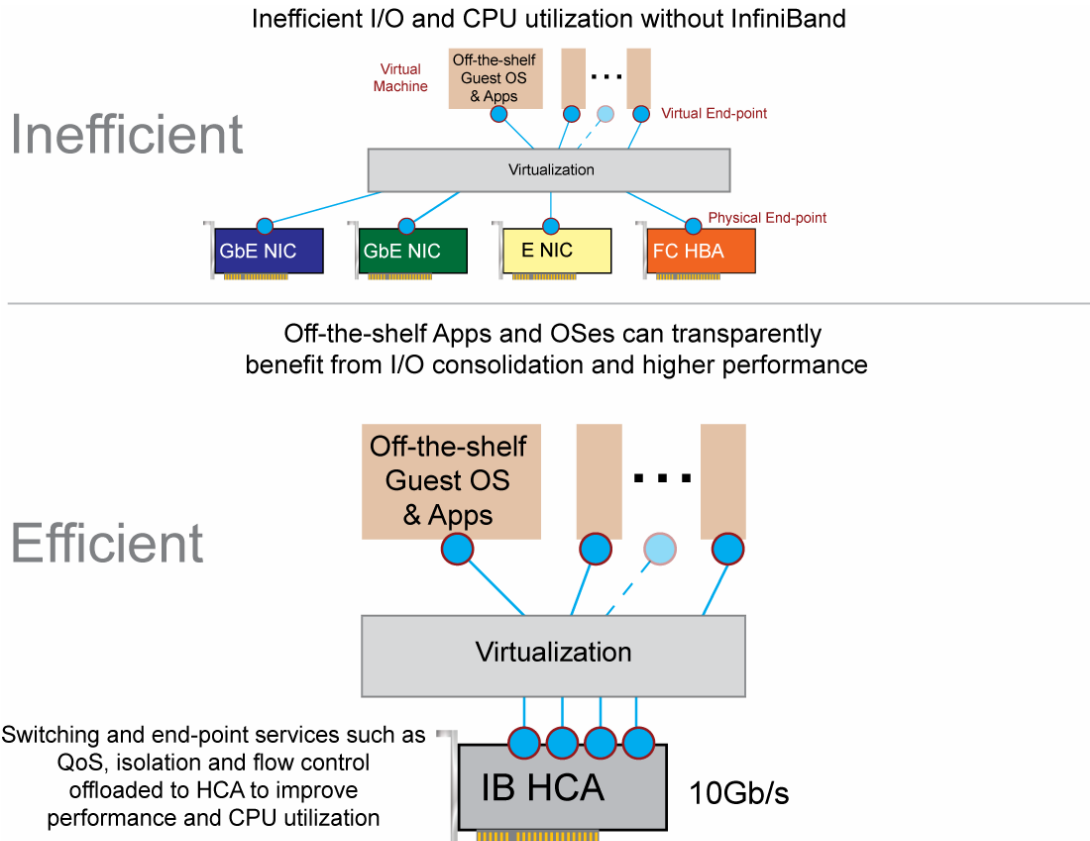
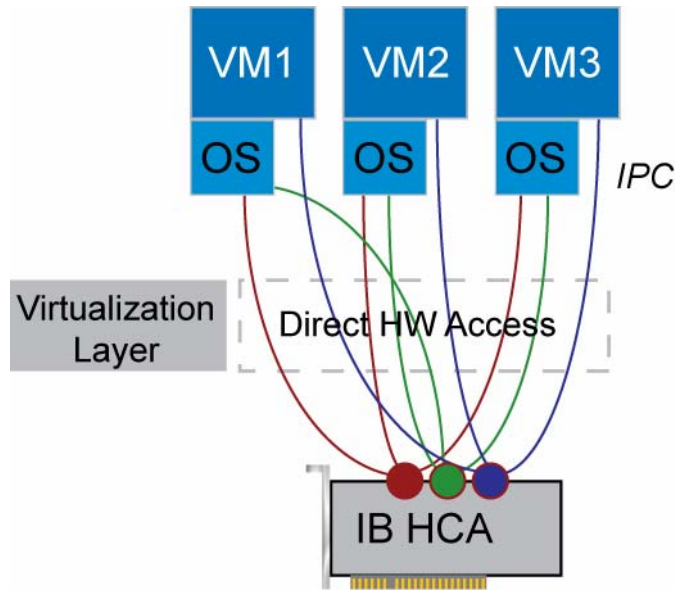


Figure 8: I/O consolidation with virtualization – enable transparent deployment of off-the-shelf applications

- The channels-based I/O architecture along with transport, RDMA, QoS, congestion control and checksum functions implemented in InfiniBand hardware complements strong trends toward pass-through and paravirtualized virtual machine architectures. Full-offload of hypervisor functions frees CPU cycles for processing application and middleware services and can increase resource utilization by enabling more virtual servers per physical server resource (see figure 9 below). RDMA and InfiniBand aware virtual machines or guest operating systems can further scale out leveraging full pass-through implementations possible with Xen and other virtual machine based architectures.



- Direct HW access for I/O
- One 10Gb/s card for Network, Storage, and IPC
- Scale out using clustering (IPC)

Figure 9: Pass-through virtualization with InfiniBand

Virtualization Solutions over InfiniBand

VmWare, Xen and Microsoft Virtual Server 2005 R2 based virtualization solutions over InfiniBand are currently under development. Mellanox and other InfiniBand vendors are developing software within the VmWare Community Source Program to add support for InfiniBand based networking, storage and virtual machine migration features in the VmWare ESX Server platform. Cisco offers its VFrame software that enable virtualization and lowers TCO. Novell/SUSE recently announced virtualization capabilities with its SUSE Linux server OS through inclusion of Virtual Iron's paravirtualized solution that supports InfiniBand.

10Gb/s InfiniBand Adapter at a price that matches Enterprise Gigabit Ethernet Pricing

With \$125 volume OEM-pricing available for Mellanox InfiniBand adapters, the barrier to consolidation is completely removed. IT managers can now consolidate to "one wire", and at the same time enjoy significant savings on hardware and manageability costs, and reduce power requirements across the data center (see figure 10 below):

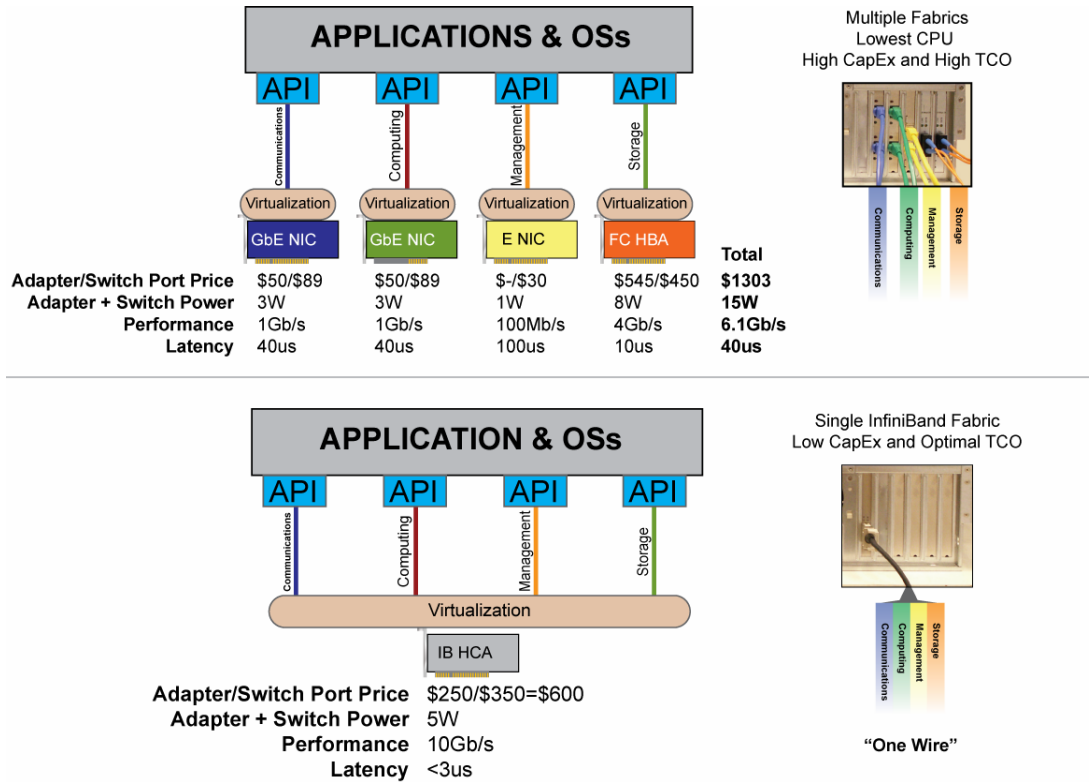


Figure 10: Compelling cost, power and performance benefits using InfiniBand HCA

* Prices are for end-users, Ethernet pricing from Dell'Oro Jul05, FC pricing from IDCDec05

Conclusion

The need for agility in data center resource deployment is a key requirement today. Sharing of the physical resource pool using virtual server technologies requires all physical servers to meet SLA requirements of the most demanding applications and at the same time reduce total cost of ownership. High bandwidth, low latency and I/O consolidation benefits provided by InfiniBand technology enable data centers to cost-effectively consolidate over one fabric, and at the same time meet the needs of the most demanding applications. Off-the-shelf data center applications can transparently leverage those performance and consolidation benefits using virtual server technologies over InfiniBand. InfiniBand HCAs and software can support multiple, intelligent endpoints that can provide dedicated and granular QoS and security services to virtual servers, as well as individual networking, storage, IPC and management applications, eliminating the need for deploying different fabrics for different services. Finally, aggressively-priced Mellanox InfiniBand adapters that support all of the above capabilities offer a compelling value proposition to IT managers who value data center agility and lowest total cost of ownership.



Links to other relevant white papers and presentations:

1. QoS and Congestion Control:
http://www.mellanox.com/pdf/whitepapers/deploying_qos_wp_10_19_2005.pdf
2. Scaling 10Gb/s Clustering at Wire-Speed:
<http://www.mellanox.com/pdf/whitepapers/scaling10gbsclusters.pdf>
3. SDP:
http://www.mellanox.com/pdf/whitepapers/RAIT_SDP%20Zcopy.pdf
4. Implementing Xen over InfiniBand:
http://www.mellanox.com/pdf/presentations/xs0106_infiniband.pdf