



PCIe Sharing: Breaking the “1 to 1” model without breaking transparency

Most data center administrators will agree that the issue of achieving higher performance, scale and virtualized server density is limited by lack of sufficient I/O and memory --CPU isn't often the bottleneck. Furthermore, administrators have over-provisioned I/O in fear that they will not be able to sufficiently support and supply the required to meet the service levels of their application.

For example, many administrators provide each server with its own dedicated 10 Gigabit Ethernet (10 GbE) network interface card (or two), a dedicated network switch port, and cable. This practice is quite common and creates a “static” 1 to 1 relationship between resource and server. One might challenge that this seems counterintuitive, as the de facto practice in today's data center to pool and share resources (compute, network, & storage) – via virtualization.

These types of I/O adapter resources are PCIe (PCI Express) based and allow servers to connect to high-speed networks and high performance storage. PCIe is a well-known, well-understood and universally accepted server-based technology, used to connect and access resources such as network cards, Fibre Channel host bus adapters (HBAs), RAID (redundant array of independent disk) controllers, and most recently PCIe-based SSDs (solid state drives).

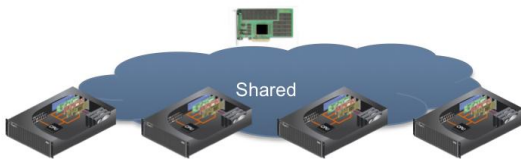
What Virtensys brings to the table is the ability to pool high value, in-demand I/O resources and share them across multiple servers, instead of dedicating a single resource to each server. This can be done without sacrificing the native capabilities of the I/O resource, retaining its initial value to the administrator, such as offloading checksum operations through hardware acceleration or improving how the card buffers information when sharing resources amongst virtualized servers.



1:1 Adapter to Server:

- Increased Management
- Poor Resource Utilization
- Expensive
- Fixed Performance

Sharing provides precisely the **capacity** and **performance** a server/application needs **when** it needs it



Shared PCIe:

- Centralized Management, Resource, & Data Access
- High Resource Utilization
- More Affordable
- Higher Performance

With Virtensys, servers have access to a virtualized instance of a traditional physical I/O resource (such as Intel's x520 10 GbE network controller). Sharing the physical I/O resource means that communications between connected servers can achieve higher performance, improved resource utilization and drastically lower their TCO.

Another interesting aspect of the shared resource model is the value of centralized provisioning and management. A server can now gain access to an I/O resource that it previously could not -- for economic or logistical reasons.

Imagine provisioning each server in your data center with its own physical PCIe-based SSD adapter; most would agree that this would be cost prohibitive. However, imagine pooling two or four of these same resources in a centralized appliance, and then sharing them as a virtualized PCIe-based SSD to multiple server hosts; now each has its own portion of the resource or depending on the configuration, all server hosts can see the resource as a shared pool, though connected via PCIe.

With this model, the performance benefit alone is a huge improvement, such as block copies within a shared logical device or datastore, to support operations within a clustered database or virtual machine live migration between server hosts. The operation can be achieved at local speeds/feeds as if the storage resource was directly attached and local to the server host itself.

Let's take a closer look at the concepts of PCIe sharing through a Virtensys PCIe sharing appliance, covering not only hardware connectivity (server to appliance to PCIe I/O adapter to upstream infrastructure), but also the methods of provisioning virtual adapters to a physical server.

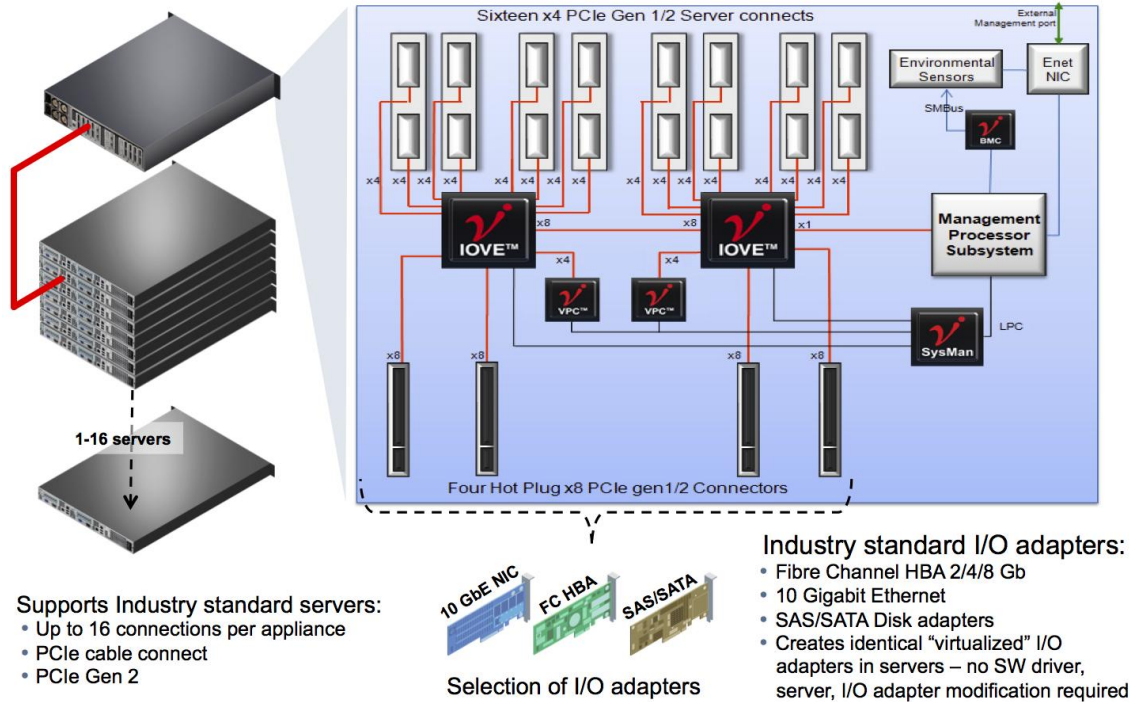
Architecture and Implementation

Rack mounted servers are physically connected to a Virtensys VIO-4000 series appliance through a resilient extension card, otherwise known as a VIO-RX. This card is a low power, low profile, PCIe x4 Gen 2 adapter that provides an active extension of the server's PCIe bus into the appliance. The key feature of this card is that it provides a stateful PCIe connection between the server and the appliance which prevents against cable faults such as a loss of PCIe cable connectivity. What this means is that in the event of a loss of PCIe connectivity, the server will not react as if a PCIe card was physically removed while the server was powered on. There will be a loss of network or storage connectivity, but not an operating system or hypervisor fault.



This resilient adapter card connects to the appliance through a standard PCIe x4 cable to a server connector card, a PCIe x8 dual port card, which provides connectivity to two servers. There are four server connector cards per "zone" in a VIO appliance, and each zone currently offers connectivity to two PCIe I/O adapters. This works out to eight servers for every two I/O adapters, for a total of sixteen servers connected to four I/O adapters.

Each zone in a VIO is comprised of a Virtensys I/O Virtualization Engine (IOVE) which provides a 16 port PCIe multi-root aware switch; a high speed switching fabric that gives 64 lanes up to 320Gb/s of non-blocking bandwidth, and is fully compliant to the PCI-SIG MR-IOV specification. The IOVE handles the separation of the control and data planes. This enables each I/O device to access multiple host memory spaces, effectively separating direct memory access (DMA) to data buffers and control structures.

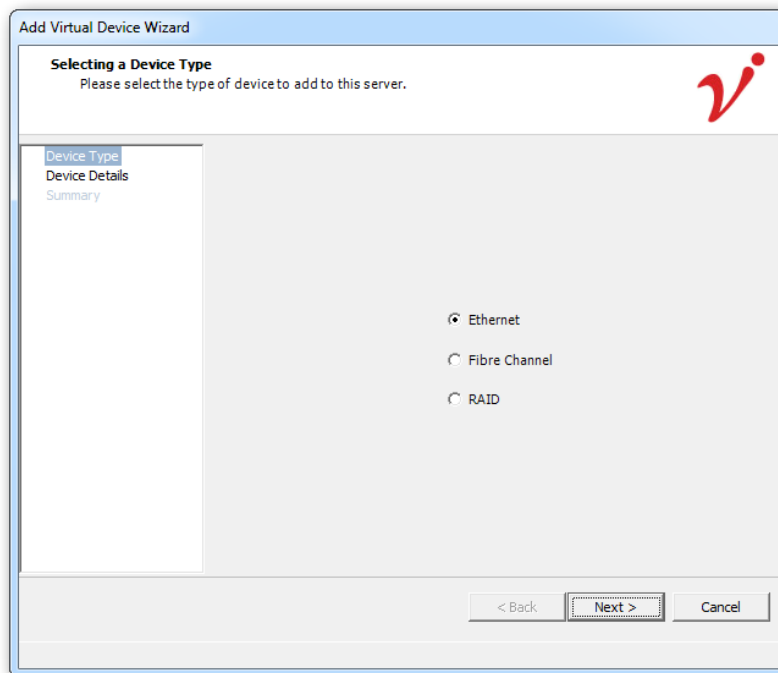


The “brain” of each zone is the Virtensys Virtual Proxy Controller (VPC). Each controller operates on the control path to provide the virtualization of multiple physical devices, and is a proxy between the I/O device and the connected server. The VPC can pre-fetch data from host memory in anticipation of a read operation from a device, or from device registers anticipating a read operation from a host. Since the VPC is implemented as a hardware state machine rather than through software and context switching, register access latency on a virtualized I/O device is comparable to the same access on a physical I/O device, and in some cases can actually improve the throughput of a virtual I/O device compared to a physical one.

When the virtualized version of a physical device is presented to a server, it’s as if the device is physically installed in the server. For example, a virtualized 10Gb Ethernet adapter will be seen as a PCIe device in the server’s BIOS, and a virtualized Fibre Channel HBA will go through its normal initialization processes during server boot – just like a physically installed adapter would. In addition, these virtualized adapters do not require any additional Virtensys-proprietary drivers nor is it necessary to boot into the host’s operating system or hypervisor to be able to operate the adapters, which makes features such as PXE (preboot execution) or boot from SAN a possibility.

Management Approach

Each VIO-4000 series appliance comes with its own web-based interface allowing for the pairing of physical servers with virtual adapters. While this functionality is necessary and important, a true strength of the VIO-4000 series is the VMware vCenter™ Plugin for Virtensys and its companion, a comprehensive PowerShell module. The VMware vCenter Plugin for Virtensys is designed to administer multiple VIO appliances (of varying model types) simultaneously within a virtual data center. Each appliance can be individually configured to provide various combinations of 10GbE, 8 Gbps Fibre Channel, or RAID (depending on appliance). This plug-in takes its place alongside other vendor specific plug-ins that offer a unified view of appliance management and VMware hosts/guests.

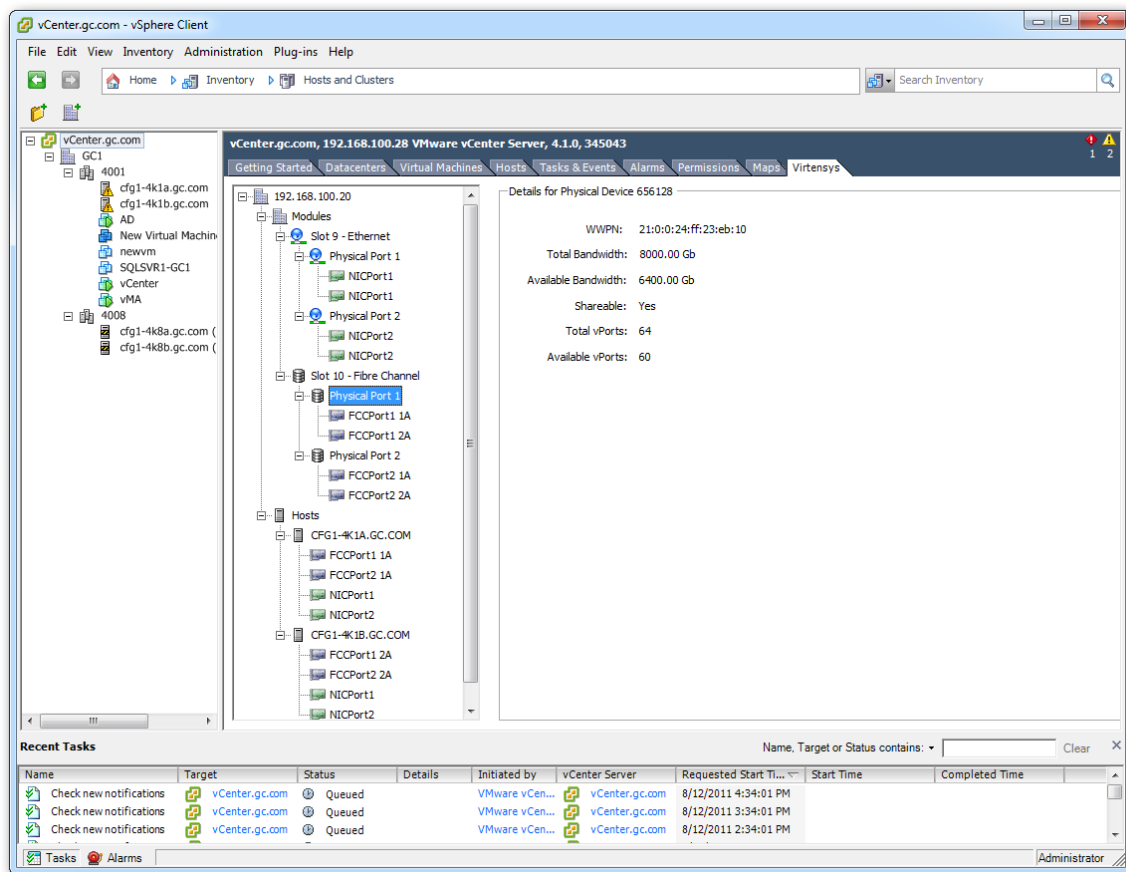


The Virtensys PowerShell module is designed to administer one appliance at a time, but its true strength is offering a fast and convenient means of adding or removing servers, virtualized adapters, or adjusting bandwidth according to scheduled events in the datacenter such as a backup window.

All three management points offer a comprehensive view into the managed network and fabric identities of the virtualized adapters, and the ability to migrate I/O profiles between VIO appliances.

Provisioning and Using vAdapters

The process of provisioning virtualized adapters to servers was designed to be simple and straightforward. A server connection is defined (typically using its fully-qualified domain name), and a physical adapter is chosen. From that pairing a virtualized adapter is created. MAC addresses or WWPNs are generated from and reside on the VIO-4000 series appliance and, in the case of 10GbE, a minimum bandwidth allocation is defined; the default is 10%, or 1Gb. Additional burst bandwidth is available on-demand, but it will never be less than the minimum allocated. Minimum bandwidth allocations can also be adjusted on the fly as the need arises.



For virtualized RAID, the process is similar; the physical adapter is configured to provide the desired RAID level and then divided by eight, the maximum number of servers available per zone. When a server is paired with a LUN, the administrator is given the option to make it bootable.

Servers provisioned with virtual adapters will “see” and begin to use these adapters as soon as the machine is booted. To the server, operating system, or hypervisor, virtualized adapters appear no different than a physically installed adapter would. VMware vSphere 4, as an example, will recognize

virtual 10Gb Ethernet adapters as standard vmnics. All of the functionality offered by vSphere to a physical adapter is identical to what is available to a virtual adapter; features such as multipath iSCSI, virtual switch (VST) and external switch VLAN tagging (EST), load-based teaming, vMotion over 10GbE uplinks, Fault Tolerance, and more, are supported. In the case of virtualized 8Gb Fibre Channel HBAs, they are also treated as standard vmhbas, and fully support different multipathing options like Round Robin, Most Recently Used, and Fixed. Virtualized storage appear as local datastores, and in the case of VMware vSphere 5 any virtualized PCIe SSD or SATA SSD LUNs are fully recognized as SSD and new features such as SSD host-based caching are fully supported.

Summary

For Virtensys the value of its intellectual property is in the ability to share PCIe-based adapters through the use of a high performance PCIe-based switch fabric and hardware virtualization layer. I/O virtualization is merely a “use case” of the Virtensys technology. In fact, Virtensys isn’t fixated on a specific I/O connectivity technology winning the battle in the virtual data center (example Fibre Channel vs. 10 GbE). Instead Virtensys is focused on providing greater access to a diversity of PCIe based resources (NICs, HBAs, RAID, PCIe SSDs, and other types of data center peripherals) through the use of its unique technology.

Sharing in the data center has been proven technology approach for decades, seen as a better way to achieve scale, resource utilization and even improve performance. As we expand beyond what we know can be shared (server, network, and storage) to other areas of the data center, we realize that it is prime time to also leverage sharing for PCIe based resources.