Cut I/O Power and Cost while Boosting Blade Server Performance

1.0 Shifting Data Center Cost Structures

Data center cost structures are shifting, from equipment-based to space, power and personnel-based. Total cost of ownership (TCO) and green initiatives are major spending drivers as data centers are being refreshed. Topping off those challenges is the rapid evolution of information technology as a competitive advantage through business process management and deployment of service-oriented architectures. Blade server and I/O technologies play a key role in meeting many of the spending drivers — provisioning capacity for future growth, efficient scaling of compute, LAN and SAN capacity, and reduction of space and power in the data center — all while reducing TCO and enhancing data center agility.

1.1 The Need for More I/O Capacity

While most blade servers are equipped today with Gigabit Ethernet or 1 Gb/s bandwidth, four key trends are quickly driving the need for more I/O bandwidth per server, typically 10Gb/s or higher:

- **Adoption of multi-core CPUs in servers:** Quad and eight-core CPUs are becoming common in servers, allowing more applications to run on each server, thus driving the need for 40Gb/s I/O bandwidth per server. Being able to allocate I/O bandwidth and other services to applications is critical to enable applications to deliver required services for meeting BPM and SOA goals.

- **Adoption of SAN:** Storage area networking requires storage access and services to be delivered over the I/O adaptor on the server. Popular database servers and server virtualization software utilize SANs to deliver value-added and cost-effective services. SANs demand high levels of I/O services, e.g., low tolerance to data loss and high-bandwidth.

- **Server Virtualization:** When servers are virtualized, multiple virtual machines and

guest operating systems execute on each server. This enhances data center agility. High-bandwidth I/O in excess of 40Gb/s is required to meet the needs of LAN and SAN capacity shared across multiple virtual machines as well as for virtual machine and storage or file system mobility related value-added functions provided by server virtualization software providers.

- **Sharing of server resources:** When providing cloud computing services, servers can't be dedicated for specific applications, e.g., front-end web services, middleware applications, database and storage servers etc. These server silos have typically required varied levels of I/O capacity, with the storage server tiers demanding the most capacity and the front-end web servers requiring the least. With server virtualization and virtual machine mobility, servers are shared across all application tiers. I/O capacity provisioned per server is dictated by the most demanding apps which may reside in one of the servers.
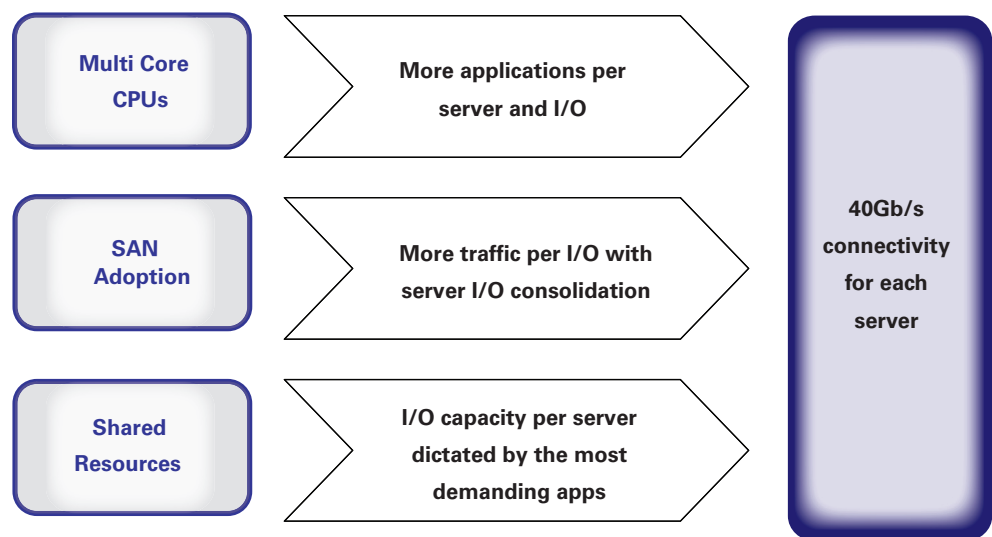
Multi Core CPUs	**More applications per server and I/O**	
SAN Adoption	**More traffic per I/O with server I/O consolidation**	**40Gb/s connectivity for each server**
Shared Resources	**I/O capacity per server dictated by the most demanding apps**	

Figure 1: Demonstrates the need for more I/O capacity

1.2 Power Consumption – the Number 1 Problem

Servers are the fastest growing consumer of power. In data centers today, energy costs exceed server purchase cost (Source: Electronics Cooling Magazine, Feb. 2007). There are two primary approaches for power reduction:
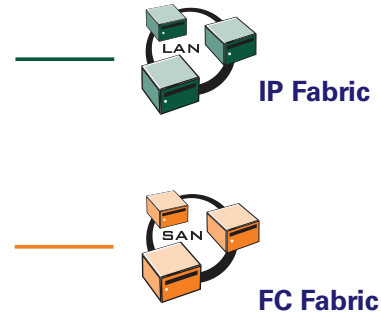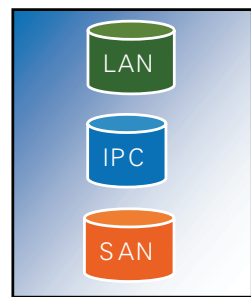
- **Reduction in power consumption per server:** Blade servers, the fastest growing server market segment (source: IDC), are delivering on their promise to deliver lower power and space-saving solutions.

- **Reduction in number of servers needed:** Server clustering enables use of cheap commodity servers to deliver SMP and mainframe-like performance and capacity scaling. When compute capacity can be scaled linearly with the addition of servers to the cluster, the highest levels of efficiency can be achieved.

High-performance I/O plays a key role in both of these approaches. Blade servers lack I/O real estate (PCI slots), thus require higher bandwidth I/O PCI adapters that can carry all data center traffic types (I/O unification) while still consuming the least power. Efficient clustering requires high-bandwidth and low-latency I/O to enable higher levels of clustering efficiency

1.3 I/O Unification Becoming Critical

Data center servers need to handle three classes of traffic – LAN/WAN (Local and Wide Area Networking), SAN (Storage Area Networking) and IPC (Inter Processor Communication or server-to-server messaging). Each of the three classes requires different I/O service levels. Today, the dominant I/O technologies serving LAN/WAN, SAN and IPC are Ethernet, Fibre Channel and InfiniBand respectively. All three types of I/O adapters are needed to deliver needed services, resulting in more management (e.g., multiple networks, cabling complexity), and more power and space consumption (requiring multiple I/O adapters and network infrastructure equipment). This leads to more dollars being spent while being less green. These problems are further exacerbated in blade servers that are compact and lack power and space for multiple I/O adapters, making I/O unification (where all three traffic classes are delivered using the same I/O adapter over a single cable) critical.

Virtualization Software

Figure 2: Multiple I/O Fabrics in the Data Center

1.4 Service Levels Dictate Unified I/O Suitability

Gigabit Ethernet, 10 Gigabit Ethernet, Fibre Channel and InfiniBand are popular I/O technologies that deliver varying degrees of LAN/WAN, SAN and IPC services.   Gigabit Ethernet is moderately suitable for LAN/WAN traffic, but because of its low reliability (data loss and software based retransmissions) and high latency characteristics, it is not suitable for SAN and IPC traffic. 10 Gigabit Ethernet can be excellent for LAN/WAN connectivity and with use of TOE (TCP Offload Engines), can deliver iSCSI based SAN traffic. However, it still suffers from high per port cost (especially in the infrastructure), unreliability and inefficient scalability (for example spanning tree restrictions), making it unsuitable for Fibre Channel storage connectivity over SAN (the most widely deployed storage systems) and IPC. Fibre Channel delivers excellent services for SAN connectivity, but lacks capabilities needed for LAN/WAN and IPC traffic. InfiniBand delivers very high-bandwidth, low-latency and reliability that matches or exceeds what Fibre Channel provides for SAN, Gigabit Ethernet and 10 Gigabit Ethernet for LAN/WAN, and is best-in-class for IPC traffic. Unified I/O in servers can be achieved using a single 40Gb/s InfiniBand adapter, while connectivity to Ethernet-based LAN/WAN and Fibre Channel-based SAN networks can be achieved through use of module InfiniBand to Ethernet or Fibre Channel gateways that enable cost-effective SAN and LAN capacity provisioning and scaling.

Connectivity Type	LAN	IPC (server-server)	SAN
Gigabit Ethernet	◑ Inadequate bandwidth	◔ High latency, unreliable	◔ Packet drops, low bandwidth
10 Gig Ethernet	● 	● High latency unreliable, limited scalability	◑ Unreliable, not suitable for FC over Ethernet
Fibre Channel	○ SAN-only fabric	○ SAN-only fabric	●
InfiniBand	● IP fabric connectivity thru gateway (GW)	● Reliable, highest BW, lowest latency	● Reliable, high BW IB SAN or FC SAN access thru GW

Figure 3: I/O Service Levels Delivered by Different I/O Fabrics

There are a number of IEEE initiatives - some in fledgling stages – under the umbrella name Data Center Ethernet (DCE) that aim to address the feature gaps required for delivering unified I/O over 10 Gigabit Ethernet. Deployment of such features would require significant overhaul of the data center networks using new equipment from system manufacturers. Cost-effective deployment of DCE is not expected until 2011 or later according to many industry experts. InfiniBand, which provides the same services or better, on the other hand, has already gone through multiple product generations and are ready for cost-effective production deployment using servers and network equipment from Tier-1 OEMs.

1.5 InfiniBand as Unified I/O for Servers

InfiniBand I/O adapters, switches and gateways deliver cost-effective, low power, efficient, loss-less and reliable server and storage connectivity. It is the highest performing industry-standard I/O solution with 20Gb/s products in deployment since 2006 and 40Gb/s products since 2008. InfiniBand protocol software, available in all popular operating systems, deliver industry-standard application interfaces like IP, sockets, SCSI, iSCSI, NFS etc., which makes application deployment transparent to the underlying fabric. Because it delivers higher levels of performance and services, a single InfiniBand adapter can replace multiple Fibre Channel and Ethernet adapter cards from each server resulting in immediate cost, power and infrastructure equipment savings. InfiniBand, as a unified I/O, delivers specific advantages in the following areas:

- Optimization of multi-core server and storage power and performance
- Efficient capacity scaling in native OS and virtualized environments
- Delivering end-to-end Ethernet and Fibre Channel connectivity
- Price-performance advantages leading to CapEx and OpEx savings

1.6 Optimization of multi-core server/storage power and performance

Unbalanced Computing and I/O: Today's servers are dominated by multi-core architectures. Two quad core CPUs can consume up to $2800 in costs and 240W in power. Both CPU and memory access bus architectures today support 80Gb/s of bandwidth, enabling memory usage of up to 8GB which cost up to $600 and consume about 96W of power. Connecting this high-

performance compute system to the outside world can be a single or dual Gigabit Ethernet port(s) that are very cost and power effective (about $100 and 4W respectively) - hence, the popularity of Gigabit Ethernet NICs in server systems. However, with network intensive applications running on the CPU and memory subsystems, they can be bottlenecked in the I/O with Gigabit Ethernet delivering only 1Gb/s of bandwidth and 50 microseconds of application latency. This results in idle CPU and memory cycles and therefore significant wasted dollars and power.

The Band-aid Solution: The problem is further exacerbated in virtualized server environments like VMware Virtual Infrastructure, where the execution of multiple virtual machines per physical server demands significantly higher I/O throughput. As such, the use of four to six Gigabit Ethernet NICs and two Fibre Channel HBAs is a common configuration where the I/O bandwidth bottleneck is somewhat alleviated with about 12Gb/s of bandwidth per server while application latency remains at 50 microseconds. Using multiple I/O adapters in a server brings new challenges in the areas of I/O utilization, cost and power. The I/O capacity provisioned in each server may go unutilized as the individual adapters serving specific functions cannot be repurposed for SAN, LAN and IPC functions on demand. Compared to the previous scenario with Gigabit Ethernet NIC cards only, where the cost and power consumption was about $100 and 4W respectively, this scenario increases to about $2400 and 28W respectively, not to mention the additional costs of cabling and associated complexity.
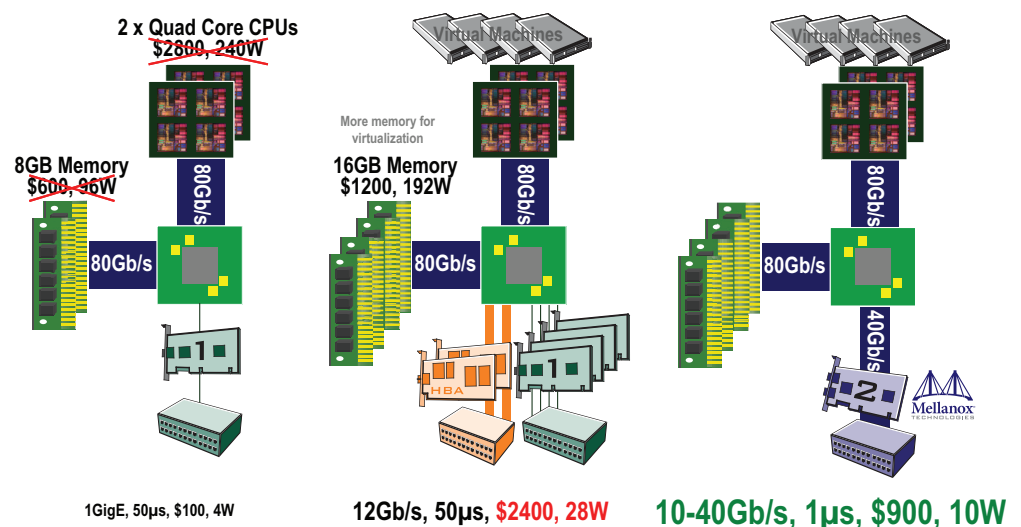

Figure 4: InfiniBand I/O Reduces Cost and Power Significantly

InfiniBand To The Rescue: When 10, 20, or 40Gb/s InfiniBand is used in the above scenarios, the power of the compute and memory systems are realized and idle cycles are avoided for network I/O sensitive applications. Because InfiniBand delivers all the services needed for unified I/O, a single or dual-port InfiniBand adapter can replace multiple adapters used in the above scenario (four to six Gigabit Ethernet NICs and two Fibre Channel HBAs). With this solution, I/O cost and power are significantly reduced to about $900 and 10W respectively, and application latency is brought down to about 1 microsecond. I/O utilization is significantly improved as channel I/O (where the I/O adapter resources are partitioned and can be dedicated to applications or virtual machines) and quality of services features enable repurposing of available I/O adapter bandwidth for different functions. This solution also delivers the benefits of "one-wire" for delivering I/O from the servers, significantly reducing cable management complexity.

1.7 Efficient capacity scaling in native OS and virtualized environments

When compute, LAN or SAN capacity needs to be scaled in the data center, and multiple Ethernet and Fibre Channel I/O adapters are used per server, the following are some basic steps that IT managers need to take:

- **Compute Scaling:** More physical servers need to be added to increase compute capacity. Each new server comes with its own new I/O ports. The trio of servers, SAN and LAN sides of the data center needs to be reconfigured.

- **SAN and LAN Scaling on Servers:** In this case, more Ethernet and/or Fibre Channel adapters need to be added to each server. This assumes free PCI slots are available, which may not be the case in blade servers. When new adapters are added on the servers, there are new server I/O ports and cables that need to be connected to the LAN and SAN sides, which requires reconfiguring of LAN and SAN infrastructures.

- **SAN and LAN Scaling in the Infrastructure:** In this case, more LAN and/or SAN capacity needs to be added in the infrastructure. This results in more Ethernet and Fibre Channel Switch ports becoming available to servers to connect to, requiring reconfiguring of the server side to avail of the added infrastructure capacity.

When InfiniBand is used to deliver a unified I/O solution, the above capacity scaling complexities are eliminated. The key is the separation of compute, LAN and SAN traffic management, where each of the three islands in the datacenter can be independently scaled and managed. Gateways that enable virtualization of Ethernet and Fiber Channel over InfiniBand are available today and their functionality is depicted in the figure below.
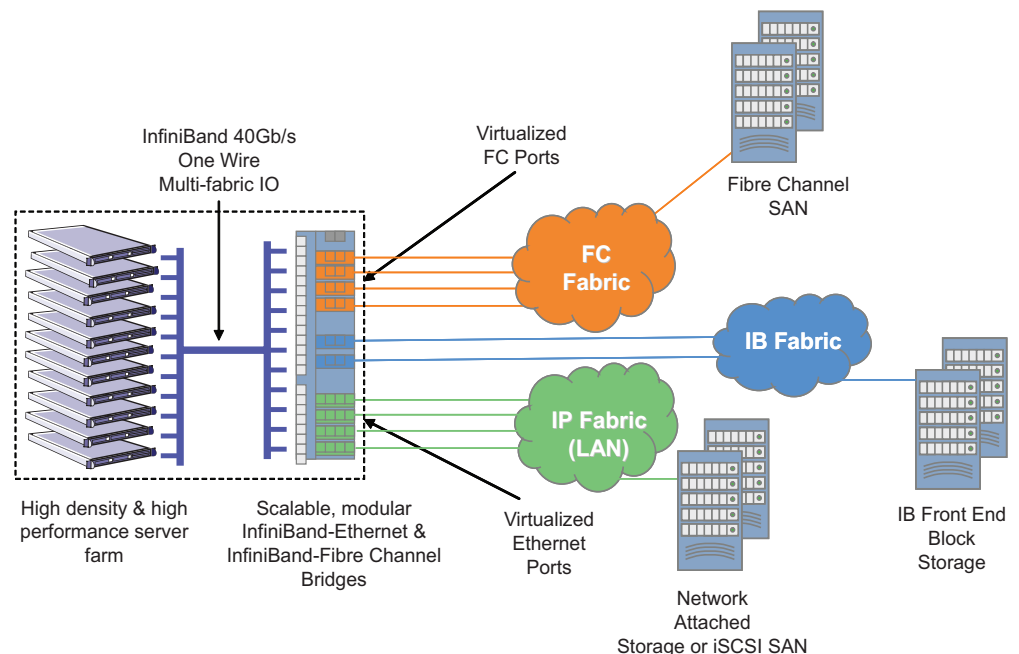

Figure 5: Unified "One-Wire" I/O using InfiniBand

There are significant benefits of this solution:

- **Unified I/O based savings of up to 50% CapEx and OpEx:** The same high-bandwidth InfiniBand adapter is used for compute, LAN and SAN traffic. Adapter I/O resources can be provisioned in each adapter to provide the segmentation as available when multiple adapters are used, but with the added flexibility that those resources can be repurposed on demand. This eliminates cost of under-utilized NICs, HBAs, and cabling complexity.

- **Delivering end-to-end Ethernet and Fibre Channel connectivity:** Servers using InfiniBand adapters expose IP/Ethernet interfaces to applications for LAN traffic. Similarly, they expose SCSI or iSCSI interfaces to applications for SAN traffic. This applies to both native OS and virtualized server environments. InfiniBand to Ethernet and Fibre Channel gateway solutions available from leading OEMs provide virtualized Ethernet and FC ports for connectivity to existing LAN and FC SAN infrastructures. Together, seamless end-to-end Ethernet and Fibre Channel connectivity is achieved while improving performance and utilization.

1.8 Price-performance advantages leading to CapEx and OpEx savings

This section describes realization of the above benefits in two usage scenarios, one is a database server application using Oracle 10g RAC, and one is a virtualized server application using VMware ESX Server version 3.5.

Power, Floor Space and TCO Advantage for Database Applications

HP and Oracle benchmarked the impact of InfiniBand I/O on Oracle 10g RAC applications using the popular TPC-H benchmark (www.TPC.org). The tests used a HP c-Class 7000 Blade Server, where server and storage blades were used for a first "all-blade" server and storage benchmark. Eight compute servers were used as database servers and 16 storage servers were used for storage SAN access over unified InfiniBand I/O. HP compared the results versus rack servers and other economical SAN solutions and proved that this configuration delivered record price-performance TPC-H benchmark (QphH = 39.613, Price/QphH = USD 12.57) while bringing significant savings – 30% power/cooling savings, 20% floor space savings and 40% savings on three-year TCO (total cost of ownership) for the storage subsystem. In late 2008, HP and Oracle introduced the Exadata, a high-performance database appliance that uses InfiniBand as the unified fabric. This "datacenter in a box" system is based on a unique architecture that takes full advantage of InfiniBand technology.

Unified I/O Delivering 50% I/O Cost Reduction, Higher I/O Utilization and 30% More Green

- When InfiniBand is used with VMware ESX Server version 3.5, IT managers can scale out their virtual machines (VMs) transparently while saving significantly on I/O related costs. When multiple Ethernet and Fibre Channel I/O adapters are replaced with InfiniBand I/O adapters to service the VMs and other ESX server I/O functions, it results in up to 50% I/O purchase and maintenance cost savings and 30% reduction in I/O power consumption (based on calculations conducted by Mellanox using I/O per port list prices from a tier 1 networking system vendor). Cabling costs have been cut to a third, and flexibility of deployment and I/O utilization has been significantly enhanced by enabling dynamic provisioning of virtual NICs and HBAs over InfiniBand.

Mellanox®
TECHNOLOGIES

350 Oakmead Parkway
Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com