

A White Paper by Neterion

THE MISSING PIECE OF VIRTUALIZATION™

Eliminating the I/O Bottleneck with IOV in Virtualized Servers

Written by:

Philippe Levy, Ravi Chalaka and Greg Scherer

April 2008



Executive Overview

The benefits of server virtualization for enterprise data centers are broadly recognized: server consolidation, better utilization of available compute resources, dynamic transfer of applications and isolation for both development and production environments. In short, reducing the cost and increasing the flexibility of the IT infrastructure. But virtualization faces some limits. IT managers commonly place applications in two categories, when it comes to virtualization: the “good” and the “bad” candidates. Typically, I/O-intensive and latency-sensitive applications like databases or ERP systems fall into the latter category and are excluded from virtualization projects.

Recent developments in multi-core processing have greatly enhanced the systems performance, making it possible to run more virtual machines per server, but the fundamental I/O bottlenecks still exist. A bigger network pipe does not necessarily remove these I/O bottlenecks because virtualized systems face challenging network overhead issues due to the extra layer of Hypervisor software between the virtual machine and the I/O hardware. As a result, many mission-critical applications are not currently being categorized as “good” candidates for virtualization.

I/O Virtualization (IOV) implemented in hardware changes everything. With a combination of new, industry-standard technologies and the utilization of 10 Gigabit Ethernet, silicon-based IOV eliminates the I/O bottleneck. With a multi-channel, hardware-based I/O architecture, a properly implemented IOV solution allows even the most performance-demanding applications to be virtualized. By virtualizing the I/O subsystem with IOV, the limits of virtualization are overcome by allowing more VMs per system and more applications to participate: IOV truly expands the benefits of virtualization across the datacenter – and results in increased cost savings in so far, untapped areas.

WHY VIRTUALIZE?

In a virtualized environment, system administrators are able to capture underutilized resources and re-allocate them to constrained applications. Resources can be dynamically allocated and load-balanced as the characteristics of traffic and applications change over time. Hardware can be transparently replaced or upgraded with minimum downtime. Utilizing existing resources more efficiently in this way leads to reduced infrastructure cost, better utilization of IT assets, lower power consumption, reduced cooling requirements – and inevitably, lower total cost of ownership.

Technical benefits

On a technical level, virtualization solves a fundamental problem: isolation of the upper layer resources (“guest” OS’s and applications) from the lower layer elements (compute hardware.) This is not a new development. IBM implemented this concept several decades ago

in their mainframes – “VM” (Virtual Machine) was the name of IBM’s operating system on these machines. What is new, however, is that this powerful concept is now available on volume servers and blades, deployed by the millions in enterprise datacenters around the world. This “mass virtualization” is both a significant opportunity and a measurable challenge. IT managers must find a way to optimize all the elements of virtualized systems (like the I/O components) in order to actually gain – and not lose – from the upgrade. Table 1 summarizes the technical benefits of server virtualization.

Economic benefits

On an economic level, server virtualization allows IT managers to realize savings of unprecedented magnitude. From server acquisition to management costs; from IT architecture to increased flexibility needed to manage a shifting workforce, virtualization brings about opportunities to reduce

Feature	Benefit
Full Isolation	Events within one virtual machine cannot impact another one
Multi-Platform	Maintain flexibility of choice in operating systems and software
No Rewrites	Supports legacy and new applications
Transparency	No changes in end-user environment
Software/Hardware Independence & Mobility	Dynamic, cross-system remapping of software to hardware resources

Table 1 – Technical benefits of server virtualization

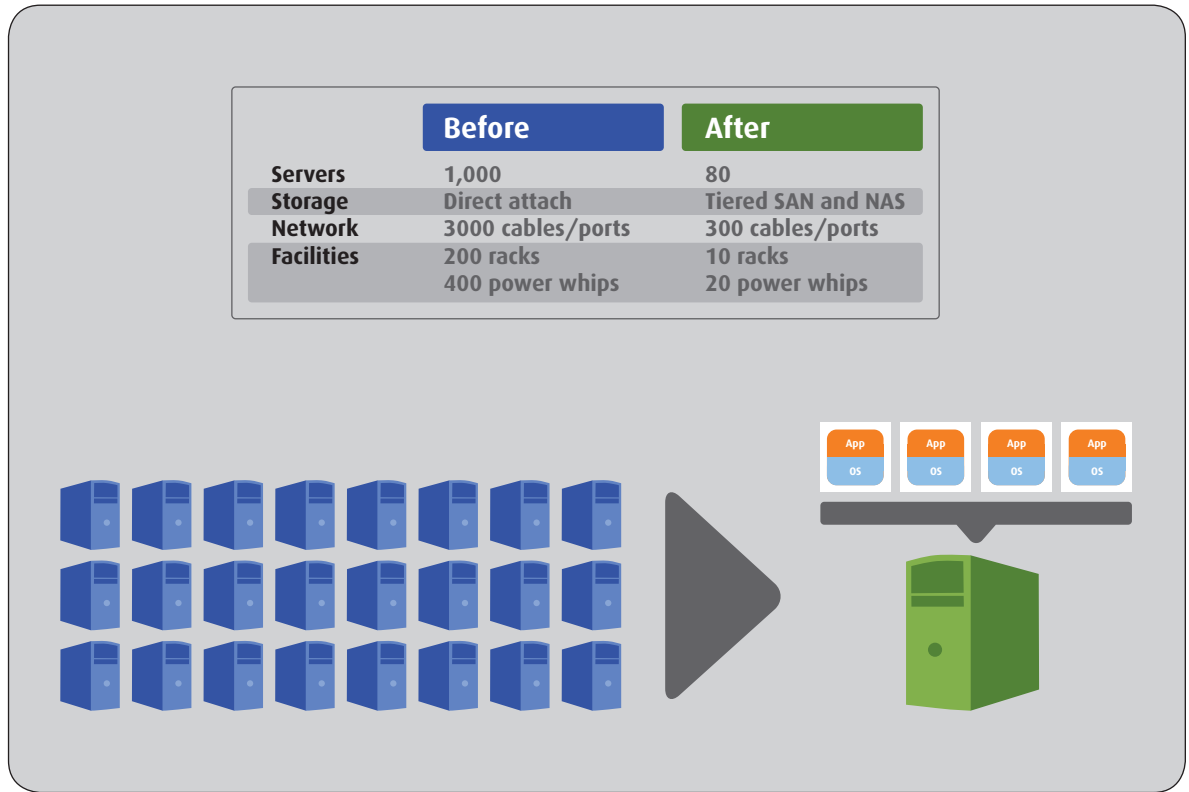


Figure 1 – Before/after server virtualization – VMware case study at a utility company

costs across the board – both immediately and over time. Enterprise datacenter managers have certainly recognized its great value. VMware reports that 100% of the Fortune 100 companies in the world are using its software today. Its website lists countless customer case studies, like the one shown in Figure 1, which illustrates the high level gain for customers. In a nutshell: simplification, consolidation, cost savings.

NEW TECHNOLOGIES DRIVE ADOPTION OF VIRTUALIZATION

As virtualization makes greater inroads in the datacenter every day, designers of computer systems have started optimizing their architectures and tackling the weak links that stand in the way of better, faster virtualization. Along with enhancements in software (both virtualized operating systems and the applications,) customer demand for virtualization has led the industry to optimize the hardware as well to improve overall system performance. For example, Intel and AMD have implemented virtualization-specific features in their most recent processors. In addition, these

processors now feature multiple cores (two, four or even eight processing units on a single piece of silicon). By their nature, multi-core processors lend themselves very well to virtualization: each core can be assigned one or several VMs (with their guest OSs and applications), physically separating them from one another, further strengthening isolation between processes.

Following a similar path, I/O architectures have been redesigned to support virtualization from end-to-end, making industry-standard servers evermore suitable to run virtualized operating systems. These new I/O architectures, together with multi-core CPUs, help boost the performance of virtualized applications as well as the server-to virtual machine ratio itself.

High speed Ethernet

A server running virtualization software like VMware would typically host multiple network interfaces, most likely Gigabit Ethernet, each running at one gigabit per second (Gbps). With higher-speed networking, like 10 Gigabit Ethernet, IT managers can replace multiple Gigabit connections with one physical link, running at 10 Gbps.

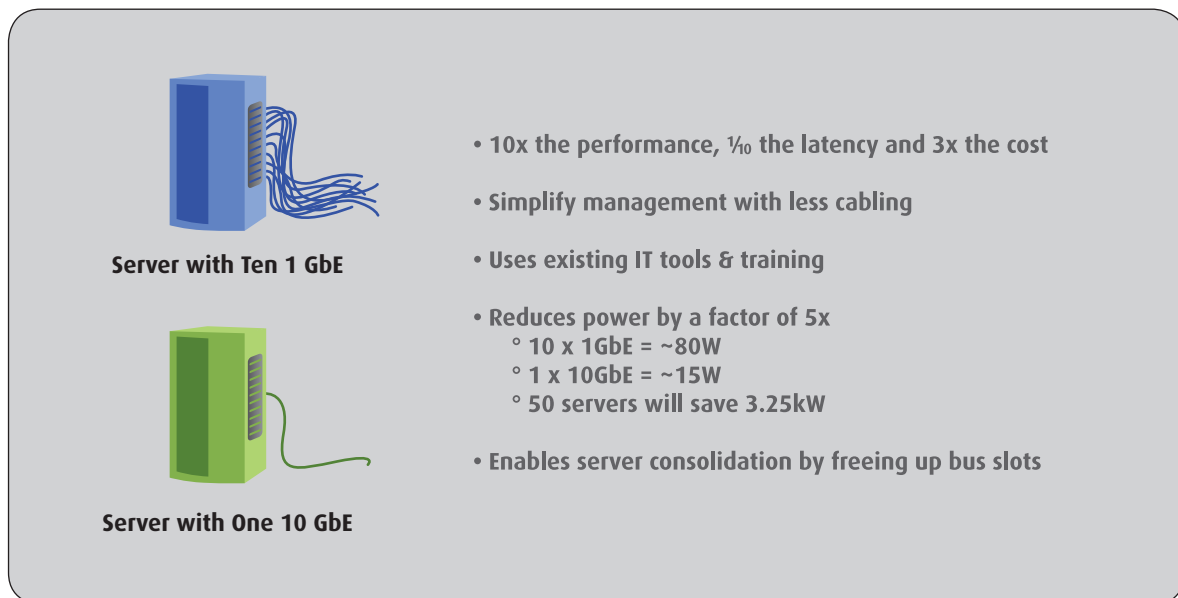


Figure 2 – Benefits of 10 GbE over GbE based servers

While it is theoretically possible to run VMware at Gigabit speeds, it is not recommended for running I/O-intensive applications or a larger number of VMs. The aggregate amount of bandwidth required by these applications running in Virtual Machines, as well as the need for speed in transfer operations like VMotion will simply choke the system. As a matter of fact, most system experts at VMware today recommend 10 Gigabit Ethernet in virtualized servers, if nothing else, to future-proof the installation. In short, if you choose to deploy virtualization, you should seriously consider 10 Gigabit Ethernet. The benefits of 10 GbE over Gigabit are summarized in Figure 2.

While upgrading a server to 10 GbE will provide significant relief on the I/O bottleneck, it is not sufficient in a virtualized environment.

IOV – The missing piece of virtualization

10 Gigabit Ethernet is a fat pipe. It will provide a faster pathway for the large datasets of any given application. But in a virtualized world, multiple VMs and applications compete for network access at a fast rate and through an additional virtualized software layer, called the “Hypervisor.” The Hypervisor is the gateway between the Virtual Machines (VMs) and the physical hardware. The more Virtual Machines per hardware platform, the more the Hypervisor is interrupted in order to handle I/O requests. And as the load increases, the system

becomes overtaxed and unable to sustain I/O performance. Functions need to be added at the I/O subsystem level to parse the traffic efficiently and preserve the Quality of Service (QoS) necessary for each application. This is where IOV comes into play.

What is IOV?

The basic principle of IOV is to share an I/O component among various compute elements, like Virtual Machines. Implementing IOV at the network level means making the interface appear to the host server like a multitude of independent interfaces.

The objective of IOV is to enable virtualized servers to perform the I/O functions as fast as non-virtualized servers, without sacrificing any of the benefits of virtualization. In order to do that, IOV architectures require truly independent I/O channels. This can only be accomplished by building separate hardware paths right through the silicon, inside the network controllers’ core structure. With an I/O architecture optimized for virtualization, 10 Gigabit Ethernet can significantly boost the network performance while making management simpler in a virtualized environment.

IOV at 10 Gbps speed

The fundamental benefit of combining IOV with 10 GbE is to increase I/O performance, while limiting CPU overhead, maintain QoS for all virtualized applications through isolation and

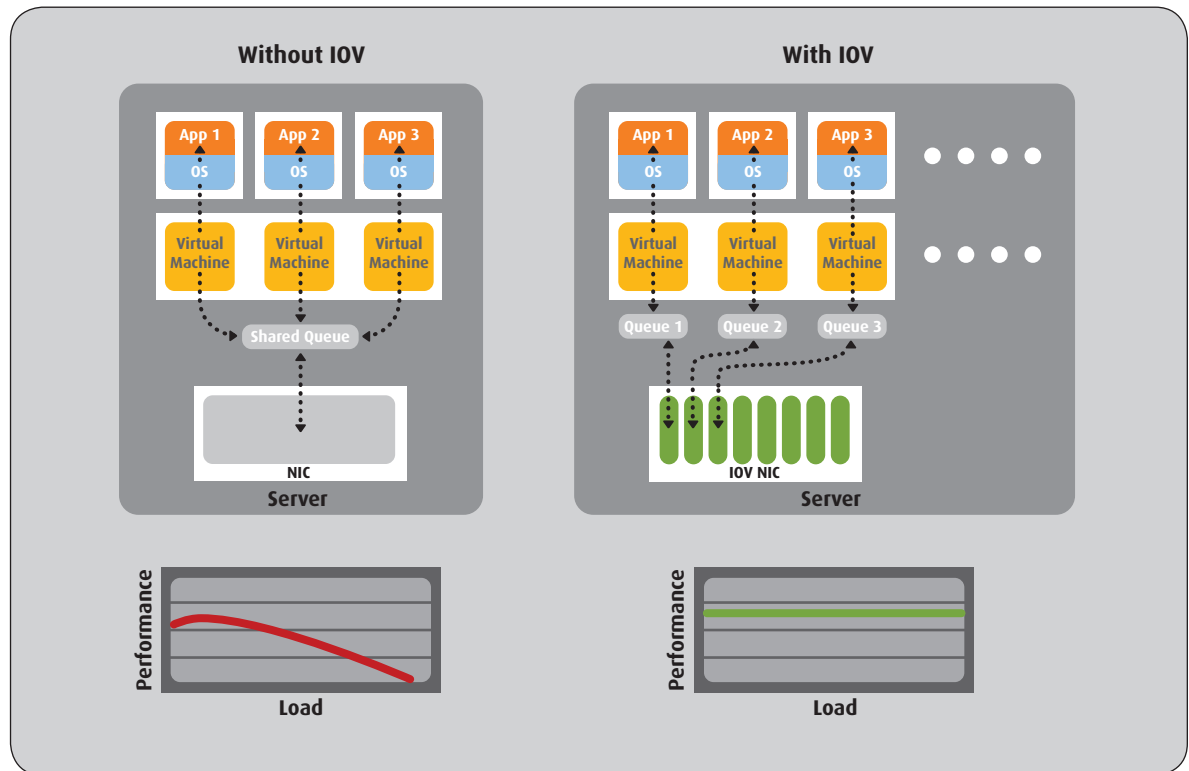


Figure 3 – IOV delivers higher performance for more virtualized applications

dynamic provisioning and increase the reliability of the virtualized server, even as the load increases. In short: enabling a significant increase of VM instances without compromise.

Let's take a look at an illustration, Figure 3. In a "non-IOV" situation, the 10 GbE adapter becomes a bottleneck as guest OS's contend through the Hypervisor for I/O access; the network performance will decrease as the load increases, which is not sustainable – especially if the VMs include I/O-intensive applications. On the right side, with an Ethernet adapter with IOV functionality, each VM and its guest OS can be assigned an independent I/O channel, which bypasses virtually all of the Hypervisor overhead and therefore removes the I/O bottleneck. This enables IT to support more VMs and to virtualize I/O-intensive applications, without impact to I/O or application performance.

INDUSTRY-LEADING IOV ARCHITECTURE

As mentioned previously, the mass adoption of virtualization in industry-standard environments can be seen both as a great opportunity and a challenge.

The opportunity has the enormous potential of associating low-cost, off-the-shelf standard server components from the x86 world with leading-edge, enterprise class virtualization software like VMware. But the challenge lies in the details of combining these two worlds. Now that virtualization is not the sole purview of mainframes anymore (which enjoyed highly integrated hardware and software,) it is absolutely key to carefully select I/O components which, without proper IOV support, will undermine the integrity of the Virtualized solution. The I/O subsystem is the crucial link forming the gateway to and from the server and integrating into the virtualization software.

An industry standard

Neterion is the first and only vendor to offer a 10 GbE adapter that complies with the industry-standard PCI-SIG Single-Root (SR) and Multi-Root (MR) IOV specifications. The I/O Virtualization working group of the PCI-SIG standard organization (Peripheral Component Interconnect – Special Interest Group) was created to address the I/O bottleneck in virtualized servers by extending the PCIe specification. The PCI-SIG IOV working group is co-chaired by IBM and HP

with significant contribution from Neterion, the only network I/O vendor to contribute.

The SR-IOV 1.0 specification allows multiple VMs in a Single-Root complex (host CPU, chipset and memory) to share a PCIe IOV endpoint without impact on performance. This specification covers the way in which I/O is configured and allocated as well as the handling of errors, events and interrupts. MR-IOV (1.0 to be released by mid 2008) provides Multi-Root complexes (independent processor systems, like blade servers or standalone rack servers) to share multiple PCIe IOV endpoints without impacting performance or reliability. To support such a topology, the I/O devices must support enhanced PCIe routing and separate registers for storing a number of hierarchies associated with multiple independent servers.

Major virtualization software vendors are expected to include SR-IOV support in upcoming revisions of their virtualization OS. Additionally a number of blade server vendors have endorsed the MR-IOV standard and are expected to release compliant products in the future.

All IOV architectures are not created equal

Neterion’s line of 10 GbE adapters offers a multi-channel design built directly in the silicon that is absolutely unique in the industry. It features totally independent hardware-based transmit and receive paths that can be reset independently, support true

QoS, and provide complete resource isolation for guest Virtual Machines.

Several 10 Gigabit Ethernet vendors have approached IOV the superficial way: in firmware rather than ASIC hardware. Neterion’s I/O Virtualization features are entirely wired-in hardware, to deliver the most I/O performance and isolation for virtualized servers.

Figure 4 illustrates the fundamental difference between firmware- and hardware-based architectures. Although firmware solutions separate channels to multiple guest VMs via a front-end, or superficial virtual path, the I/O is handled by a common back-end, usually implemented via one or more microprocessors and a lot of firmware. This doesn’t allow for true isolation and QoS on a per channel basis. From an isolation standpoint, if a single VM ‘hangs’ and stops queuing receive buffers to an adapter, this will eventually cause a superficially virtualized adapter to hang all of its channels; this is commonly referred to as a head-of-queue-blocking issue. On the other hand, with the Neterion IOV advantage, all channels are independent and therefore unaffected by the hung VM.

Similarly, in order to provide true QoS, true separation of the channels in hardware must be employed to allow different levels of bandwidth and latency guarantee. Additionally, when a firmware-based channel or VM needs to be reset or rebooted, all of the channels must be reset,

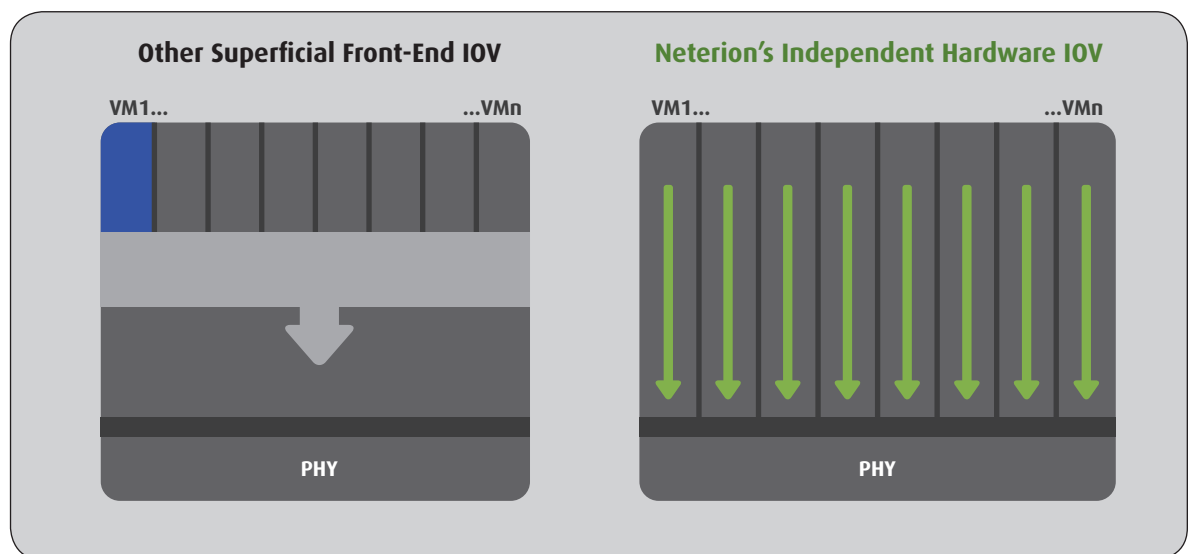


Figure 4 – Traditional firmware-based approach vs. Neterion’s hardware IOV architecture

Feature	Benefit
Cost	Replace 10 Gigabit interfaces with a single 10 Gigabit adapter, reducing complexity and administration costs
Flexibility	Dynamically allocate bandwidth across the various components of the system (different OS images or different blades in a chassis, or both) instead of limiting the bandwidth to a fixed amount per component
Performance	An IOV architecture (provided it has hardware assists) reduces the processor-intensive software layers required to pool or trunk the network interfaces together
Isolation	Every OS image is presented with its own independent I/O path; it is completely unaware that it is sharing resources with other OS images. It cannot harm if true hardware protection is implemented in the network interface
Reliability	With a second 10 Gigabit interface in the system, users can implement redundancy and fail-over capabilities, or double the size of the network pipe (20 Gigabit) to share among the components of the system. Optionally, a combination of fail-over and higher bandwidth, all dynamically managed and “virtualized,” is achievable

Table 2 – Benefits of Neterion’s IOV architecture

requiring everything configured on that adapter to be reinitialized. Such a superficial layer of multi-path is far from ideal and users will see real performance degradation and interruption of service during reset conditions.

In contrast, a true, hardware-based IOV architecture such as Neterion’s provides users with a superior solution, delivering flexibility and manageability for sustained performance. Its fully independent I/O channels can be reset or rebooted without interrupting operations running on other channels, ensuring that I/O Virtualization is ready for prime time. Each independent channel has separate receive and transmit paths, DMA engines and interrupts to deliver maximum VM I/O performance and isolation.

IOV provides dynamic provisioning for QoS

Virtualization at the I/O level also results in higher flexibility for IT managers, who can dynamically allocate the bandwidth across the various OS instances and applications, instead of being confined to a fixed stream for each. The bandwidth allocation process of a single 10 GbE connection across multi-channel I/O can follow a prioritization and QoS algorithm, for example by line of business, or time of day, etc., enhancing the experience for the end-users of these applications.

IOV enables isolation for reliability

IOV also ensures system reliability through hardware isolation. Isolation is critical from a protection perspective in that one application or operating system image cannot adversely affect the performance or reliability of another. One of virtualization’s primary strengths is that network components are unaware that their access to system resources has been virtualized. For example, an OS image assumes that it is the sole owner of an I/O port even though this port may be shared among several OS images across multiple CPUs or blades. To be truly isolated, each OS image must be able to treat each system resource as if the resource solely belongs to that OS image, even though these resources are actually shared. Such complete isolation is essential for easing deployment and management of new applications.

IOV and redundancy

Finally, it is always possible to add a second 10 Gigabit Ethernet port in the system to provide for redundancy and fail-over capabilities, or double the size of the network pipe (20 Gigabit), or a combination of both.

Table 2 summarizes the benefits of Neterion’s hardware-based IOV and 10 Gigabit Ethernet for virtualized environments.

Network support in VMware ESX Server

In the fall of 2007, VMware released ESX 3.5, bringing significant enhancements to support 10 Gigabit Ethernet, and alleviating network bandwidth bottlenecks in virtualized servers. These enhancements are:

- Update of the I/O Stack to be SMP/Multi-Core friendly (indispensable to support higher network throughputs)
- Reduction of memory copies
- Support of Network Offloading techniques: TCP Segmentation Offloads (TSO, also called LSO, Large Send Offloads), Checksum Offloads
- Support of Jumbo frames (packets larger than 1,500 bytes, up to 9,600 bytes or more)
- Addition of the NetQueue feature and its API for physical NICs:
 - Multiple receive queues, each associated with its own MAC address(es)
 - The MAC address(es) associated with each receive queue can be changed dynamically without resetting the entire pNIC
 - Each queue having one interrupt vector, which requires MSI-X support in the NIC
 - Each capable of being quiesced independently by the Hypervisor
 - Default receive queue for receiving packets that are not classified into any other receive queue
 - A configurable option to strip VLAN tag, if present, from a received packet's Ethernet header
 - Physical NIC assist to insert VLAN tag, with VLAN Id supplied out-of-band

VMware ESX NetQueue increases performance

One of the key new features in ESX 3.5 (and 3i) is **NetQueue**, a technology specifically designed for 10 Gigabit Ethernet. Shortly put, it allows Virtual Machines to scale the network traffic over multiple processors, improving the overall network

performance. It is important to note that ESX 3.5's networking enhancements require hardware support in the controller. Without the proper hardware, performance improvements are marginal.

Neterion Ethernet adapters are the only adapters currently certified to support VMware ESX 3.5's NetQueue technology. The two companies closely collaborated in this effort, as VMware selected Neterion as their development platform for the networking features of ESX 3.5. Together, NetQueue and Neterion's 10 GbE technology allows users to individually allocate bandwidth to each of the dedicated hardware channels, providing a complete end-to-end virtualization solution. These independent channels allow Virtual Machines to control the virtual paths as if they were their own independent adapters. Neterion's 10 GbE adapters allow one physical adapter to behave like multiple adapters and each can be easily managed, load-balanced and reset independently. A single channel can also support one or more VMs under the control of the system Hypervisor. Since each channel is essentially its own virtual adapter, the Hypervisor has the flexibility to reconfigure each channel independently as needed without interrupting operations running on the other hardware channels. With NetQueue, functions such as traffic classification can be offloaded from the software Hypervisor to the network controller hardware. Without NetQueue, the response time of applications running on VMs is impacted because of the large number of CPU cycles that must be dedicated to managing Ethernet traffic flow. This limits the number of VMs, and their associated guest OS's and applications that can effectively be run on a physical server. Neterion's support of NetQueue, in hardware, provides the I/O throughput and reduced CPU utilization that users require to capitalize on their server investment. This extends IT users' ability to do more with fewer physical resources.

As a result, the NetQueue-Neterion combination delivers outstanding performance and enterprise reliability. A benchmark recently developed

Neterion Ethernet adapters are the only ones currently certified to support VMware® ESX 3.5's NetQueue™ technology. The two work literally hand in hand, as VMware selected Neterion® as their development platform for the networking features of ESX 3.5.

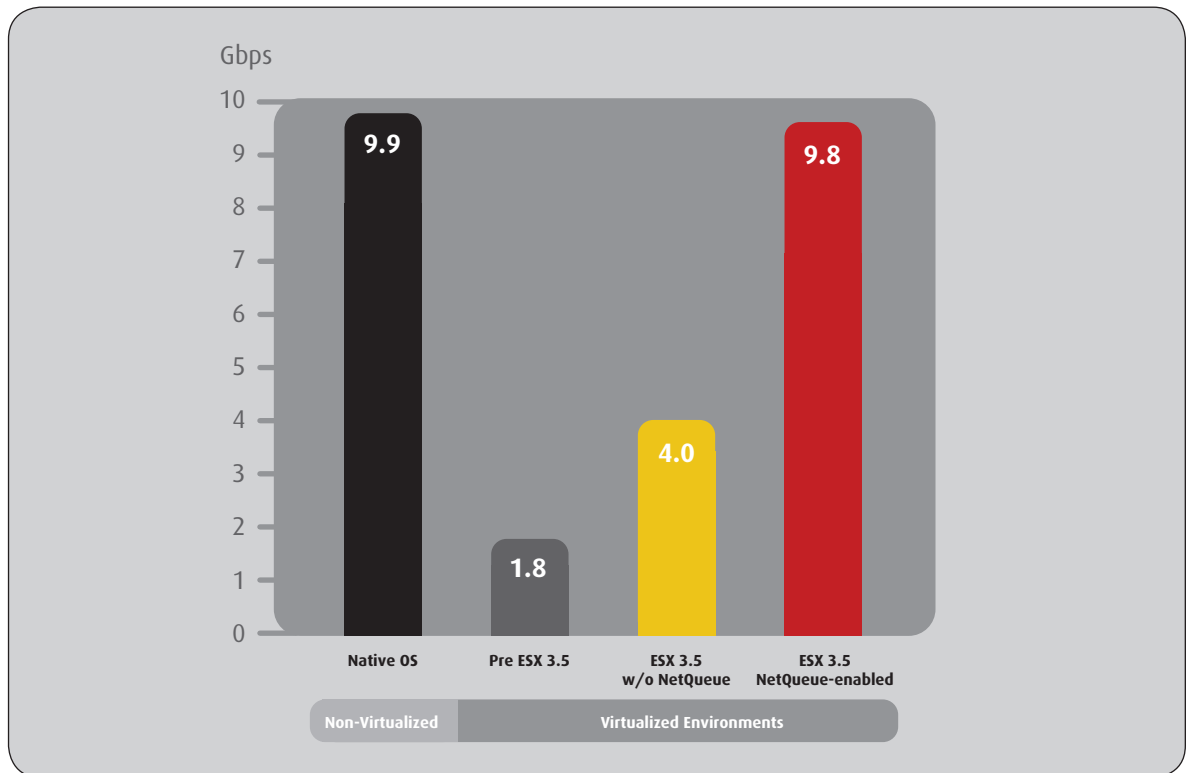


Figure 5 – 10 GbE I/O throughput in virtualized server and non-virtualized Servers

by VMware, IBM and Neterion (see Figure 5), demonstrated full line rate 10 Gbps speed on a virtualized server, for the first time in the industry. Prior to ESX 3.5 with NetQueue and Neterion's hardware-based IOV support, the virtualized OS maxed out its network I/O throughput at less than two gigabits per seconds. NetQueue represents a large scale development effort to remove that bottleneck. Today, 10 Gigabit Ethernet is as fast and transparent in VMware's ESX 3.5 as it has been in native, single-instance operating systems. This opens the door to virtualizing all the I/O-intensive applications that have been waiting to take advantage of the many benefits of Server Virtualization.

Virtualization without compromise

Today's advent of VMware's NetQueue, combined with Neterion's IOV architecture, is enabling a sharp increase in the number of VMs per system – by sheer mechanical effect of removing the I/O bottlenecks in the system. Neterion's 3rd generation of IOV Ethernet adapters also support Direct Hardware Access. When virtualized Operating Systems vendors include support for Direct Hardware Access this offers limitless possibilities. Direct Hardware Access

will further enhance throughput and latency, with linear scaling of network performance, bypassing the Hypervisor altogether. Neterion's IOV architecture, with fully independent hardware I/O paths built directly in silicon, is the only 10 GbE ready today for the Direct Hardware Access breakthrough.

Cost analysis

Beyond the technical benefits of IOV, the cornerstone of its value proposition is the savings it enables, both in upfront acquisition costs and in Total Cost of Ownership (TCO). This is best illustrated by a case study of a real-life situation.

An IT department is engaged in a server consolidation project, that will implement virtualization at some level. The initial state with no consolidation or virtualization of the compute resources includes 32 applications running on 32 physically different servers. Eight of these applications are "I/O intensive", which traditionally – in pre-IOV world – tags them as "bad candidates", excluding them from the virtualization project. However, if the IT department chooses to implement 10 Gigabit Ethernet with IOV technologies, these applications

could become “good candidates”. The following provides a comparison of two outcomes.

Solutions A: “Partially virtualized”

A partially virtualized solution, that would not use hardware-based IOV, looks as follows:

- Single virtualized server with 24 applications
 - Dual-socket quad-core x86 server
 - 24GB of RAM (1GB of RAM per VM)
 - 10 Gb Ethernet adapter (with no virtualization functionality enabled)
- Eight servers remaining untouched, running
 - native OS for the I/O intensive applications
 - Each with single-socket quad-core x86 processor
 - 4GB of RAM
 - Each server features dedicated Gigabit Ethernet connection

Solutions B: “Fully virtualized”

In a fully virtualized scenario, all of the 32 applications, including the 8 I/O-intensive, would be consolidated into 1 virtualized server, using the following configuration:
Single virtualized server with 32 applications

- Quad-socket quad-core x86 server
- 32GB of RAM (1Gb of RAM per VM)
- Neterion’s 10 GbE adapter

Note that this solution specifically requires a Neterion adapter, due to its unique ESX 3.5 NetQueue support and I/O virtualization features, to handle all applications’ I/O needs, including the performance sensitive ones.

Table 3 estimates a 30% reduction in acquisition cost of solutions B over solution A (\$20K out of a total \$65K). In addition to acquisition cost, the IT department will enjoy the TCO savings of the reduced maintenance, power, cooling and space required by 9 servers versus 1. Using TCO calculation methodology from VMware, this represents another 34% saving over a period of 3 years (\$84K out of \$250K).

The bottom line is simple: choosing the “Fully virtualized” option yields over 30% reduction in both upfront acquisition and long-term TCO costs over the “Partial” scenario. This savings is in addition to savings derived from implementing a virtualized environment and fully justifies rolling out a 10 Gigabit Ethernet based solution, using VMware’s and Neterion’s latest technologies offering enhanced I/O virtualization support.

Configuration	Solution A	Solution B	Difference
	Partially Virtualized Servers with other 10 GbE	Partially Virtualized Servers with Neterion’s 10 GbE	Savings
	9 Servers 1 x 2-socket with 24GB RAM + 8 1-socket with 4GB RAM	1 Server IBM x3850 with 32GB RAM	
Base systems cost	\$48,000 1 x \$22K + 8 x \$3K	\$32,000	\$16,000
VMware license	\$13,500	\$9,500	\$4,000
Windows OS license	\$3,500	\$3,500	\$0
Acquisition Cost	\$65,000	\$45,000	\$20,000
Datacenter Power	2,913W / \$6,205	637W / \$1,357	2,276W / \$4,848
Datacenter Cooling	3,645W / \$7,764	796W / \$1,695	2,849W / \$6,068
Space	\$7,750	\$3,100	\$4,650
Server Provisioning	\$10,454	\$116	\$10,338
Server Admin	\$217,725	\$159,665	\$58,060
TCO over 3 years	\$249,897	\$165,933	\$83,964

Table 3 – Cost comparison of solutions A and B

Summary

With I/O Virtualization, IT managers are now free to virtualize *virtually* all applications... The combination of new I/O technologies, fully backed by industry standards, and the ten-fold increase in bandwidth of 10 Gigabit Ethernet, gets even the most I/O-intensive applications prepared to reap the benefits of virtualization.

For proper implementation, IT managers must carefully choose the right IOV-optimized components. Like VMware's new ESX 3.5 release, with built-in NetQueue capabilities. It brings the I/O performance of virtualized servers to the same level as native, single-instance, operating systems. Adding 10 Gigabit Ethernet into the mix allows servers to run at 10 times the speed of older servers in native environments, which is clearly indispensable to future-proof IT infrastructures today. But to enjoy the highest levels of performance in virtualized environments, 10 GbE adapter technology with hardware IOV optimization must be employed.

Neterion's line of 10 GbE adapters includes hardware-based IOV functions that deliver maximum value. Offering fully independent hardware I/O paths built directly in silicon, Neterion is the only network adapter vendor currently certified to support VMware ESX 3.5's NetQueue technology. And with support for the industry-standard PCI-SIG Single-Root and Multi-Root IOV, plus Direct Hardware Access functions, Neterion enables a new generation of high performance data centers.

For the IT industry, virtualization is a game-changing, high return-on-investment opportunity. But like solving a puzzle, it requires all the elements to come together to be complete. I/O-intensive applications used to fall through the cracks of virtualization projects. Something was missing. It is now here, available through VMware and Neterion; it's called IOV – **the piece of the puzzle you have been missing.**