



SERVER CONNECTIVITY

Brocade 1860 Fabric Adapter I/O Virtualization and Virtual Switching

To keep pace with dynamic business requirements, organizations are transitioning their data centers to private cloud architectures to enable them to consolidate, scale, simplify, and automate their IT resources to increase business agility while reducing capital and operational expenditures. To help address the new demands derived from server virtualization, the Brocade® 1860 Fabric Adapter simplifies and optimizes I/O in virtualized environments.

BROCADE

CONTENTS

Overview	3
Introducing the Brocade 1860 Fabric Adapter	3
Brocade AnyIO Technology	4
Unmatched Flexibility and Performance	5
Brocade vFLink I/O Virtualization (IOV)	5
Improving I/O Performance for Virtual Machines	6
Direct I/O	6
Single-Root I/O Virtualization (SR-IOV)	8
Virtual Switching	9
Virtual Machine Optimized Ports (VMOPs)	9
Hardware-Based Virtual Ethernet Bridge (VEB)	10
Virtual Ethernet Port Aggregator (VEPA)	10
Edge Virtual Bridging (802.1Qbg)	10
Summary	12

OVERVIEW

With the primary goal to consolidate servers and make a more efficient use of their resources, such as CPU and memory, organizations have been deploying server virtualization for several years. Although initially implemented at a small scale, and mainly for test and pre-production purposes, over the last few years organizations have started to deploy virtualization to higher degrees, increasing virtual machine (VM) densities and virtualizing more critical applications. This has been possible due to the ever-increasing CPU speeds and memory capacities available from server vendors, as well as to new technologies in Intel and AMD server chipsets to improve performance for virtual workloads.

However, business requirements are more dynamic than ever in today's globalized economy, and organizations are looking for ways to achieve more business agility without sacrificing service levels. As such, they are looking to evolve their IT architectures towards the notion of private cloud, where pools of resources—compute, storage, and network—can be dynamically provisioned on-demand to respond to such changing requirements, enabling faster time-to-market and better response times to competitive threats and other market dynamics.

Server virtualization has become a key building block for the private cloud, and as organizations take their virtualized environments to the next level—by consolidating a higher number of applications and mission-critical workloads—new networking requirements arise in terms of performance and virtualization awareness.

Brocade® has been delivering innovative technologies for years to help organizations meet the new demands of the highly virtualized data center, while at the same time enabling them to virtualize mission-critical applications and increase server consolidation ratios with greater confidence. Technologies like Brocade Server Application Optimization (SAO) have allowed Brocade Fibre Channel [Host Bus Adapter](#) (HBA) customers to extend essential fabric services, such as Quality of Service (QoS), all the way to the server and VM level. Virtual Machine Optimized Ports (VMOPs) in [Brocade Converged Network Adapters](#) (CNAs) offload essential networking tasks from the most common hypervisors in the industry to enable line-rate performance in 10 Gigabit Ethernet (GbE) environments. Brocade HBAs and [CNAs](#) provide the necessary performance for real-world applications to support the highest virtualization ratios per server with peace of mind. In addition, Brocade Network Advisor provides unified management of all of these resources, from Storage Area Network (SAN) to Local Area Network (LAN) and server connectivity products, under a centralized single pane of glass with VM visibility and integration into industry-leading third-party management and orchestration frameworks.

INTRODUCING THE BROCADE 1860 FABRIC ADAPTER

Taking these concepts to the next level, Brocade has introduced the Brocade 1860 Fabric Adapter, a new class of server connectivity product that includes a set of features and technologies designed to help organizations simplify and optimize server connectivity and I/O in virtualized environments.

As described, the increasing virtualization ratios and the more powerful servers that are capable of driving new levels of I/O are creating unprecedented pressure on network connections. Today, organizations typically deploy an Ethernet-based network for TCP/IP communications within the data center and with the outside world, as well as—on many occasions—a Fibre Channel-based SAN to enable efficient sharing of storage resources, which is essential in any virtualized data center. For storage, virtualization has been creating a demand for higher performance. This has been the primary driver in the transition from 4 Gbps to 8 Gbps Fibre Channel technology over the past few years. On the LAN side, however, 10 GbE technology has not been readily available in an affordable manner until relatively recently. The result has been a great proliferation of 1 GbE Network Interface Cards (NICs) being installed in servers dedicated to virtualization, with the corresponding consequences in terms of power and cooling, cabling, and management complexity.

With the emergence of Data Center Bridging (DCB) and Fibre Channel over Ethernet (FCoE), the lines between the traditional Ethernet-based LAN and Fibre Channel-based SAN are starting to blur, particularly at the access layer, with to the introduction into the market of CNAs—for rack-based and blade servers—and top-of-rack and embedded FCoE switches. In addition, virtualization has essentially moved the access layer of the network into the servers, limiting VM visibility from the point of view of the network. This has made it impossible to assign networking policies

with VM granularity. This has also impacted performance, because every I/O operation has to go through the hypervisor, which increases CPU utilization and hinders virtualization scalability. The Brocade fabric adapter technology extends essential fabric services from Fibre Channel and Ethernet fabrics all the way to the VM and application level (see Figure 1). It optimizes I/O performance in virtualized environments while enabling seamless application mobility for the highest levels of operational flexibility.

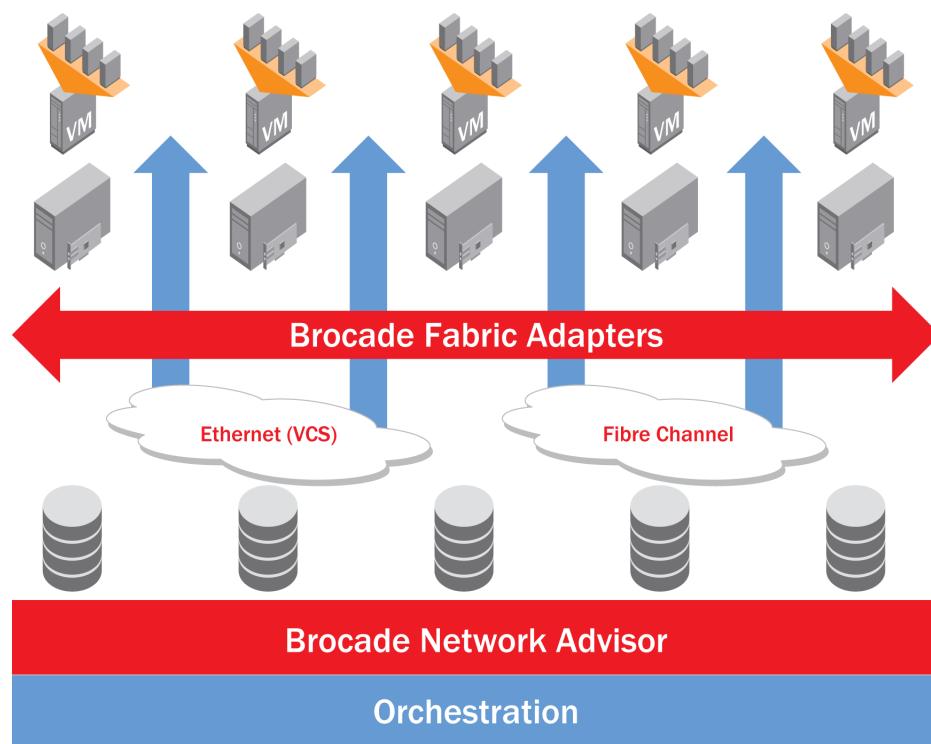


Figure 1: Brocade fabric adapter technology extends fabric services to the applications

Brocade AnyIO Technology

Brocade fabric adapters are multiprotocol, to support both the Ethernet and Fibre Channel networks that organizations have deployed. The ultimate expression of this multiprotocol capability is the new Brocade AnyIO technology, which enables the Brocade 1860 Fabric Adapter to combine a Fibre Channel HBA, a CNA, and a NIC in a single product and to extend essential fabric services to the VM and application level. It supports native 16 Gbps Fibre Channel as well as 10 GbE DCB for TCP/IP, FCoE, or Internet Small Computer System Interface (iSCSI), and it can run all protocols simultaneously in a single card. Users have the flexibility to choose, on a port-by-port basis, the connectivity protocol that is the most appropriate for their applications and their business requirements, without compromises and without any licensing. This unprecedented flexibility allows organizations to standardize on a single adapter for all their connectivity needs.

Each port on the Brocade 1860 can be configured in any of the following modes:

- **HBA mode:** Appears as a Fibre Channel HBA to the operating system (OS). It supports 16/8/4 Gbps Fibre Channel when using a 16 Gbps SFP+ and 8/4/2 Gbps when using an 8 Gbps SFP+.
- **NIC mode:** Appears as a 10 GbE NIC to the OS. It supports 10 GbE with DCB, iSCSI, and TCP/IP simultaneously.
- **CNA mode:** Appears as two independent devices, a Fibre Channel HBA (using FCoE) and a 10 GbE NIC to the OS. It supports 10 GbE with DCB, FCoE, iSCSI, and TCP/IP simultaneously.

Unmatched Flexibility and Performance

In addition, the Brocade 1860 Fabric Adapter provides organizations with unmatched flexibility to support their dynamic server connectivity needs in highly virtualized environments. With escalating performance requirements to support the application consolidation ratios that are required, and with so many options for storage connectivity—including Fibre Channel, FCoE, iSCSI, and NAS—organizations can standardize on a single adapter for all their servers and have peace of mind, knowing that they can connect to the network and storage technology required for each application.

Supporting line-rate 16 Gbps Fibre Channel and 10 GbE, the Brocade 1860 is ideal for bandwidth-intensive applications such as backup, video editing and rendering, and live VM migrations—including vMotion or storage vMotion in VMware ESX environments. In Fibre Channel mode, the Brocade 1860 supports N_Port trunking as part of SAO. N_Port trunking can aggregate two 16 Gbps Fibre Channel links into a single logical 32 Gbps link with frame-level load-balancing for the highest levels of link utilization and transparent, automatic failover and failback for high availability.

Additionally, the Brocade 1860 can deliver over one million I/O operations per second (IOPS) for block storage applications independent of the protocol of choice—Fibre Channel, FCoE, or iSCSI. This makes the Brocade 1860 an ideal solution for I/O intensive applications such as e-mail, databases, Online Transaction Processing (OLTP), Virtual Desktop Infrastructure (VDI), and any application deployed on Solid State Drive (SSD) storage technology. With such unparalleled performance capabilities, organizations can deploy mission-critical transactional applications in virtualized environments with confidence that the fabric adapter will be able to handle the load.

BROCADE vFLINK I/O VIRTUALIZATION (IOV)

As mentioned previously, as organizations evolve their architectures towards a private cloud, they are increasing the VM densities and virtualizing more mission-critical applications. This is driving increased performance requirements from the network and from the connection between the server and the network. Users currently deploy a large amount of 1 GbE NICs to be able to support these performance demands, but also because they need to provision network connectivity for multiple purposes, ranging from management to backup, production, or live VM migration. They need to keep these different network connections independent and isolated, generally connected to separate VLANs within the network, to guarantee service levels and QoS. Even with the emergence of 10 GbE, organizations have been hesitant to consolidate this large number of 1 GbE interfaces for fear of losing network granularity, isolation, and not being able to guarantee QoS for each of those networks.

Additionally, many servers will also have Fibre Channel HBAs—generally two for high-availability purposes, but sometimes more—to connect to the storage network. The result is an unmanageable adapter and cable sprawl coming out of the back of the server, leading to increased management complexity.

Brocade vFLink technology allows a single Brocade 1860 Fabric Adapter to logically partition a physical link into as many as four virtual fabric links (vFlinks). On a dual port adapter, a total of 8 virtual NICs (vNICs) or virtual HBAs (vHBAs¹) can be created. This is achieved by replicating the adapter at the PCIe bus level and presenting multiple PCIe physical functions (PFs) to the OS layer, similar to the way a multi-port NIC presents itself as multiple independent devices or a dual-port CNA presents itself as four devices (two NICs and two HBAs). The operating system does not need to have any special support for vFLink; it will just see each vNIC or vHBA as a separate physical I/O device, and it will know how to operate it as long as the appropriate driver is present. When configured as 16 Gbps Fibre Channel, a single physical port can support up to four vHBAs. When configured as 10 GbE, a physical port can support any combination of up to four vNICs and vHBAs. In this case, vHBAs are achieved by using FCoE protocol.

¹ Multiple vHBA support available in a future driver release

Portions of the total available bandwidth—10 Gbps for Ethernet ports and 16 Gbps for Fibre Channel ports—can be allocated to these virtual fabric links in 100 Mbps increments. Each vNIC can be assigned to a different VLAN in order to maintain the isolation and QoS for the different networks. This allows organizations to consolidate up to ten 1 GbE NICs into a single 10 GbE interface while maintaining the management granularity they require (see Figure 2).

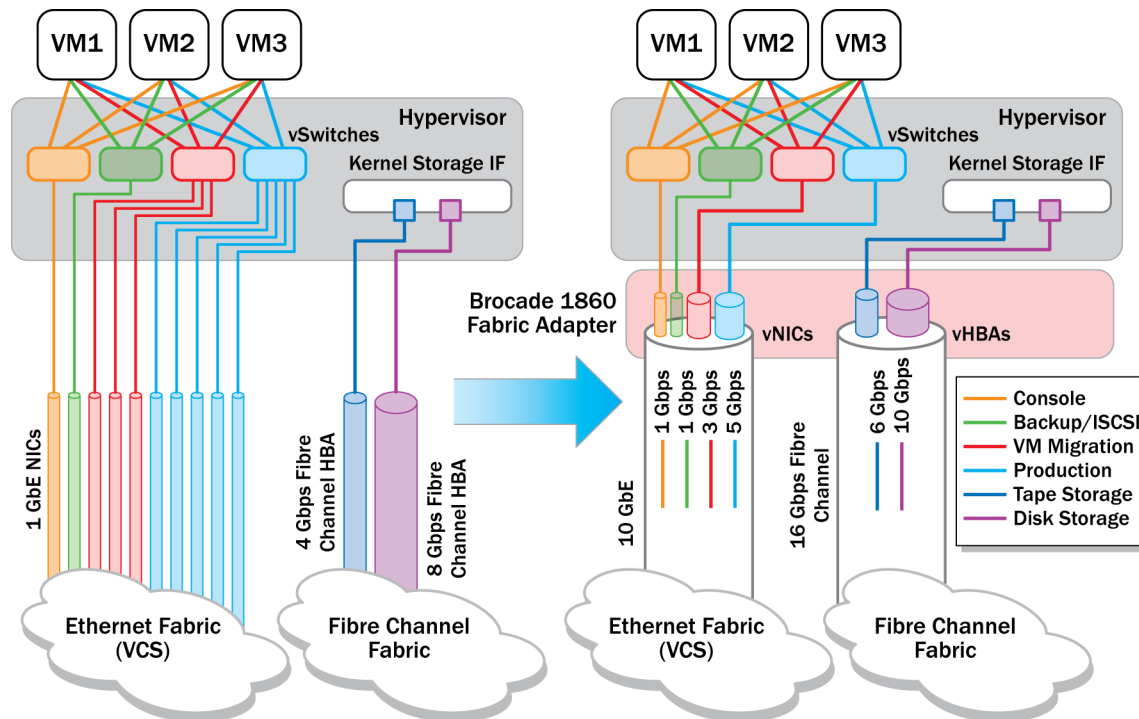


Figure 2: Adapter consolidation with Brocade vLink technology

IMPROVING I/O PERFORMANCE FOR VIRTUAL MACHINES

In a virtualized environment, the hypervisor typically controls the physical adapters and presents emulated I/O devices—standard SCSI block devices or network interfaces—to the virtual machines. From an IP perspective, the hypervisor establishes a Virtual Ethernet Bridge (VEB) or vSwitch to allow the VMs to communicate with each other or with external devices. From a storage perspective, there is no need for inter-VM communications and thus no requirement for a virtual switch, but every I/O from a VM still has to go through the hypervisor, which has to translate virtual guest addresses into physical host addresses. At very high transfer speeds—such as 16 Gbps Fibre Channel or 10 GbE—a bottleneck can quickly occur, impacting application performance and limiting VM scalability.

Direct I/O

In order to improve I/O performance, hypervisors typically have the option to grant VMs direct access to hardware resources. This is known as direct I/O, and in VMware ESX environments it is implemented by means of a feature called VMDirectPath I/O. Direct I/O relies on server chipset technologies like Intel Virtualization Technology for Directed I/O (VT-d) or AMD I/O Virtualization (AMD-Vi), which implement I/O Memory Management Units (IOMMUs), and PCI technologies such as Address Translation Services (ATS) to take care of translating memory addresses from virtual to physical.

This removes the hypervisor involvement in I/O processing and enables near-native performance for those VMs. However, directly assigned I/O devices cannot be accessed by more than one VM at a time, thus requiring dedicated hardware resources. I/O virtualization technologies like vFLink can help alleviate this problem by logically partitioning the adapter and directly mapping vNICs or vHBAs to VMs, enabling a better sharing of physical adapters with direct I/O (Figure 3).

However, current implementations of direct I/O—whether to a physical or virtual (vNIC/vHBA) adapter—do not support the use of certain advanced virtualization technologies, such as live VM migrations. For this reason, direct I/O is typically used in a limited fashion for very specific VMs that absolutely need the performance advantage and do not require these advanced features. In addition, the number of virtual I/O devices that can be achieved with PF-based I/O virtualization does not provide enough scalability to support the ever-increasing virtualization densities that organizations require.

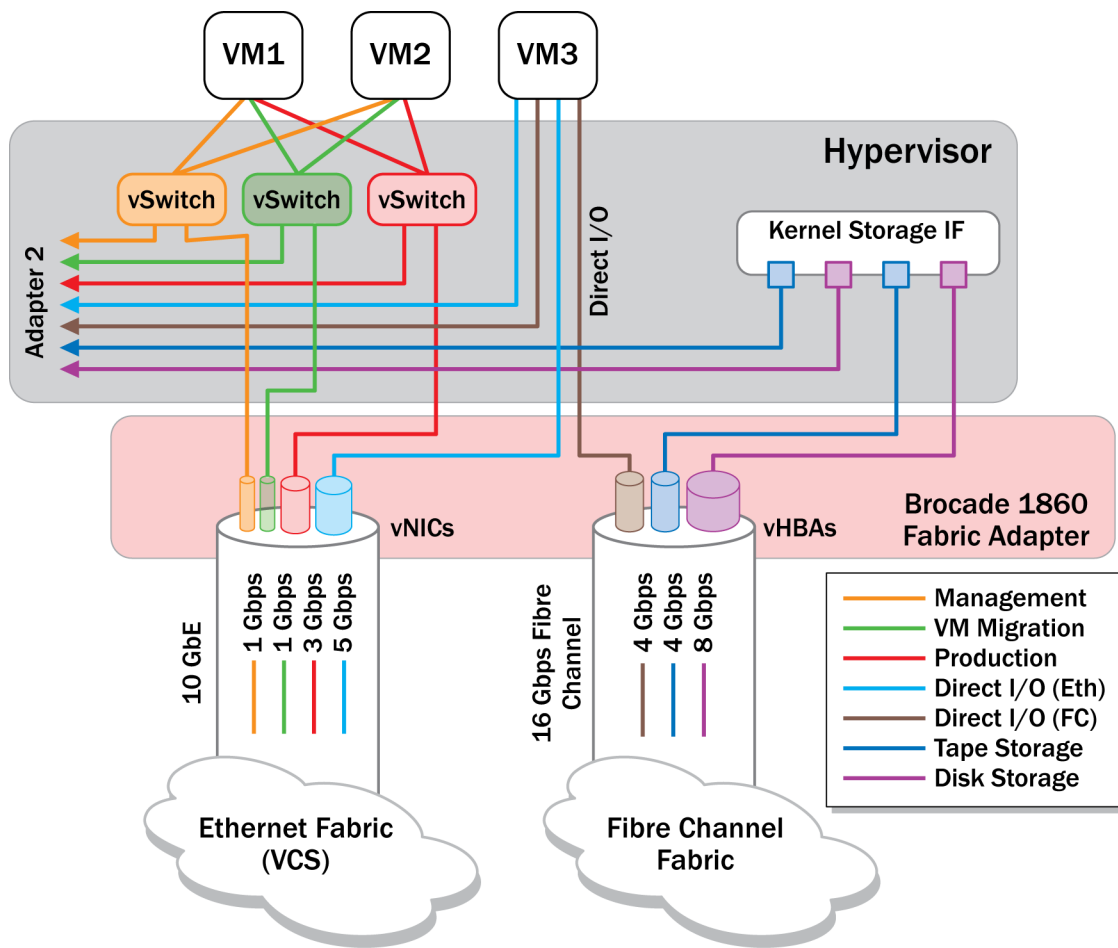


Figure 3: vFLink and direct I/O

Single-Root I/O Virtualization (SR-IOV)

In order to address some of these challenges, the PCI Special Interest Group (PCI-SIG) has developed the SR-IOV specification to define a scalable and standard mechanism for virtualizing I/O devices. SR-IOV allows a PCIe device to be virtualized, by introducing the concept of PCI virtual functions (VFs). VFs are lightweight PCI functions that can be used only to move data in and out of the device, and that have a minimum set of configuration resources. The SR-IOV specification allows scalability to virtually thousands of VFs; however, real-world implementations will likely be limited to a few hundred.

VFs can be directly mapped to virtual machines using direct I/O technology, while the hypervisor retains control of the PF, which requires what is called a “split-driver” model. A standard device driver is loaded on the hypervisor to control the PF, where all the configuration of the physical adapter, including the creation and management of VFs, occurs. On the VMs, a lightweight version of the device driver is loaded to handle the I/O operations through these limited-function VFs (see Figure 4). The VFs cannot be treated like full PCIe devices, and for this reason SR-IOV must be supported not just by the adapter, but also by the server BIOS and the hypervisor.

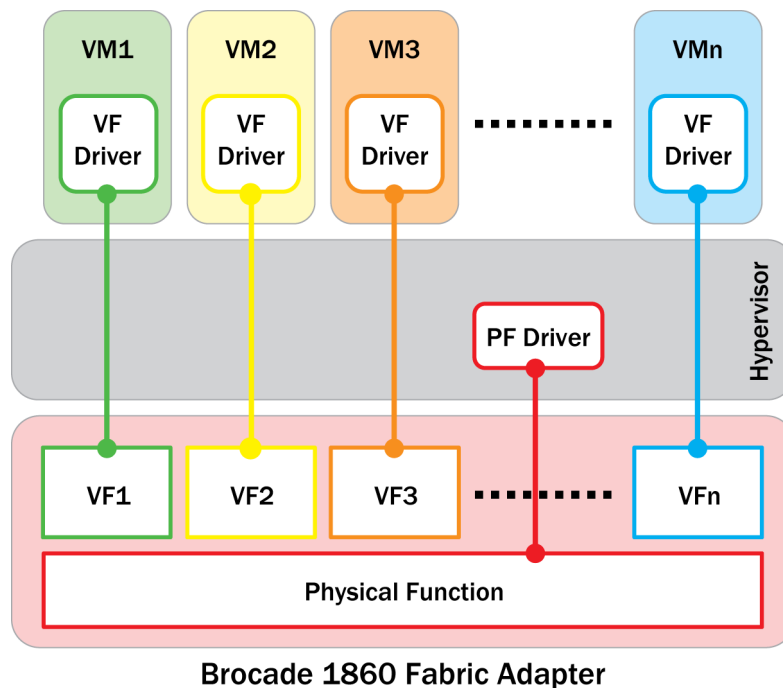


Figure 4: Single-Root I/O Virtualization architecture

Most modern servers support SR-IOV or will be able to support SR-IOV in the future with a BIOS update. However, hypervisor support for SR-IOV is not common today, and it will become more widely available in 2012. At the same time, hypervisor vendors are expected to deliver enhancements to current direct I/O technologies in order to leverage SR-IOV to its fullest potential without sacrificing advanced features like live VM migrations. While direct I/O does not really require SR-IOV, practical implementations in real-world environments will require more scalability than what PF-based I/O virtualization provides, making SR-IOV an important element of any solution.

The Brocade 1860 Fabric Adapter hardware supports the PCI SIG SR-IOV specification with up to 255 VFs, extending Brocade vFLink technology to provide a much more efficient sharing of the adapter. Software support for this feature will be added with a future driver release once SR-IOV is supported on the most popular hypervisors in the industry, such as VMware ESX and Microsoft Hyper-V.

VIRTUAL SWITCHING

As mentioned in the previous section, a software-based VEB implemented in the hypervisor provides inbound/outbound and inter-VM network communication. This places an additional burden on the server CPU, as the hypervisor needs to inspect all Ethernet packets and filter them, based on MAC address and VLAN tagging, in order to determine the destination VM for each packet. As a result, every incoming packet requires two CPU interrupts, one for the hypervisor to inspect the packet and one for the CPU core that has the affinity with the destination VM. In addition, even in today's multi-core CPUs, only one core is used for these incoming packet classification and sorting tasks, which can be overloaded and become a bottleneck for overall system performance.

Virtual Machine Optimized Ports (VMOPs)

A first approach at addressing this challenge was the implementation by the hypervisors of multi-queue technologies, such as VMware NetQueue and Microsoft VMQ. These enable the creation of dedicated I/O queues that can be leveraged to deliver incoming packets directly to the affinity CPU core for the destination VM, thus reducing by half the number of CPU interrupts for every I/O operation. In order to do this, the NIC needs to take over the packet inspection and the Layer 2 classification and sorting based on MAC address and VLAN tagging from the hypervisor. The Brocade 1860 Fabric Adapter—as well as the Brocade 1010 and 1020 CNAs—support these technologies through the implementation of VMOPs, with up to 64 queue pairs per port to help free the CPU from this burden and enable near-line-rate performance in 10 GbE environments (see Figure 5).

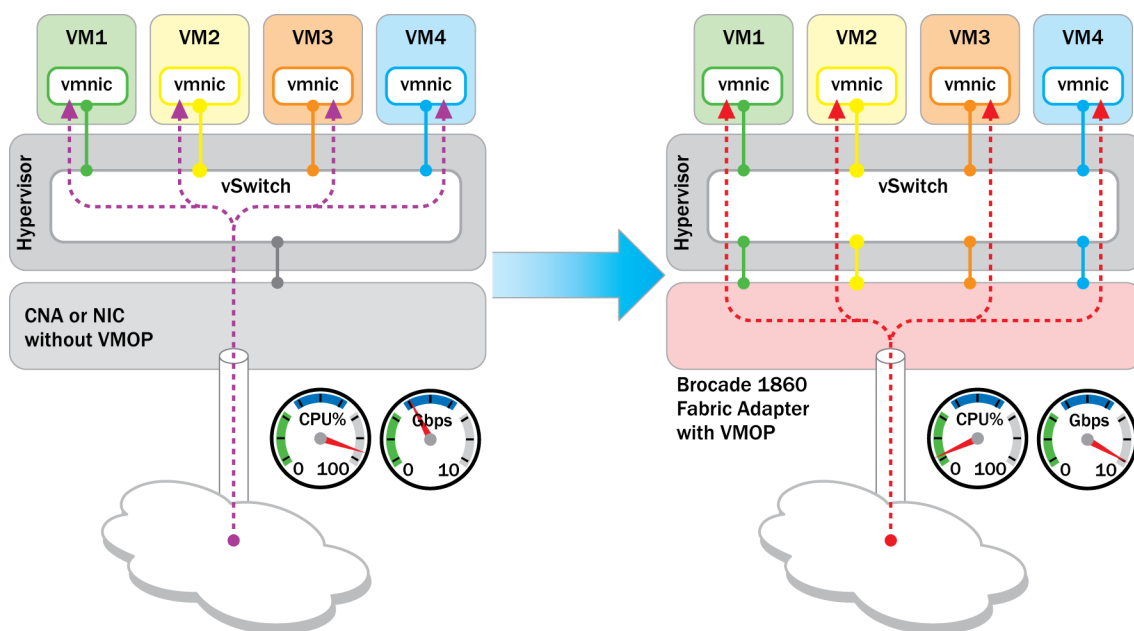


Figure 5: Virtual Machine Optimized Ports (VMOPs)

Hardware-Based Virtual Ethernet Bridge (VEB)

However, as the number of VMs per server and the performance demands for each VM continue to grow, the hypervisor still needs to be involved to handle VM traffic switching, particularly for traffic between VMs inside the same physical server. This continues to be a burden for server CPUs. As SR-IOV becomes widely available, and direct I/O technologies mature to overcome their current limitations, VMs will directly access the virtualized hardware resources and completely bypass the hypervisor and vSwitch altogether. This will enable native I/O performance for VMs, and all inter-VM traffic switching will be offloaded to a fully integrated virtual switch or Virtual Ethernet Bridge (VEB) residing inside the I/O adapter. The adapter will then be responsible for providing both inter-VM and inbound/outbound communication. The advantages of this approach are the following:

- Packets are switched directly in the adapter with no hypervisor involvement, providing high performance, low latency and low CPU utilization.
- No special support is required from the access layer switch, since inter-VM traffic continues to be switched inside the server.

Virtual Ethernet Port Aggregator (VEPA)

While an adapter-based VEB provides the best performance and accomplishes the task of offloading network switching functions from the hypervisor, improving I/O performance and alleviating CPU utilization, it still provides limited visibility into the VMs and their traffic for an organization's network management team. In addition, an integrated VEB inside an adapter will typically provide fairly basic L2 networking services, and not the comprehensive set that an enterprise-class top-of-rack 10 GbE switch would provide.

In order to address this, a second approach to handling inter-VM switching is VEPA, where all VM-generated traffic is sent out of the adapter to an external switch, essentially moving the demarcation point of the network back to the physical access layer. The external switch can then apply filtering and forwarding rules to the VM traffic, and it can also account for and monitor the traffic with the same management tools that network administrators are accustomed to. Traffic destined for VMs within the same physical server can be forwarded back on the same physical port via a "reflective relay" process known as "hairpin turn." Since this process is not typically allowed by the Spanning Tree Protocol (STP) to avoid network loops, access layer switches need to support this new feature in order to support VEPA. In essence, VEPA extends network connectivity all the way to the applications, making VMs appear as if they were directly connected to the physical access layer switch. VEPA can also work with software-based VEBs inside a hypervisor, requiring minimal changes to the software in order for the vSwitch to forward every packet to the outside switch. In that case, SR-IOV is also not a requirement for VEPA. In such a scenario, however, the hypervisor will still be involved in every I/O operation and will continue to be a bottleneck for network performance.

Multi-channel VEPA is an additional enhancement to VEPA to allow a single Ethernet connection to be divided into multiple independent channels, where each channel acts as a unique connection to the network. Multi-channel VEPA uses a tagging mechanism called Q-in-Q (defined in IEEE 802.1ad, sometimes referred to as 802.1QinQ), which adds an "outer" VLAN tag (also called a service tag or S-Tag) in addition to the standard VLAN tag. This method requires Q-in-Q capability from both the adapter and the access layer switch, which is usually supported in hardware. As such, switches that do not currently support Q-in-Q cannot be easily upgraded to support multichannel VEPA.

The benefit of multi-channel VEPA is that it allows a combination of VEPA for VMs where strict network policy enforcement and traffic monitoring is important, and hardware-based VEB for high-performance VMs where minimal latency and maximum throughput is a requirement.

Edge Virtual Bridging (802.1Qbg)

The term Edge Virtual Bridging (EVB) refers to a standard being developed by the IEEE 802.1 working group as IEEE 802.1Qbg. It defines, among other things, how external switches and VEBs—software- and hardware-based—can talk to each other to exchange configuration information. It defines the standard for VEPA and the Virtual Station Interface Discovery Protocol (VDP)—sometimes referred to as Automatic Migration of Port Profiles (AMPP)—that can be used to automatically associate and de-associate a VM to a set of network policies, sometimes referred to as "port profile."

VDP can automate the migration of a network state or profile ahead of a VM's migration across servers in order to guarantee a successful live migration without disruptions due to inconsistent network settings on the destination switch or port.

The Brocade 1860 Fabric Adapter supports both approaches for offloading inter-VM traffic switching from the hypervisor (see Figure 6), including standard and multi-channel VEPA. Since any practical implementation of both approaches requires SR-IOV and direct I/O in order to bypass the hypervisor, actual software support for these technologies will be available coinciding with SR-IOV support.

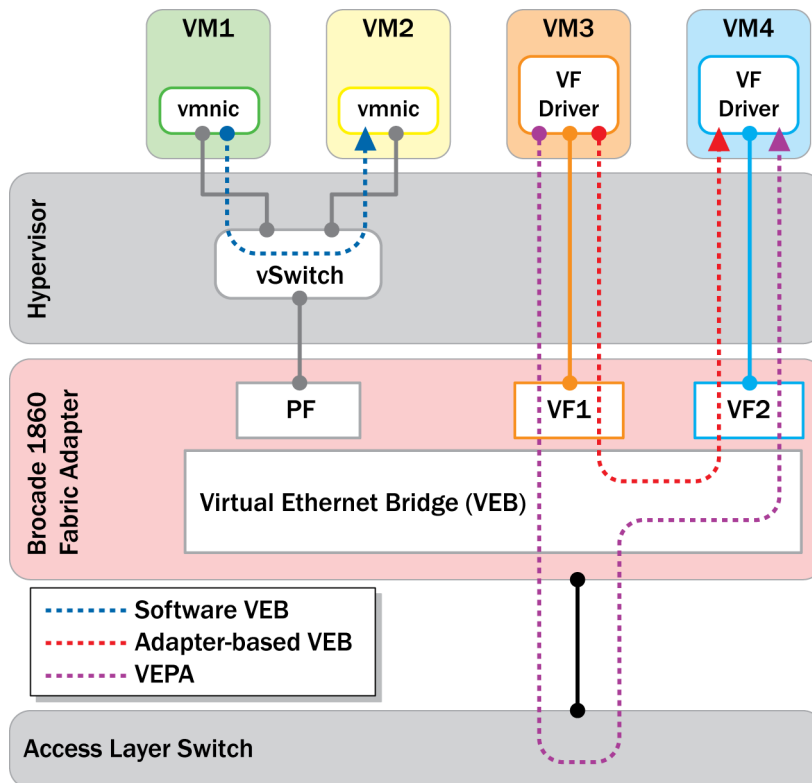


Figure 6: Switching options for virtualized environments

In traditional virtualized environments, management of physical and virtual networking is fragmented, as the software vSwitch is typically managed by the server administrator, whereas the physical network is managed by the network administrator. The network administrator has no visibility into the software switch management, and is unable to enforce networking policies on inter-VM traffic within a physical host. By offloading the switching from the hypervisor onto the adapter or the access layer switch, management of physical and virtual switching can be unified under a single management application by the network administrator. With tools like Brocade Network Advisor that unify the management of adapters and switches, unified management of physical or virtual networking can be achieved in both hardware-based VEB and VEPA scenarios. This approach returns control of all network management duties to the network administrator and greatly simplifies operations.

SUMMARY

Brocade continues to lead in technology innovation to help organizations solve real challenges in their day-to-day operations. With the introduction of the Brocade 1860 Fabric Adapter, organizations can dramatically simplify server connectivity and optimize I/O performance in their virtual server environments. This will help them scale their server virtualization to the levels required in order to evolve their architectures towards a private cloud. For more information on the Brocade 1860 Fabric Adapter, visit www.brocade.com/adapters.

© 2011 Brocade Communications Systems, Inc. All Rights Reserved. 07/11 GA-TB-375-01

Brocade, the B-wing symbol, DCX, Fabric OS, and SAN Health are registered trademarks, and Brocade Assurance, Brocade NET Health, Brocade One, CloudPlex, MLX, VCS, VDX, and When the Mission Is Critical, the Network Is Brocade are trademarks of Brocade Communications Systems, Inc., in the United States and/or in other countries. Other brands, products, or service names mentioned are or may be trademarks or service marks of their respective owners.

Notice: This document is for informational purposes only and does not set forth any warranty, expressed or implied, concerning any equipment, equipment feature, or service offered or to be offered by Brocade. Brocade reserves the right to make changes to this document at any time, without notice, and assumes no responsibility for its use. This informational document describes features that may not be currently available. Contact a Brocade sales office for information on feature and product availability. Export of technical data contained in this document may require an export license from the United States government.