Broadcom Ethernet Network Controller Enhanced Virtualization Functionality

A Dell Technical White Paper

Third party information brought to you, courtesy of Dell.

July 2010

Contents

Virtualization Overview 2
Challenges in Virtualization 2
Virtualization Phase 1 2
Virtualization Phase 2 7
Summary 9
References 9

Figures

Figure 1. Multiple Queues in Virtual Environment 3
Figure 2. VMware Static VMDirectPath and TCP Offload in VM 4
Figure 3. TCP Offload and Windows 7 Hyper-V 5
Figure 4. iSCSI HBA in Virtual Environment 6
Figure 5. SR-IOV in a Virtual Environment 7
Figure 6. VMware Dynamic VMDirect Path 8

"Advancements in VMware® virtualization technology coupled with the increasing processing capability of hardware platforms are driving higher server consolidation ratios in data centers. To complement this trend, Broadcom's innovation in high bandwidth I/O subsystems and network fabric convergence can add value for our customers by pushing the current boundaries of performance and functionality," Brian Byun, Vice President of Global Partners and Solutions at VMware[1]

Virtualization Overview

Networking is an essential component inside a virtualization environment. Similar to other hardware resources found in the system, a network adapter is virtualized to the Virtual Machine (VM). VM vendors utilize hypervisor, also called virtual machine monitor (VMM), based architecture that hides the physical characteristics of the computing platform and allows unmodified VM to run concurrently on the host.

Virtualization presents many advantages, including the ability to enable users to consolidate computing hardware resources and allow them to run multiple virtual machines concurrently on consolidated hardware. This allows live migration of virtual machines from one physical server to another with zero downtime and continuous service availability. Virtualization also provides the user a rich set of features, I/O sharing, consolidation, isolation, and mobility along with simplified management with provisions for teaming and failover.

Challenges in Virtualization

Virtualization comes at a cost of reduced performance due to hypervisor architecture. Today's virtualization architecture includes VM with device driver, I/O stack, and applications layered on top of a Virtualization layer that includes device emulation, I/O stack, and physical device driver managing the Ethernet network controller. This additional virtualization layer adds overhead and degrades system performance by requiring higher CPU utilization and reducing bandwidth.

Virtualization Phase 1

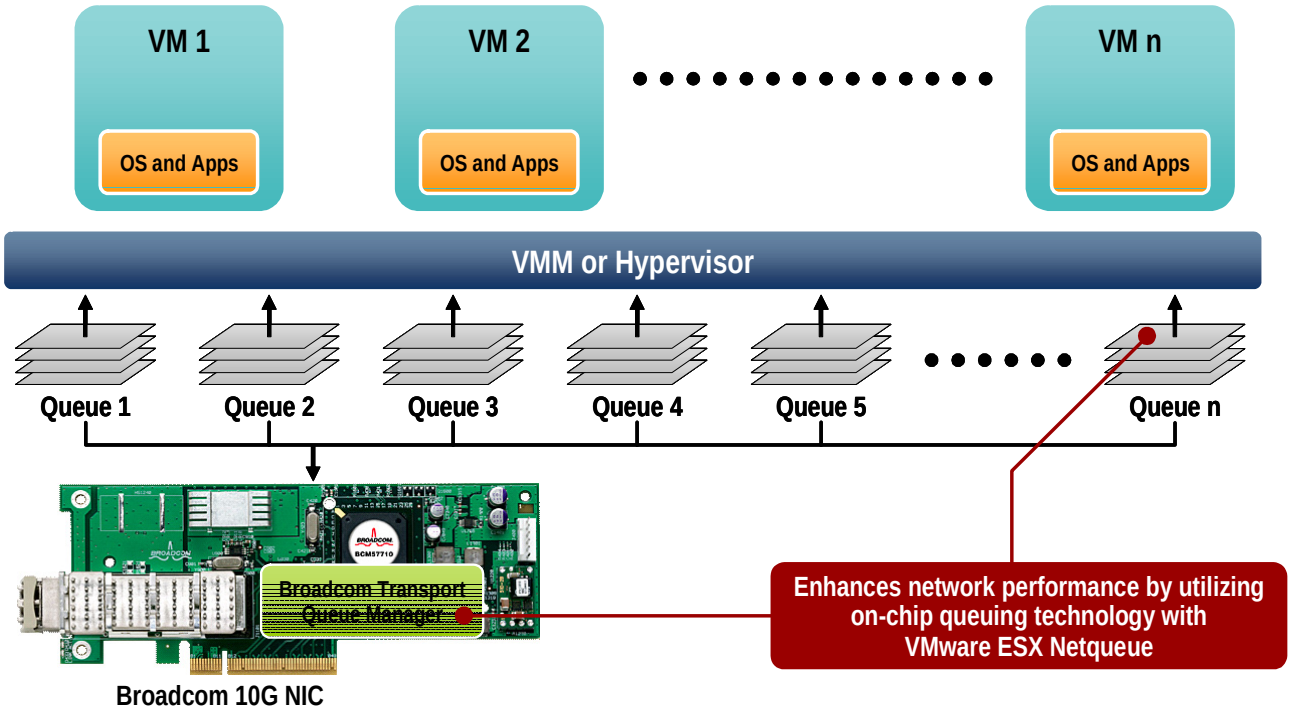
In the first phase of enhanced virtualization of Ethernet network controllers, Broadcom has closely partnered and collaborated with various VM vendors, VMware®, Microsoft®, and Xen® to remove some of the virtualization bottlenecks and improve system performance by providing additional features. Broadcom Ethernet network controllers support stateless offloads such as TCP Check Sum Offload (CSO), which enables network adapters to compute TCP checksum on transmit and receive, and TCP Large Send Offload (LSO), which allows the TCP layer to build a TCP message up to 64 KB long and send it in one call down the stack through IP and the Ethernet device driver, saving the host CPU from having to compute the checksum in a virtual environment. Jumbo frame support in virtual environments also saves CPU utilization due to interrupt reduction and increases throughput by allowing the system to concentrate on the data in the frames, instead of the frames around the data. However, performance is still limited due to the single-threaded nature of hypervisor in processing I/O and duplicate I/O copies in the virtualization layer. NetQueue support in VMware and VMQ support on Microsoft Hyper-V removes single queue bottlenecks and the use of statefull offloads such as TCP offload. iSCSI hardware-based acceleration in virtual environments is proven to provide excellent performance on VM.

[1] Excerpt from Broadcom VMWorld 2008 press release
http://www.broadcom.com/press/release.php?id=1197764

Virtualization Phase 1: Usage of Multiple Queues

The current hardware trend of increased processor core density is leading to an increased number of VMs requiring more CPU cycles to route packets to the VMs. For example, it is common today to expect to have quad core processors on each blade. That indicates 2^3 blades * 2^2 threads = 2^5 threads per chassis. By utilizing the hardware queues provided by the network controller, VM vendors have eliminated the single thread limitation in a traditional OS and have optimized the hypervisor for multiple hardware threads.

Figure 1. Multiple Queues in Virtual Environment



In both VMware and Microsoft Hyper-V, packets must traverse the hypervisor or parent partition since there is no direct path between the NIC and VMs. On the egress, packets are first copied from the originating VM for vSwitch processing. The destination MAC address and VLAN ID are looked up to determine the route. Depending on the result of route lookup, the packet can be copied to the receive queue of the other VMs and/or be submitted to the network driver for transmission. On ingress, packets are indicated to the switch, which uses the destination MAC address and VLAN ID to determine which VM or group of VMs the packets can be copied to.

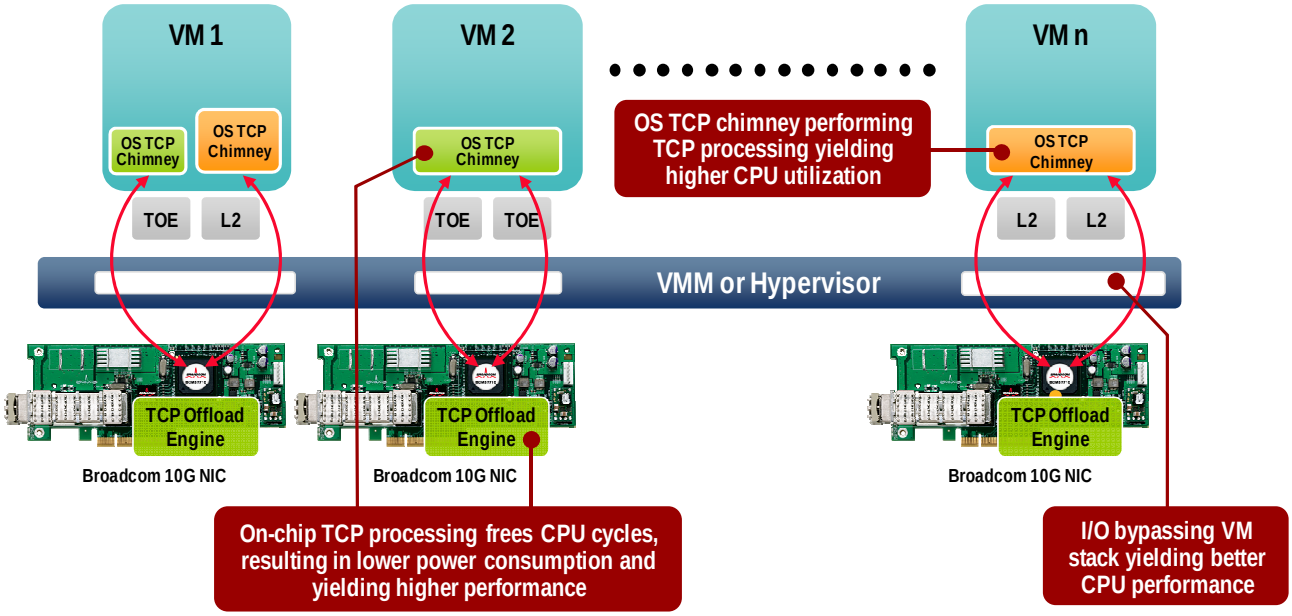
Route lookup, data copy, and filtering tasks represent additional CPU load and latency that are absent in the non-virtualized environment. The resultant overhead can significantly impact networking performance, especially at 10 Gbps. The first attempt to address these problems is to offload these tasks into a network adapter, where the transport queue manager can transmit packets from multiple queues and steer the receive packets into multiple queues. By dedicating TX/RX queue pair to a VM, the network adapter can get directly to and from the VM's memory, and vSwitch will only process the control plane operation.

NetQueue in VMware and VMQ on Microsoft Hyper-V enables Broadcom NetXtreme® I and NetXtreme II® Ethernet controllers to provide extended performance benefits that meet the demands of bandwidth intensive applications requiring high performance and higher networking throughput in a virtual environment

Virtualization Phase 1: Networking Offload

High I/O performance is required for enterprise class applications. However, the additional virtualization layer adds overhead and degrades I/O performance for additional virtualization benefits. For example, receive side processing in a virtual environment includes three I/O copies. A network I/O packet received by the network controller is copied into the receive buffers owned by hypervisor and raises the physical interrupt. Hypervisor parses the I/O packet to find a destination, copies the I/O packet into the VM receive queues, and raises the virtual NIC interrupt. The VM then copies the I/O packet to the application buffer.

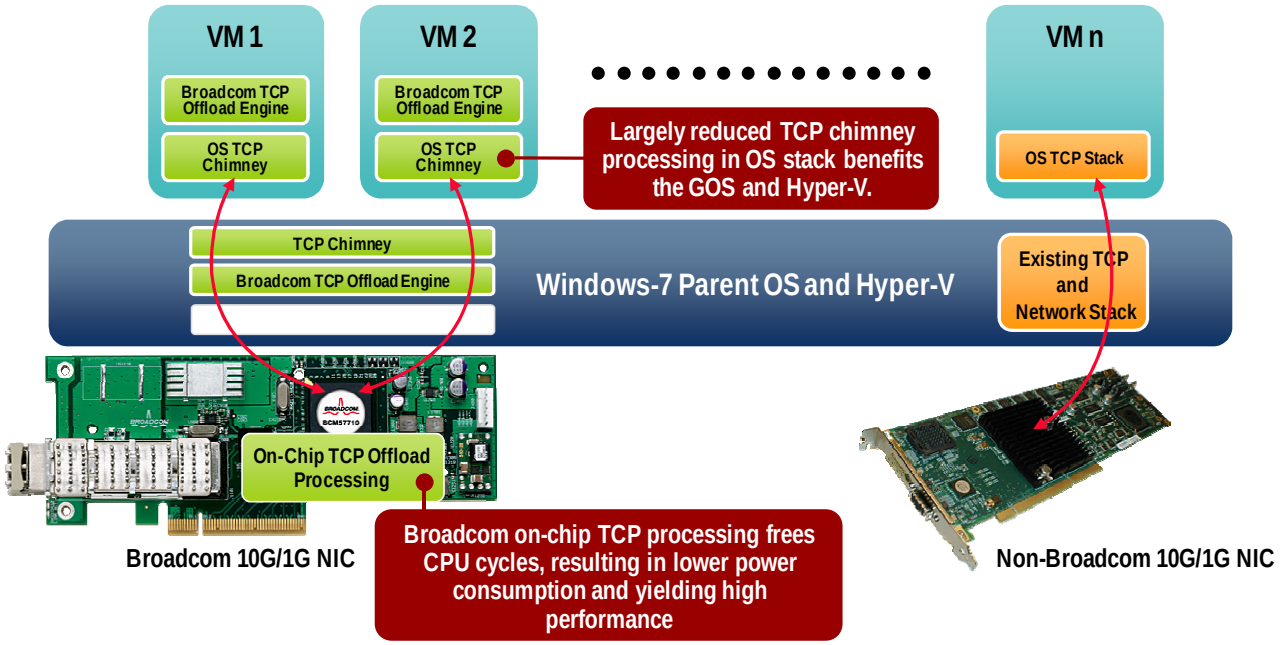
Figure 2. VMware Static VMDirectPath and TCP Offload in VM



VMware Static, VMDirectPath, or Fixed Pass-Through (FPT) enable Broadcom NetXtreme II C-NICs to be completely dedicated to high performance VM. FPT utilizes AMD IOMMU or Intel® VT-D to directly DMA data I/O from the physical device to the VM, bypassing the virtualization layer, and thereby yielding complete physical access of the dedicated Broadcom CNIC to the VM. FPT enables on-chip processing support in the C-NIC to be exposed and function within the VM. By functioning within the VM, Broadcom's C-NICs provide better throughput for the VM, as well as better use of CPU resources since the hypervisor no longer needs to process each network I/O request. In addition, VMDirectPath support allows the user to benefit from on-chip TCP offload processing in a VM, which allows server customers to realize improved networking performance.

Microsoft Hyper-V has extended its TCP offload support for VMs. TCP/IP traffic in a VM can be offloaded to a physical NIC on the host computer. TCP connection offload capability is available in the VM. Virtual NIC in the VM advertises connection offload capabilities, and VM switch offloads VM connections to the NIC.

Figure 3. TCP Offload and Windows 7 Hyper-V



TCP chimney on Windows Hyper-V saves two copies for the I/O packet on the receive and transmit path in a virtual environment. It bypasses excessive chatter between the parent partition and VM by avoiding the overhead of parent partition and VM context switch and maintaining connection state in the network controller. There is the added benefit of being able to support live migration by uploading connections to the host stack.

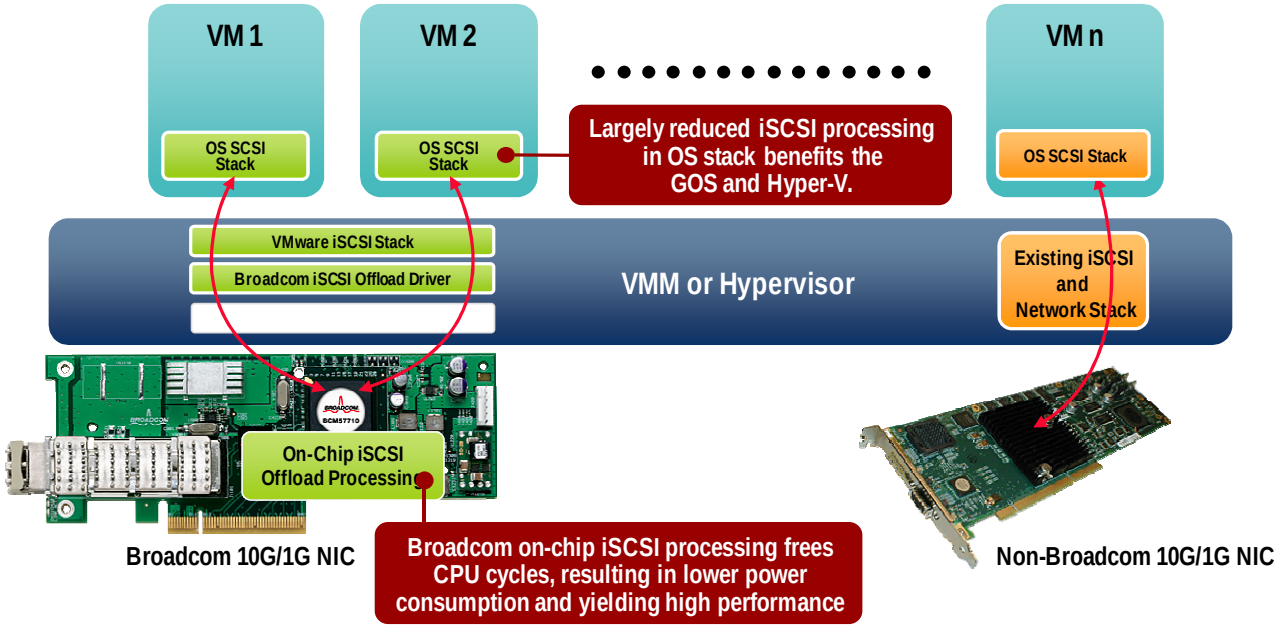
The Broadcom TCP offload engine enables on-chip processing of the TCP protocol, which frees up host CPU resources at line rate and provides impressive performance on Windows 7 Hyper-V. Freed-up CPU cycles can be provisioned to the VM to run high performance applications. It provides significant improvement over non-TOE Ethernet NICs with TCP segmentation. Hardware-based TCP offload provides more efficiency and better performance than software-based TCP offload.

Virtualization Phase 1: Storage Offload

Networked storage is crucial in virtual server environments as it provides for a smooth migration and failover of a VM from one physical server to another. iSCSI has emerged as a high-performance, easy-to-manage, networked storage technology that is popular in many virtualization deployments. As server and enterprise application customers strive to achieve density and computer resource utilization objectives for their servers and enterprise applications, Broadcom's NetXtreme II iSCSI Hardware-Based Acceleration (HBA) functionality, with support for VMware, Xen, and Microsoft Hyper-V virtualization, provides the converged functionality needed in a virtualized server environment by offering complete on-chip processing solutions that free up CPU resources and increase bandwidth and performance.

Enterprise networks are well positioned to take advantage of iSCSI functionality since it is built on top of the highly familiar TCP/IP protocols and because of the ubiquity and effectiveness of Ethernet, which allows storage content to be accessed from an Ethernet fabric. The increase from 1 Gbps Ethernet to 10 Gbps Ethernet delivers increased storage performance levels not previously achievable and provides sufficient bandwidth that permits multiple types of high bandwidth protocol traffic to coexist on the same network. As a result, a server converges networking and storage onto the same network while lowering the total cost of ownership (TCO), or uses a dedicated network for data and for storage, thereby using the same equipment for multiple purposes.

Figure 4. iSCSI HBA in Virtual Environment



Broadcom iSCSI HBA functionality enables on-chip processing of the iSCSI protocol (as well as TCP and IP protocols), which frees up host CPU resources at 10 Gbps line rates over a single Ethernet port. This functionality provides extended performance benefits that meet the demands of bandwidth intensive applications requiring high performance block storage I/O for the hypervisor, servicing all instances of the VM.

Broadcom 10G and 1G iSCSI HBA functionality as described in the figure is enabled in VMware VSphere 4.1 for Dell PowerEdge solutions by default.

Virtualization Phase 1: iSCSI Boot

iSCSI boot is a powerful technology because it allows a server to boot an Operating System (OS) over a storage area network (SAN), completely eliminating the need for local disk storage (which is the number one source of failures in computer systems). In addition to this enhanced system reliability, the use of diskless servers simplifies the IT administrator's workload by centralizing the creation, distribution, and maintenance of server images, reducing the overall need for storage capacity through increased disk capacity utilization and adding increased data redundancy through the use of data mirroring and replication. iSCSI boot is an easy way to boot to a virtual operating system over a SAN.

Broadcom 10G and 1G iSCSI Boot is enabled in VMware VSphere 4.1 for Dell PowerEdge solutions by default.

As the use of Storage Area Networks continues to grow in a virtual environment, and the benefits of moving local storage from individual servers to centrally managed storage arrays are realized by IT administrators, network boot solutions such as iSCSI boot will become a more common feature within the data center and throughout the enterprise. Broadcom, VMware, Xen, and Microsoft Hyper-V are evaluating and coordinating to bring a simple-to-configure yet richly featured iSCSI boot solution, allowing iSCSI to completely replace all forms of local storage in a virtual environment.

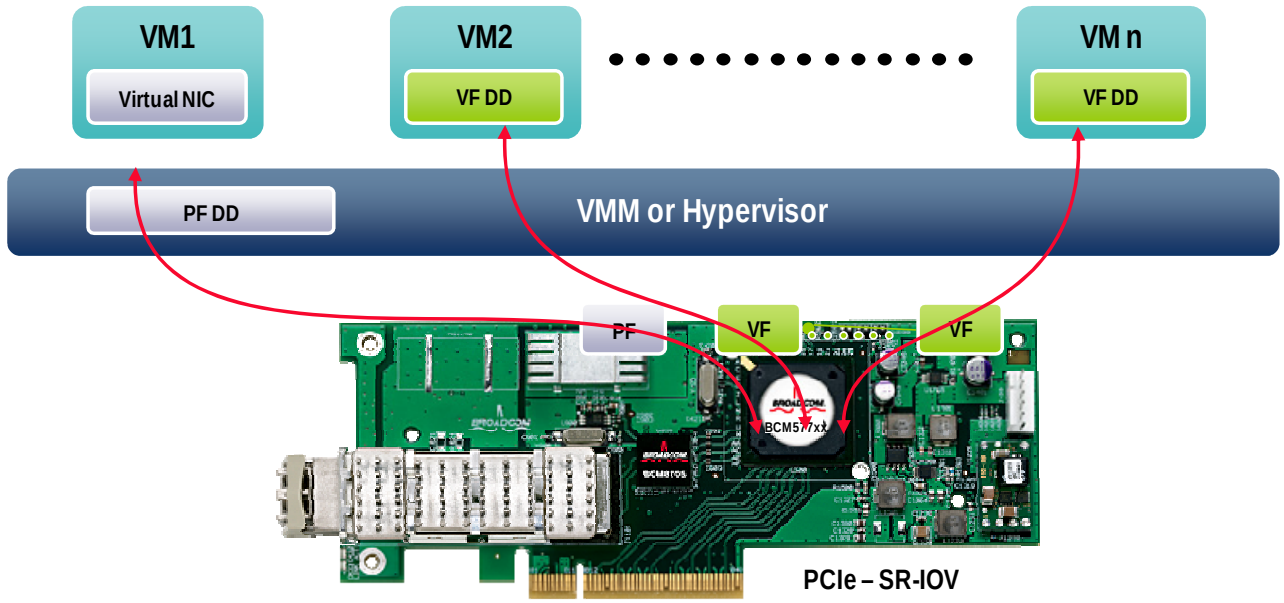
Virtualization Phase 2

In the second phase of enhanced virtualization of Ethernet network controllers, Broadcom has closely partnered and collaborated with various VM vendors, VMware, Microsoft, and Xen for Single-root I/O Virtualization (SR-IOV). SR-IOV capable Ethernet network controllers obtain benefits of I/O throughput and reduce CPU utilization while increasing scalability and sharing capabilities of the device. SR-IOV allows the direct I/O assignment of the Ethernet network controller to multiple Virtual Machines maximizing full bandwidth potential of the network adapter.

Virtualization Phase 2: SR-IOV

PCI Express® (PCIe™) SR-IOV provides guidelines on PCI I/O Virtualization and sharing technology. This specification is the basis for SR-IOV implementation of Broadcom SR-IOV capable Ethernet network controllers.

Figure 5. SR-IOV in a Virtual Environment



The PCIe SR-IOV specification defines an extension to the PCIe specification to enable multiple system images (SI) or VM to share PCIe hardware resources. The Broadcom SR-IOV device presents a Physical Function (PF) with multiple Virtual Functions (VF). A VF is a light weight PCIe function, and resources associated with the main data movement of the function are available to the VM. The VF can be serially shared between different

VMs, that is, a VF can be assigned to one VM and then reset and assigned to another VM. Also, a VF can be migrated from VF to another PF.

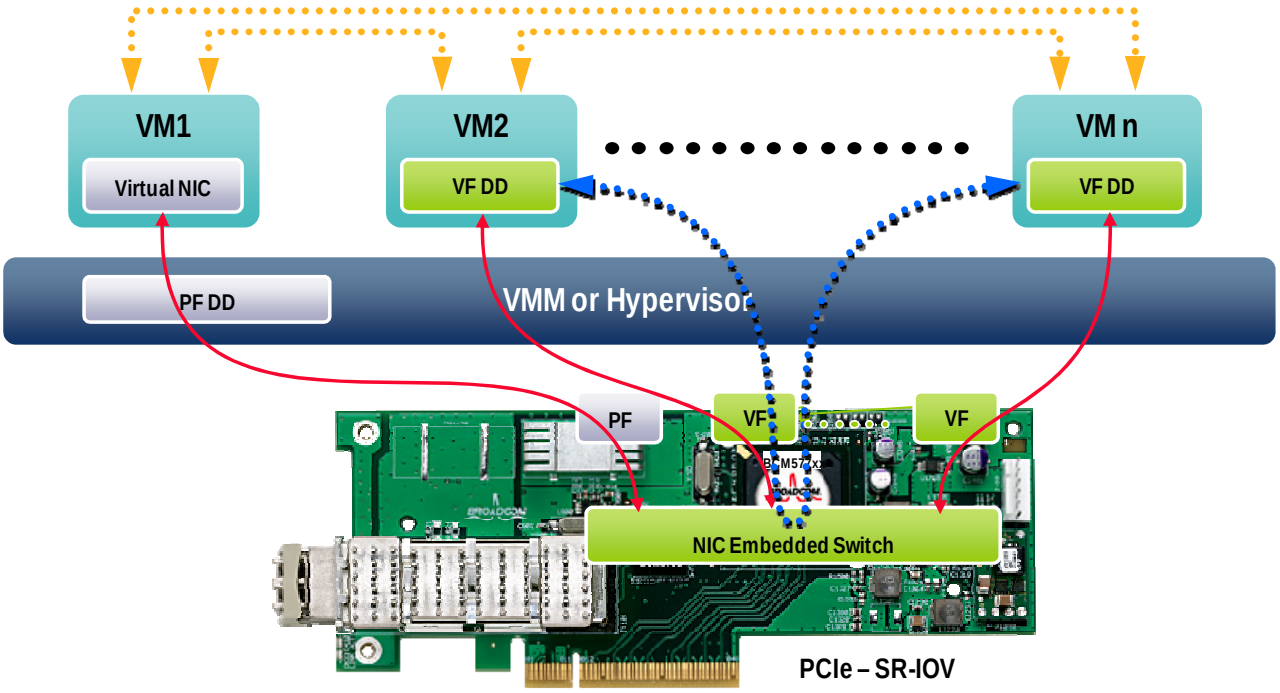
Complete support and realization of the benefits of PCIe SR-IOV involves enhancing and adding new capabilities to the entire platform and OS. Network controller device drivers that support SR-IOV will also need to be rearchitected to support new communication paths between the physical functions and virtual functions it supports.

Virtualization Phase 2: Dynamic VMware Direct Path

High-throughput and low latency are especially important in a distributed system, where the I/O latency of each node impacts both cluster and overall application performance. Low latency is required in order to preserve data coherency in large database clusters implementing scalable SR-IOV network adapters.

With VMware Direct Path network plug-in architecture and a Broadcom SR-IOV device, a VF can be directly assigned to a VM. This yields native performance and eliminates additional I/O copy in the hypervisor with the added advantage of being able to support all virtualization features including live migration or vMotion for VMware. Direct assignment of PCI devices to VM will be necessary for I/O appliances and high performance VMs.

Figure 6. VMware Dynamic VMDirect Path



With the Dynamic VMDirectPath or Uniform Pass-Through (UPTv2), the device interface is split into two parts. This enables pass-through of performance critical operations such as TX/RX producer index registers, interrupt mask register, and emulated infrequent operations in management driver running in ESX. In order to implement live migration, the VF is acquiesced and switched to emulation mode from pass-through mode, which will allow minimal device state to be checkpointed or restored. Most of the state lives in the VM memory and the guest is not aware of migration, and the migration completes flawlessly.

Support for Dynamic VMDirectPath requires a rearchitecture of the OS platform and network device driver. VMware implements a network plug-in architecture that allows pass through of performance critical parts, by partitioning VMxnet driver to include a VM-specific shell and hardware-specific module or network plug-in driver. A VM-specific shell implements the interface to the OS network stack and interacts with the hypervisor for configuration. A hardware-specific network plug-in driver interacts with hardware in the data path and uses the VM shell interface for all OS-specific calls. VMware ESX controls the network plug-in used by the shell to load the plug-in into the VM based on VF and maps the VF into VM address space.

Virtualization Phase 2: NIC Embedded Switch

I/O virtualization and sharing are also required for point-to-point and switch-based configurations. I/O virtualization and sharing will enable interoperability between VMs, VFs, chipsets, switches, end points, and bridges. A Broadcom NIC embedded switch built into the network controller enables Ethernet switching between VMs, or VF to VF, and to or from an external port.

Summary

IT needs drive convergence, higher bandwidth, and new standards enable convergence. Convergence is Virtualization's natural friend. Convergence addresses the need for flexible and dynamic I/O, multiple and variable different workloads, higher network and storage utilization, and centralized storage. Convergence minimizes the number of devices to control and to migrate.

Convergence with offload technologies and real-time flexible I/O is a necessity for Virtualization. Specialized Ethernet network controllers can significantly reduce power consumption. SR-IOV and I/O pass-through functionality for Ethernet network controllers with TCP offload engine and iSCSI HBA will provide near-native performance and reduced latency. Trends for networking and storage convergence will further accelerate the rate of virtualization adoption, reducing costs and saving power.

Dhiraj Sehgal is a senior product line manager for Ethernet controllers at Broadcom.

Abhijit Aswath is a senior product line manager for Ethernet controller software at Broadcom.

Srinivas Thodati is a senior product marketing manager for PowerEdge M-Series servers at Dell.

References

SR-IOV Networking in Xen: Architecture, Design and Implementation

http://www.usenix.org/events/wiov08/tech/full_papers/dong/dong.pdf

Single Root I/O Virtualization and Sharing Specification Revision 1.0