I/O Virtualization Using Mellanox InfiniBand And Channel I/O Virtualization (CIOV) Technology

- Reduce I/O cost and power by 40 – 50%
- Reduce I/O real estate needs in blade servers through consolidation
- Maintain interoperability with existing network infrastructures
- Maintain compatibility with applications and management software
- Scale I/O performance to native-OS levels
- I/O virtualization framework for service oriented architectures
- Complements current and future virtualization ecosystem developments with standard implementations

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

Introduction to Server Virtualization

When a physical server is virtualized, it results in multiple logical servers. Each logical server comprises a virtual machine (VM) that works over a virtualization intermediary software layer, also known as the virtual machine monitor (VMM) or hypervisor (see figure 1 below). Each VM runs its own operating system (called the guest OS) and its own instance of the underlying physical resources in a server, such as CPU, memory, I/O etc. The VMM creates and controls the VMs and resources allocated to them. It also provides a framework for virtual infrastructure management, such as a user interface for creating VMs, associating resources to them dynamically and enabling migration of the VMs across physical servers. VMM software is available from vendors such as VMware, Microsoft, XenSource, Novell and Red Hat. The latter three are based on the open source Xen VMM.

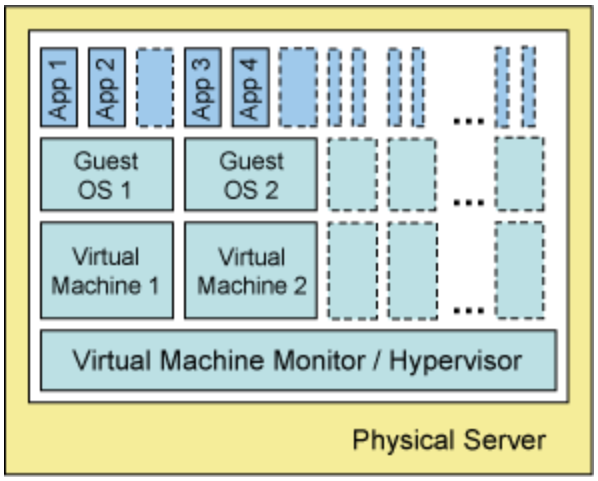

Figure 1: Virtual Server with VMM, VM, Guest OS and Applications

Benefits of Server Virtualization

Server virtualization is being increasingly deployed in enterprise data centers because of the numerous benefits that VMM software offers:

- **Improving server utilization**: When physical servers are tied to fixed applications, their utilization is significantly less than when logical servers (or VMs) can dynamically assign applications to physical servers on an as needed basis. This results in significant server consolidation, effectively doing more with less physical resources
- **Improved reliability with reduced downtime**: Unhealthy physical servers can be easily taken offline for maintenance without impacting services. This is achieved by seamless migration of the VMs from the unhealthy physical servers to another set of healthy physical servers. Once the servers have been fixed, the VMs can be seamlessly brought back to the original servers, once again without interrupting any services.
- **Dynamic resource allocation and load balancing**: Depending on workload needs, additional physical resources can be dynamically assigned to existing VMs, or new VMs can be created to address increasing workloads. In the same vein, VMs can be migrated to physical servers that have more resources available to address the increased workloads.
- **Service Oriented Architecture**: The above benefits enable creation of data center architectures where server resources can be tailored purely to the service needs of applications and users, without the need to physically redeploy, reconfigure or change servers, storage and network connectivity.

©2007 Mellanox Technologies Inc.

Overall, the above benefits improve the end user's total cost of ownership.

I/O virtualization (IOV) enables the sharing of physical I/O resources in the server by the VMs. I/O functions offered to VMs are the same as those available to physical servers, namely networking, clustering, storage area networking and management. The I/O also plays a role in how VMs are migrated across physical servers or in how new VMs are deployed in physical servers, thereby contributing to enable and enhance the benefits listed above. Multiple types of IOV implementations are possible today with available VMM software. Future enhancements are planned by the VMM software vendors in conjunction to other key players in the ecosystem, such as CPU and chipset vendors, I/O vendors and standards-bodies.

I/O Virtualization Options Today

VMM software vendors support three broad categories of I/O virtualization (IOV) methods, or some hybrids of them:

- Fully virtualized, software-based I/O virtualization
- Native I/O virtualization, which is a combination of software and hardware-based I/O virtualization
- Pass-through, hardware-based I/O virtualization

The three approaches have their pros and cons and they are discussed below. Currently shipping Mellanox InfiniHost III-based InfiniBand adapters support the first two methods. InfiniBand adapters based on the Mellanox ConnectX-based architecture supports all three methods.

Fully Virtualized, Software-based I/O Virtualization

In this method, the VMM is implemented entirely in software. The VMM virtualizes a physical I/O adapter (such as a SCSI or FC HBA, InfiniBand HCA or Ethernet NIC) into multiple virtual I/O adapters (i.e., virtual HBAs, HCAs and NICs respectively) that are then assigned to VMs. See figure 2 below. To ensure compatibility with legacy operating systems and I/O stacks, the VMM software sometimes maintain legacy virtual adapter interfaces in the VM even when new and faster physical adapters are used. By doing so, The VMM transparently makes the additional bandwidth available in faster adapters to the VMs, without requiring any changes in the VMs. A specific example of this is with InfiniBand HCAs where the VMM exposes the legacy virtual NIC and virtual HBA interfaces to the VMs, as shown in the figure 3 below.

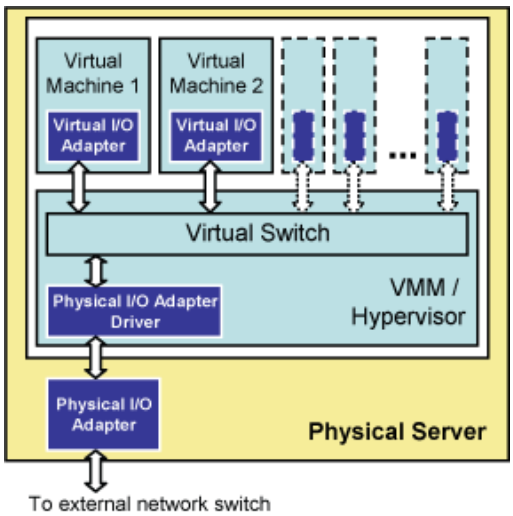

Figure 2: Software-based I/O Virtualization

The virtual adapters are enabled with the same set of resources (such as network or LUN addressing) as are available in the physical adapters. To enable communication between the VMs and steering of traffic between the physical adapter and the virtual adapters, the VMM software implements a shared-memory software switch, also called a virtual switch. In some VMM software implementations, the virtual switch also acts as a layer upon which all virtual infrastructure management functions are implemented.

In this software-based I/O virtualization method, the VMM is always in the path between the VMs and the physical I/O adapter. This has advantages and disadvantages:

Advantages:

- Enhances legacy support and stability of VMs and applications within VMs
- Benefits of dynamic load balancing and reduced down time through hot migration of VMs are available with VMM software that support this method of I/O virtualization
- Better server utilization than using non-virtualized servers

Disadvantages:

- Higher CPU utilization because of the I/O functions that are provided by the VMM software
- Limits number of VMs per server, and therefore limits server utilization benefits compared to the other two IO virtualization methods
- VMs suffer from lower I/O bandwidth and higher I/O latency compared to the other two IO virtualization methods or use of non-virtualized servers

Use of Mellanox CIOV in Software-based I/O Virtualization

Mellanox InfiniBand adapters with Channel I/O Virtualization (CIOV) technology can be used with the software-based I/O virtualization method and can enjoy the advantages listed above. In addition, CIOV enables the Mellanox InfiniBand adapter to be:

- Shared for both networking and storage area networking functions
- Shared as multiple physical adapters

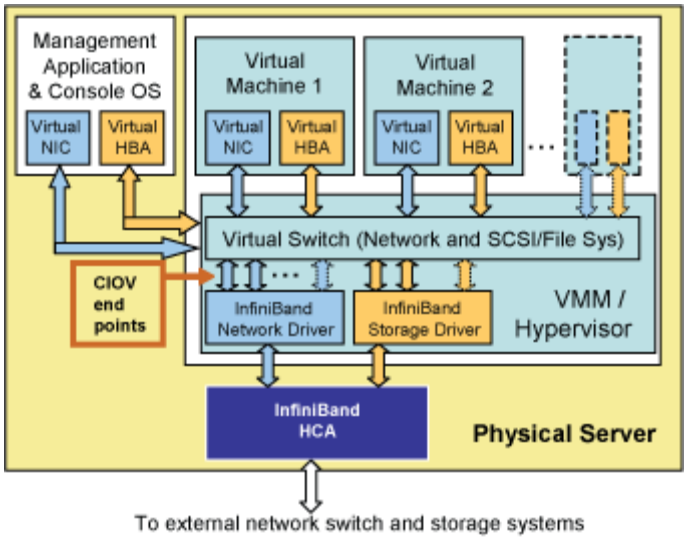

Figure 3: Mellanox CIOV with Software-based I/O Virtualization

Channel I/O Virtualization (CIOV) technology in Mellanox InfiniBand adapters enable dividing up of the I/O resources in the same adapter, maintaining full hardware based isolation, protection and quality of service. This amounts to applications seeing one adapter as multiple adapters, with the CIOV interfaces appearing as end points for each of those adapters. Such end points in the adapter can be applied to VMs as well as different I/O functions such as clustering, networking or storage. See figure 3 above.

In software-based I/O virtualization, the CIOV end points cannot be directly exposed to the VMs. However, they can be exposed to the lower edge of the VMM, or more specifically to the lower edge of the virtual switch and virtual SCSI and file system layers in the VMM. For networking, a single or multiple InfiniBand layer 2 interfaces (using a variation of IPoIB – Internet Protocol over InfiniBand) can be exposed to the virtual switch using such end points. For storage area networking (SAN), a single or multiple LUN interfaces (using SRP - SCSI RDMA Protocol or iSER – iSCSI Extensions for RDMA) can be exposed to the virtual SCSI and file system layers in the VMM. The VMs still see virtual NIC and virtual HBA interfaces, as do virtual infrastructure management consoles that reside above the virtual switch and virtual SCSI and file system layers in the VMM to provide functionalities such as dynamic allocation of physical I/O resources to the VMs and hot migration of VMs.

Mellanox CIOV Delivers Significant Performance and Cost Benefits

Performance: Mellanox InfiniBand adapters support up to 20 Gb/s bandwidth. In actual implementations, even with the overheads of the VMM being in the data path, with use of Mellanox InfiniBand adapters with Channel I/O technology, VMs can experience 3-4 times higher bandwidth than Gigabit Ethernet NICs, or 6-7 times higher bandwidth than 2Gb/s Fibre Channel adapters.

I/O Cost: First of all, since the VMs and virtual infrastructure management is transparent to the use of InfiniBand, it significantly reduces operational expenses in the areas of software qualification and IT training. Secondly, software-based I/O virtualization deployments typically use multiple NICs, either for teaming to achieve higher bandwidth, or for dedicated functions such as VM migration, production VMs, backup etc. They also use multiple Fibre Channel adapters to sustain required SAN bandwidth or for dedicated storage related functions. All of those functions can be consolidated over the same InfiniBand adapter with CIOV, delivering significant benefits – reducing I/O related expenditures and power consumption by 40% to 60% in typical environments (see Table 1 below).

©2007 Mellanox Technologies Inc.

Table 1: Estimated I/O Cost, Power, & Performance Comparison (Software-based IOV)

Server I/O Performance, Cost, Power	Using Gigabit Ethernet and Fibre Channel	Using Mellanox based InfiniBand and CIOV Technology
Networking Performance Per Adapter Equivalent To	1 GigE NIC	Up to four GigE NICs
SAN (Storage) Performance Per Adapter Equivalent To	1 Fibre Channel Adapter	Up to seven 2Gb/s Fibre Channel Adapters or up to four 4 Gb/s Fibre Channel Adapters
Total I/O Initial Purchase Cost*	About $870,000	$352,000 - $360,000
Maintenance Cost per port	Costs equivalent to about six ports per server	1/3rd to 1/6th the cost of using GigE and FC depending on whether 2 InfiniBand adapters (high availability) or one is used.
Total I/O Power Consumption*	2300 Watts	About 1700 Watts

*Cost and power analysis based on:

- Tier 1 OEM list prices and data sheets
- Building a virtualized infrastructure of 128 physical servers with server to server, SAN and LAN connectivity

(Note: For further details on cost and power savings using InfiniBand in a software-based I/O virtualization environment such as VMware Virtual Infrastructure 3, please contact your Mellanox sales representative.)

Finally, use of available InfiniBand to Gigabit Ethernet and InfiniBand to Fibre Channel gateways enable InfiniBand-based virtualized servers to fit and interoperate seamlessly in Gigabit Ethernet and Fibre Channel-based network infrastructures.

The above are further explored below.

Deployment Scenarios

The following deployment scenarios are used as examples in this discussion.

Scenario I:

The end user is trying to resolve one or more of the following issues:

- Looking at new servers and software-based IOV (such as from VMware), but there isn't enough PCI slots or real estate on the servers (e.g., blade servers) to install multiple NICs and FC I/O adapters
- Wants to reduce I/O cost and power consumption per server
- Wants the VMs and management infrastructure to operate transparently without requiring re-qualification or significant new IT training
- Wants to continue to use the existing GigE based LAN and FC based SAN setups or infrastructures

The user's current server setup looks somewhat like in figure 4 below. The user has four physical Gigabit Ethernet NICs and two Fibre Channel (FC) HBAs per server. The NICs and HBAs connect to separate switch and network segments as shown below. There are multiple ports and wires emerging from each server.

Using Mellanox InfiniBand and CIOV technology, and a set of new blade servers for example, the setup can be converted to what is shown in figure 5 below, meeting all of the end user goals listed above. Note that each server is populated with one InfiniBand HCA only, compared to six I/O adapters using GigE and Fibre Channel.

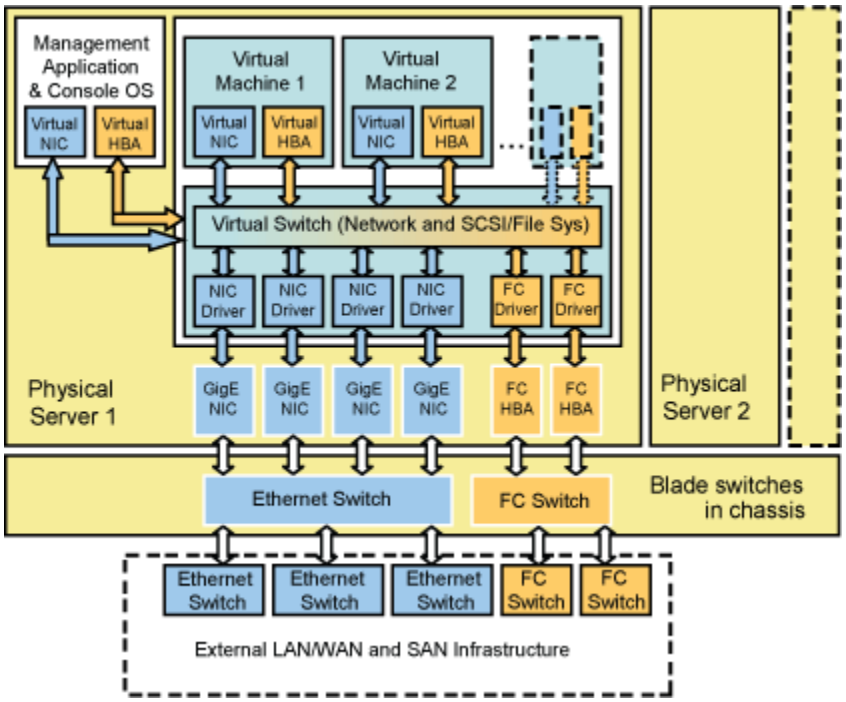

Figure 4: Software-based I/O Virtualization with multiple GigE NIC and FC adapters

Cost effective blade server solutions with internal and external InfiniBand-to-Ethernet and InfiniBand-to-Fibre Channel gateways are available from all tier 1 server OEMs. (Note: contact your Mellanox sales representative for further details on available solutions).

In typical server virtualization scenarios, the above InfiniBand-based solution can reduce I/O cost and power consumption per server significantly as mentioned earlier (section 3.1.2).

©2007 Mellanox Technologies Inc.

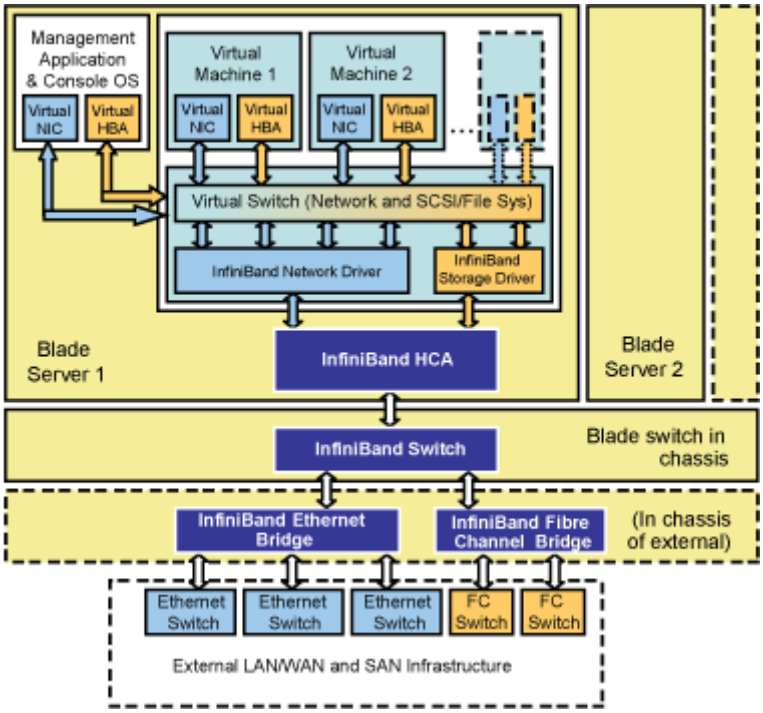

Figure 5: Software-based I/O Virtualization with InfiniBand adapters and gateways

Scenario II:

The end user is trying to resolve one or more of the following issues:

- Wants to use a higher bandwidth server-to-server connectivity option for specialized functions like migrating VMs between physical servers, or IPC (inter process communication)
- Wants to speedup SAN-centric functions in virtual server environments, for e.g., deployment of new virtual appliances, backup to SAN storage, running SAN intensive database applications within VMs etc.

In this scenario, the end user can use the higher bandwidth InfiniBand network for server-to-server connectivity. Native InfiniBand storage solutions (contact Mellanox for available solutions) can be used to connect directly with the InfiniBand based SAN. Alternatively, InfiniBand-to-Fibre Channel gateways can be used to connect to Fibre Channel SANs. See figure 6 below.

In either case, subject to availability of adequate back-end storage capacity, the InfiniBand HCA in a software-based I/O virtualization environment (such as VMware) can deliver up to 1500 MB/s of block storage I/O throughput from the VMs (divided evenly across multiple VMs).

©2007 Mellanox Technologies Inc.

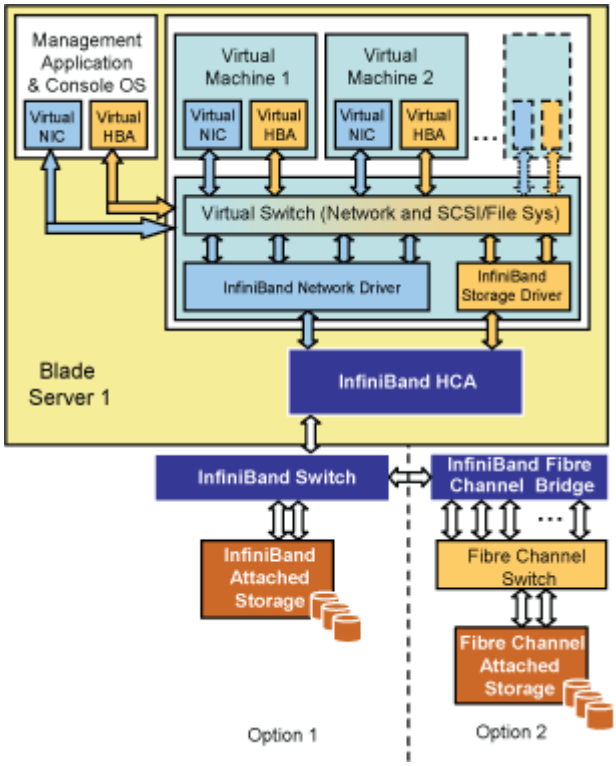

Figure 6: Use of InfiniBand for Server-to-Server and SAN connectivity functions

Native I/O Virtualization Solutions

This method is a blend of software and hardware-based I/O virtualization. The software-based VMM is still responsible for setting up the I/O resources for the virtual adapters in the VMs, and controlling those resources. This includes all memory mapping and setup operations, and isolation of I/O resources across the VMs. However, once the I/O resources are assigned, the VMs are allowed direct access to the I/O adapter for sending and receiving data, except that DMA remapping functions are still controlled by the VMM. In other words, the data path between the VM and the physical I/O adapter bypasses the VMM, especially the packet steering functions in the virtual switch. However, the control path and memory mapping functions are maintained by the VMM. This requires that the VMs install drivers specific to the physical I/O adapter. The VMM virtual switch may continue to be used for VM to VM communication, unless the I/O adapter supports switching between the end points that connect to the VMs for the data path.

The native I/O virtualization method has the following advantages and disadvantages:

Advantages:

- I/O data path performance from the VMs is accelerated significantly
- I/O data path latency can be reduced significantly
- CPU utilization is significantly better than software-based I/O virtualization, enabling more VMs per physical server, improving server utilization

Disadvantages:

- Every guest OS supported in VMs have new I/O driver stack, requires re-qualification of applications in VM.
- Migration of VMs require re-start of the VMs. Benefits of dynamic load balancing and reduced down time are lost.
- The software VMM is involved in setting up of I/O resources. This means that hardware based isolation and quality of service methods cannot be easily applied (without adequate VMM support), adversely affecting CPU utilization and I/O bandwidth when such mechanisms are implemented.

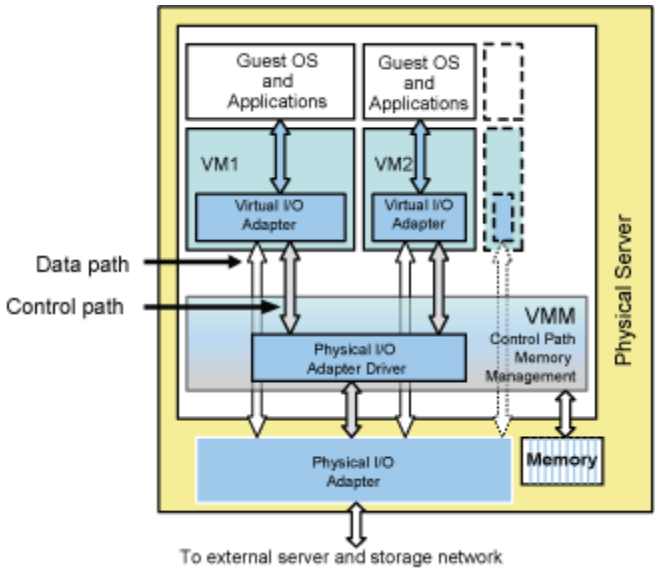

Figure 7: Native I/O Virtualization

Use of Mellanox CIOV in I/O Native I/O Virtualization

In native I/O virtualization, Mellanox CIOV benefits are available both at the VMM and VM levels. In other words, applicable benefits in the VMM are similar to those available in the software-based I/O virtualization. The following VMM functions can be implemented using CIOV:

- Multiple end points to interface directly with the VMs, eliminating the need of the virtual switch in the VMM for that functionality
- Virtual switching and steering of traffic based on packet header information

Additionally, the VMs and guest operating systems executing on the VMs can open multiple channels on the I/O adapter for data path operations. I/O consolidation benefits, i.e., use of channels for multiple traffic types become available at the VM level and each VM can independently set up its own set of channels based on application requirements.

The result is more flexibility in how the VMs access and use the physical I/O resources for data path operations. Also, since the VMM is not in the path of data path operations, RDMA and transport offload benefits available with CIOV technology become available to the VMs, resulting in near-native OS level (e.g., Linux) bandwidth and latency capabilities from the VMs.

©2007 Mellanox Technologies Inc.

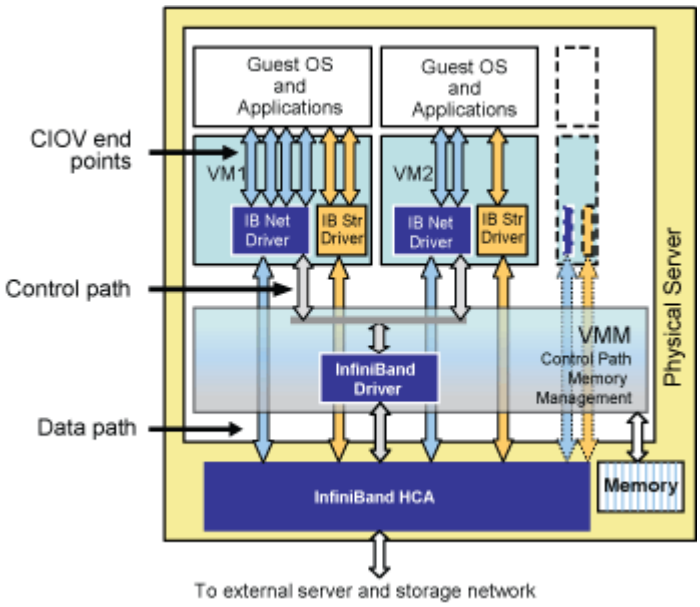

Figure 8: Mellanox InfiniBand and CIOV with Native I/O Virtualization

CIOV Enables the Best in Native I/O Virtualization Performance

Performance: As mentioned, Mellanox InfiniBand adapters support up to 20 Gb/s bandwidth. In actual native I/O virtualization implementations, it has been proven that with use of Mellanox InfiniBand adapters with CIOV technology, VMs can experience up to twelve times more bandwidth (using RDMA protocols on the VMs) than Gigabit Ethernet NICs, or seven times more bandwidth at 2 Gb/s Fibre Channel adapters. Latency is proven to be as low as $1/10^{th}$ of Gigabit Ethernet NICs. Use of non-RDMA protocols such as IPoIB on the VM can result in up to five times more bandwidth than Gigabit Ethernet NICs.

I/O Cost: The same I/O consolidation benefits as those available in software-based I/O virtualization are available. All storage, networking and other functions relevant in virtual infrastructures can be consolidated over the same InfiniBand adapter with CIOV, delivering significant benefits – reducing I/O related expenditures and power consumption by 40% to 60% in typical environments. See Table 2 below.

Finally, as in the software-based IOV scenarios discussed above, use of available InfiniBand to Gigabit Ethernet gateways and InfiniBand to Fibre Channel gateways enable InfiniBand based virtualized servers to fit and interoperate seamlessly in Gigabit Ethernet and Fibre Channel based network infrastructures.

Table 2: Estimated I/O Cost, Power, Performance & Comparison (Using Native I/O Virtualization)

Server I/O Performance, Cost, Power	Using GigE and Fibre Channel	Using Mellanox InfiniBand and CIOV Tech
Networking Performance Per Adapter Equivalent To	1 GigE NIC	Up to twelve GigE NICs
SAN (Storage) Performance Per Adapter Equivalent To	1 Fibre Channel Adapter	Up to seven 2Gb/s Fibre Channel Adapters or up to four 4 Gb/s FC Adapters
Total I/O Initial Purchase Cost*	About $870,000	$352,000 - $360,000
Maintenance Cost per port	Costs equivalent to about six ports per server	1/3rd to 1/6th the cost depending on whether two HCAs are used (for high availability) or one is used.
Other Cost Savings Due to Lower CPU Usage for I/O Processing	None because of no native I/O virtualization support	Reduced CPU utilization can result in up 30-50% more VMs per physical server. This can result in use of fewer physical servers and significant savings
Total I/O Power Consumption*	2300 Watts	About 1700 Watts

*Cost and power analysis based on:

- Tier 1 OEM list prices and data sheets
- Building a virtualized infrastructure of 128 physical servers with server to server, SAN and LAN connectivity

(Note: For further details on cost and power savings using InfiniBand in I/O paravirtualization environments such as with Novell and XEN, please contact your Mellanox sales representative.)

Deployment Scenarios

Scenario I:

The end user is trying to resolve one or more of the following issues:

- Improve performance and latency of applications when they run within VMs
- Reduce the number of I/O adapters needed per server
- Reduce I/O cost and power consumption per server
- Continue to use the existing GigE LAN and FC based SAN setups or infrastructures

Using Mellanox InfiniBand and CIOV technology, and a set of new blade servers for example, the setup such as is shown in figure 9 below can be built, meeting all of the end user goals listed above. Performance from VMs can reach near native Linux levels. Note that each server is populated with one InfiniBand HCA only, compared to six I/O adapters when GigE and Fibre Channel adapters are used.

©2007 Mellanox Technologies Inc.

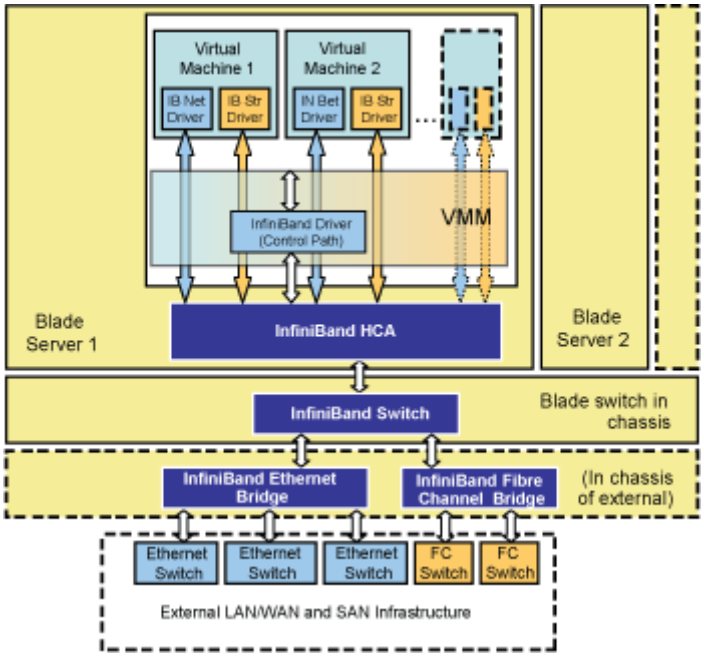

Figure 9: Use of InfiniBand in a Native I/O Virtualization Deployment Scenario

Pass-through, Hardware-based I/O Virtualization

Pass-through or hardware-based I/O virtualization takes I/O virtualization to the next level by enabling complete bypass of the VMM, especially with respect to memory operations. As in the native I/O virtualization mode, the VMs are able to pass and receive data directly from the VMs, without involving the VMM. In addition, I/O adapters working in this mode are able to export part of the I/O resources (as configured by the VMM at set up time) directly to the VMs.

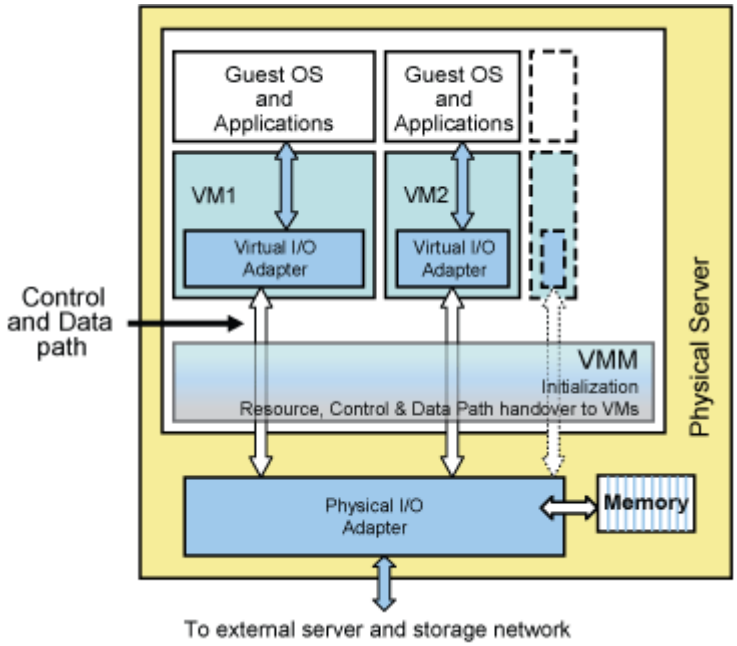

Figure 10: Hardware-based I/O Virtualization

The I/O adapter- as shown in figure 10 - or the CPU chipset implements DMA remapping functions and relieves the VMM from performing those functions. DMA remapping intercepts the I/O adapters attempts to access system memory and uses I/O page tables to remap the access to that portion of system memory that belongs to the target VM. The VMM is involved only in ensuring that the DMA requests of the VMs are isolated from one another.

Use of Mellanox CIOV in Hardware-based I/O Virtualization

Mellanox CIOV implementation in its ConnectX family of adapters implement required functions in the I/O adapter to make hardware-based I/O virtualization possible with existing and installed based on servers. See figure 11 below. Each resource in the adapter can be associated with a VM. The association is done via protection domain allocation and supports hundreds of VMs executing concurrently. Association of resources with VMs enables execution of not only the data path operations (as in the paravirtualized mode), but also control and configuration operations directly from the VMs. The memory management and DMA remapping capabilities in the adapter enable the VMs to register memory with the I/O device. The following VMM functions can be implemented using CIOV:

- Ensure that the end points exposed to the VMs are fully independent, each with its own comprehensive list of resources settings (including stateless or full transport offload services, RDMA and other transport services, interrupts etc.)
- Ensure enablement of quality of service policies from the VMs
- Multiple end points to interface directly with the VMs, eliminating the need of the virtual switch in the VMM for that functionality
- Virtual switching and steering of traffic based on packet header information

The result is ultimate flexibility in how the VMs access and use the physical I/O resources for data path and control path operations – similar to a native OS and non-virtualized environments.

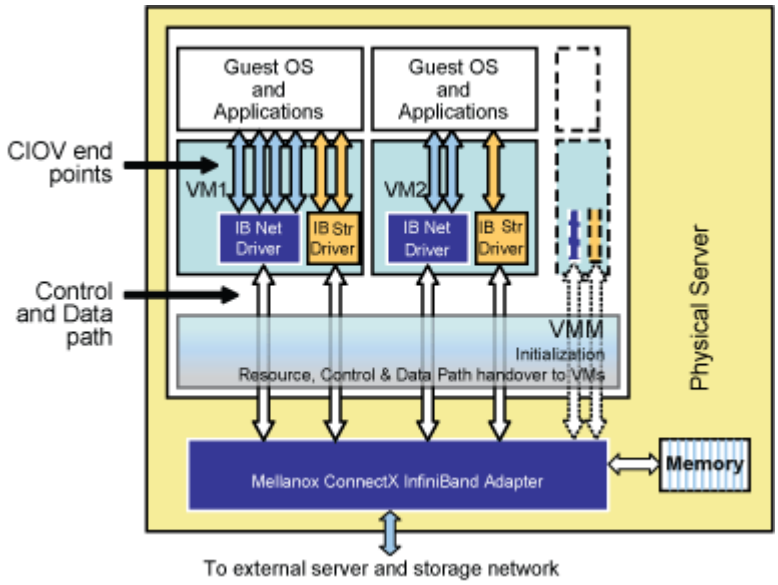

Figure 11: Hardware-based I/O Virtualization using InfiniBand and CIOV

Addressing the Future

Service Oriented I/O

Service oriented data centers are evolving and any of the above I/O virtualization methodologies can be deployed to enable a server and storage network to deliver service oriented functionality. It is critical that the I/O fabric connecting the servers and storage elements:

- Is able to deliver all data center connectivity functionality (such as clustering, communications, storage and management) in a seamless and coherent manner using flexible and virtualized I/O end points, and
- Without requiring physical redeployment of network elements.

Mellanox InfiniBand solutions with CIOV technology are prime for delivering true service oriented I/O in data center environments.

In Line with Ecosystem Trends

It is important to note that VMM software vendors must enable use of hardware-based I/O virtualization functions implemented in I/O adapters. Server virtualization technology is expected to evolve rapidly in the next few years, with the ecosystem of CPU, chipset, I/O adapter and VMM software vendors doing their bit in enhancing capabilities. PCI SIG driven IOV efforts relate to defining and standardizing the functionality of I/O adapter end points and DMA remapping functions. It is worthwhile to note that major x86 CPU vendors are introducing processors that offer hardware assistance for CPU and memory virtualization. Major features include:

- VM interrupt handling assistance
- Support for multiple logical CPUs – to speed context switches between VMM and VMs
- Assist I/O virtualization using integrated I/O memory management unit (I/OMMU), including support for DMA remapping

Mellanox CIOV technology is built with the above evolutions in mind. For example, when needed, the migration of DMA remapping to the CPU chipset will be enabled, while maintaining IOTLB (I/O Translation Look-aside Buffer) based cache for recent address translations to allow pre-fetching of translated addresses. The goal of CIOV technology is to work in tandem with enhancements in CPU, chipset and VMM software, and assisting where needed to improve performance, flexibility and resource utilization.

©2007 Mellanox Technologies Inc.

Conclusion

Server virtualization technologies offer many benefits that enhance agility of data centers to adapt to changing business needs, while reducing total cost of ownership. Virtualization technologies continue to evolve, as the ecosystem of software and hardware suppliers continue to enhance their products. Current IOV implementations are either software-based or paravirtualized. Future enhancements will shift I/O virtualization techniques to completely hardware-based, enabling higher performance and better resource utilization.

The Mellanox CIOV technology available in its InfiniHost III and ConnectX family of InfiniBand adapters offer significant values in all I/O virtualization environments, and are ready for future ecosystem enhancements. The following table summarizes some of those benefits in the software-based and native IOV environments (in comparison to building a similar server and storage network with Gigabit Ethernet and Fibre Channel):

Table 3

	Software-Based IOV	Native IOV
Net throughput per adapter	3-4 times GigE	10-12 times GigE
SAN throughput per adapter	10 times 2Gb/s FC	10 times 2Gb/s FC
I/O Initial Purchase Cost	Up to 60% cheaper	Up to 60% cheaper
I/O Maintenance Cost (per port)	Up to 1/6th the cost	Up to 1/6th the cost
I/O Power Consumption	30-40% less	30 – 40% less
Reduced CPU Utilization (fewer physical servers)	None	30 – 50% more VMs per physical server
Ethernet LAN & FC SAN interoperability	Yes	Yes
Legacy I/O Support in VMs	Yes	No
Management transparency	Yes	No
Recommended VMM Software*	VMware Virtual Infrastructure 3	Novell XEN 3.x

*Contact Mellanox sales representative about availability

Hardware-based I/O Virtualization with ConnectX solutions fill further enhance the above benefits.

©2007 Mellanox Technologies Inc.