White Paper

Virtual Machine Networking: Standards and Solutions

What You Will Learn

With the advent of server virtualization, two basic assumptions of data center network design have changed: multiple OS images (or virtual machines) are now allowed to transparently share the same physical server and I/O devices, and the relationship between an OS image and the network is now dynamic. The access layer of the network extends further to support local switching between different virtual machines within the same server, thus invalidating the traditional assumption that each network access port corresponds to a single physical server running a single image. Further complicating the picture, each virtual machine can be moved from one physical server to another within the data center or even across data centers.

One option for network virtualization is to implement a software switch as part of the hypervisor. Another option is to enable the switching function to be performed by an external switch. Both approaches offer advantages and should be looked at as complementary options for different workload profiles and data center designs.

This document describes the two main standards for virtual networking and briefly discusses existing Cisco® solutions.

Definitions: Virtual Embedded Bridge, Reflective Relay, Multichannel, and Port Extension Capabilities

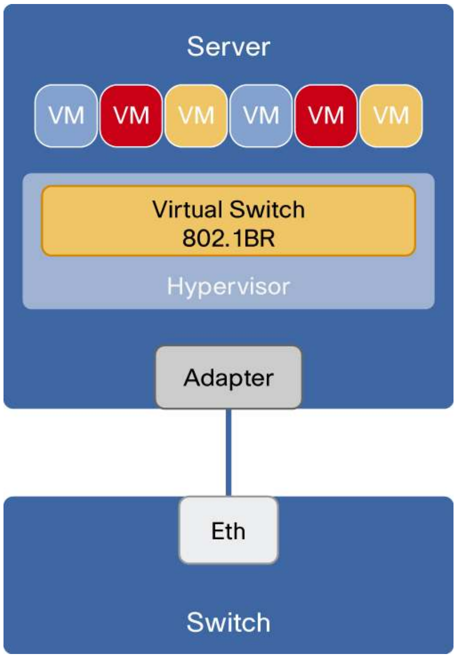
Cisco and other major industry vendors have made standardization proposals in the IEEE to address networking challenges in virtualized environments. The resulting standards tracks are IEEE 802.1Qbg Edge Virtual Bridging and IEEE 802.1BR Bridge Port Extension. Note that these standardization efforts focused on the specification of the data plane functions of a virtual network switch (or bridge) without addressing the network-to-hypervisor management functions that are necessary to deliver a deployable and supportable solution.[1]

Virtual Embedded Bridge

A virtual embedded bridge (VEB) is an embedded bridge used in virtualization applications (Figure 1). Note that there is no separate standard for a VEB. A VEB may be a fully standards-compliant IEEE 802.1Q Ethernet bridge that resides within the hypervisor. Alternatively, a VEB may be hypervisor aware and may take a streamlined approach; for example, it may get MAC addresses from the hypervisor instead of learning them. VEBs provided by hypervisor vendors are often managed through hypervisor management tools instead of by traditional bridge management functions.

[1] Bridge Port Extension standardization effort initially started under the IEEE 802.1Qbh working group. IEEE then decided to move it under the 802.1BR standard, to make it independent from IEEE 802.1Q.

Figure 1. Virtual Embedded Bridge



The Cisco Nexus® 1000V Switch, as will be further discussed in a later section, is an example of an extremely advanced VEB. As such, it performs as a standard IEEE 802.1Q bridge that works with all existing bridges and servers and does not require any change to the existing frame format.

The VEB approach is extremely valuable for current and future data center environments because it enables flexible and ubiquitous deployments, dense virtual machine environments, and rapid deployment of new features.

External Hardware Switch

A complementary approach both in the standard tracks and in the solutions proposed by the market leaders enables a mode on the VEB, hypervisor, or adapter in which all traffic sourced by a virtual machine is forwarded to an external controlling switch (a traditional access-layer switch).

Within this category, a further distinction can be made between tagless (reflective relay) and tagged (multichannel and port extension) options. These options are currently under development in the IEEE within the IEEE 802.1Qbg and 802.1BR projects.

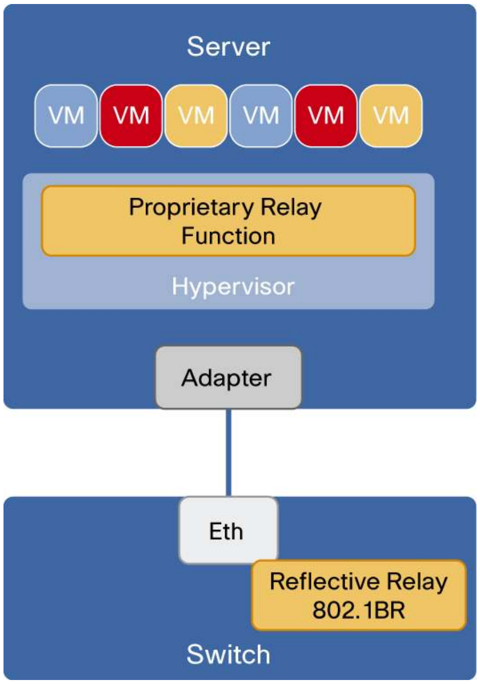
The Cisco UCS M81KR Virtual Interface Card (VIC) is an example of a complete port extension solution that uses the external controlling switch to switch virtual machine traffic. Use of an external switch has the benefit of consolidating the virtual and physical switching infrastructure into a single entity and simplifying the management infrastructure.

IEEE 802.1Qbg Edge Virtual Bridging

Reflective Relay

Reflective relay (Figure 2) is the only tagless option that uses an external controlling switch for virtual machine traffic switching.

Figure 2. IEEE 802.1Qbg Reflective Relay



In this approach, the VEB forwards all frames sourced by the virtual machines to the adjacent controlling switch. The controlling switch applies various policies on those frames and then forwards them back to the VEB. The VEB then forwards the frame to the appropriate virtual machine (or machines) based on the MAC address and VLAN ID.

IEEE 802.1Qbg specifies a function in the controlling switch that allows a packet received on a switch port to be pinned on the same port, a behavior called reflective relay. The standard does not specify how the hypervisor, network interface card (NIC), or VEB uses this relay function, hence making it vendor dependent and proprietary.

Reflective relay and the proprietary relay function are simple to implement since they do not require a tag. They partially free computing resources from advanced policy and monitoring functions by offloading these to the controlling switch, though at the cost of extra link bandwidth utilization. Most server access switches can provide this capability through software upgrade: that is, without hardware modification. This option is useful for policies that need to be applied globally across virtual machines; however, this approach has strong limitations in its ability to apply policies to a subset of virtual machines.

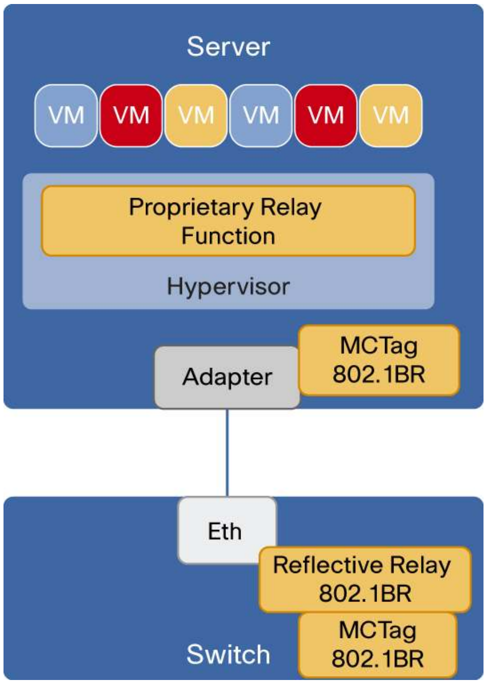
This mode requires nonstandard modifications to the hypervisor to implement the relay function. Also, although computing resources are partially freed from advanced policy and monitoring functions, this benefit is offset by the consumption of CPU resources because each packet must pass through the VEB twice. Further, from a management perspective, there are now two separate elements, a VEB and a controlling switch, that require coordination to determine where specific features are applied and implemented for the same virtual machine traffic.

Currently, no products are shipping with the proprietary relay function.

Multichannel Capability

Reflective relay operation depends on the capability of devices using it to perform source suppression of multicast frames. Such source suppression requires knowledge of all the MAC addresses in use by the device. However, this knowledge is not possible with a large class of data center applications, including firewalls and load balancers. The multichannel approach (Figure 3) overcomes some of these limitations by introducing a tag to explicitly specify the source and destination of virtual machine traffic to handle critical applications such as these.

Figure 3. IEEE 802.1Qbg Reflective Relay and Multichannel



The multichannel approach specifies a new function of the VEB and the controlling switch. The VEB continues to be managed as a standalone device. Deploying multichannel capability is somewhat equivalent to adding a switch, further complicating management.

Multichannel capability does not efficiently handle multicast frames. A separate copy of each multicast frame is required for each channel in a multichannel deployment, thus consuming additional bandwidth.

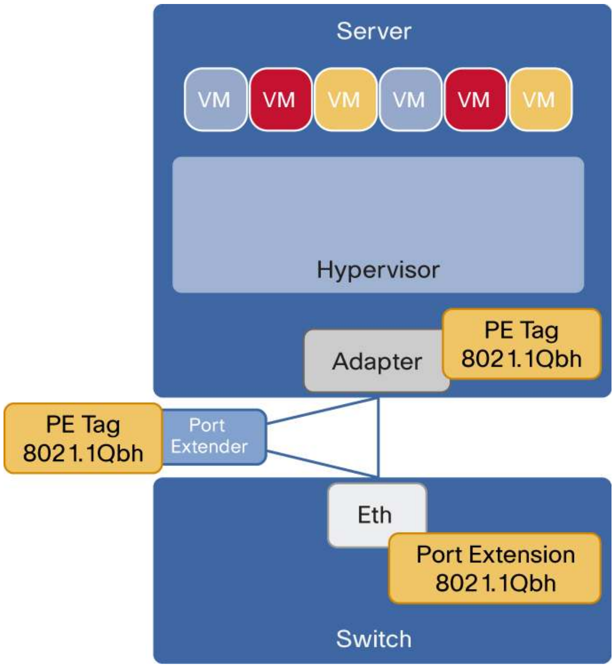
Mulitchannel capability is specified as optional by IEEE 802.1Qbg. Currently, no products are shipping with multichannel support.

Bridge Port Extension

IEEE 802.1BR introduces a new device called a port extender, One or more port extenders can be attached to a controlling switch. A controlling switch and its set of port extenders form a single extended switch. An extended switch is managed as a single entity through the controlling switch. Adding a port extender to a controlling switch is like adding a line card to a traditional switch.

Similar to the multichannel approach, port extension (Figure 4) specifies the use of a tag that is added to the frame to enable the controlling switch to identify the source extended port of a frame and to explicitly direct a frame to a destination extended port.

Figure 4. IEEE 802.1BR Port Extension



As mentioned earlier, multichannel capability requires the controlling switch to transmit the same broadcast or multicast frame multiple times: once for each channel in the multichannel deployment, forcing the switch to create and queue multiple copies of the same packet in the same egress queue and so introducing the need for additional hardware modification and entailing significant on-the-wire inefficiency. With port extenders, the controlling switch sends the frame only once with a replication control tag, and the port extender replicates the frame to the appropriate ports, making port extenders significantly more efficient.

Additionally, port extenders can be cascaded. Cascading has several benefits. It allows multiple network layers to be managed as a single layer. Also, without cascading, the controlling switch to which a server is connected is most likely the simplest device in the network hierarchy (either a blade switch or an access switch), and it may not support advanced functions that may be needed for virtual machine traffic switching (such as advanced quality of service [QoS], policy enforcement, and troubleshooting). Port extenders can be installed up the hierarchy until a switch with the desired capabilities is located.

Cisco has been shipping prestandard port extension products implementing VNTag since 2009 (Cisco Nexus 5000 Series Switches and Cisco Nexus 2000 Series Fabric Switches distributed access architecture and the Cisco UCS M81KR VIC in the Cisco Unified Computing System™). These products deliver the same capabilities provided by the port extension standard currently under development. After the IEEE 802.1BR standard is complete, Cisco is expected to be able to deliver a product that fully supports a heterogeneous environment of VNTag and standards-based solutions.

Cisco Offers Industry-Leading Virtual Machine Support

Cisco has led the way in virtualized environments specific products, making the network infrastructure virtual machine aware. It provides tools with the same level of visibility, security, and troubleshooting for virtual machines as customers are accustomed to using for physical devices.

Cisco Virtual Machine Networking helps enable:

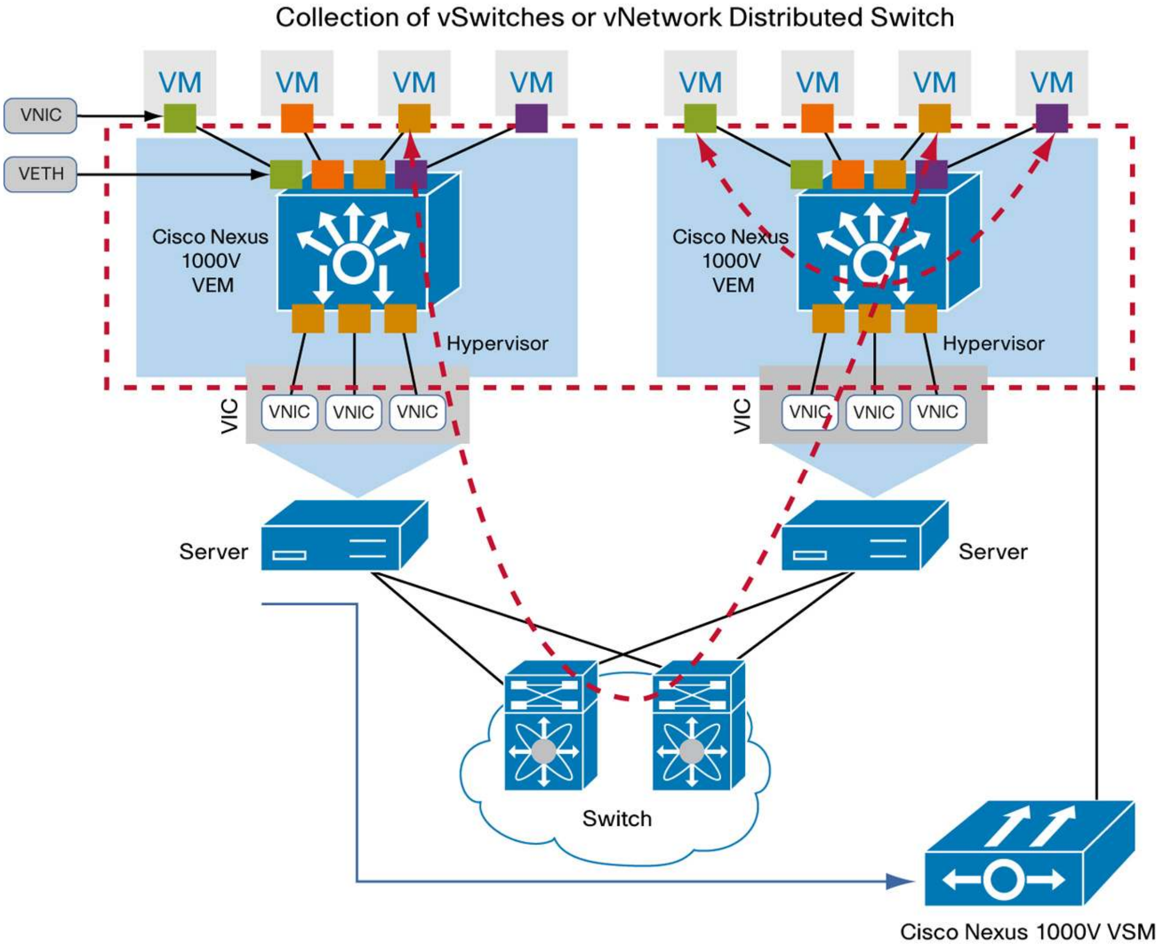
- Policy-based virtual machine networking
- Transparent network and security policy mobility with virtual machine migration
- Nondisruptive operational model, with the network administrator managing both virtual and physical networking resources with a consistent set of tools

The portfolio of Cisco Virtual Machine Networking products provides a variety of options that meet a range of customer needs, including advanced hypervisor switching as well as high-performance hardware switching. It is flexible and extensible, interoperable with developing standards, and service enabled.

Cisco Nexus 1000V

The first product to deliver Virtual Machine networking is the Cisco Nexus 1000V. The Cisco Nexus 1000V is the industry’s most advanced software switch for VMware vSphere, providing switching for up to 64 VMware ESX hosts from a single point of management. Built on Cisco NX-OS Software, it offers advanced features such as QoS, access control lists (ACLs), Encapsulated Remote Switched Port Analyzer (ERSPAN), NetFlow, control-plane security features, and integration of network services (Figure 5).

Figure 5. Cisco Nexus 1000V Series Switches Architecture



The Cisco Nexus 1000V consists of two main components that virtually emulate a modular Ethernet switch with redundant supervisor functions:

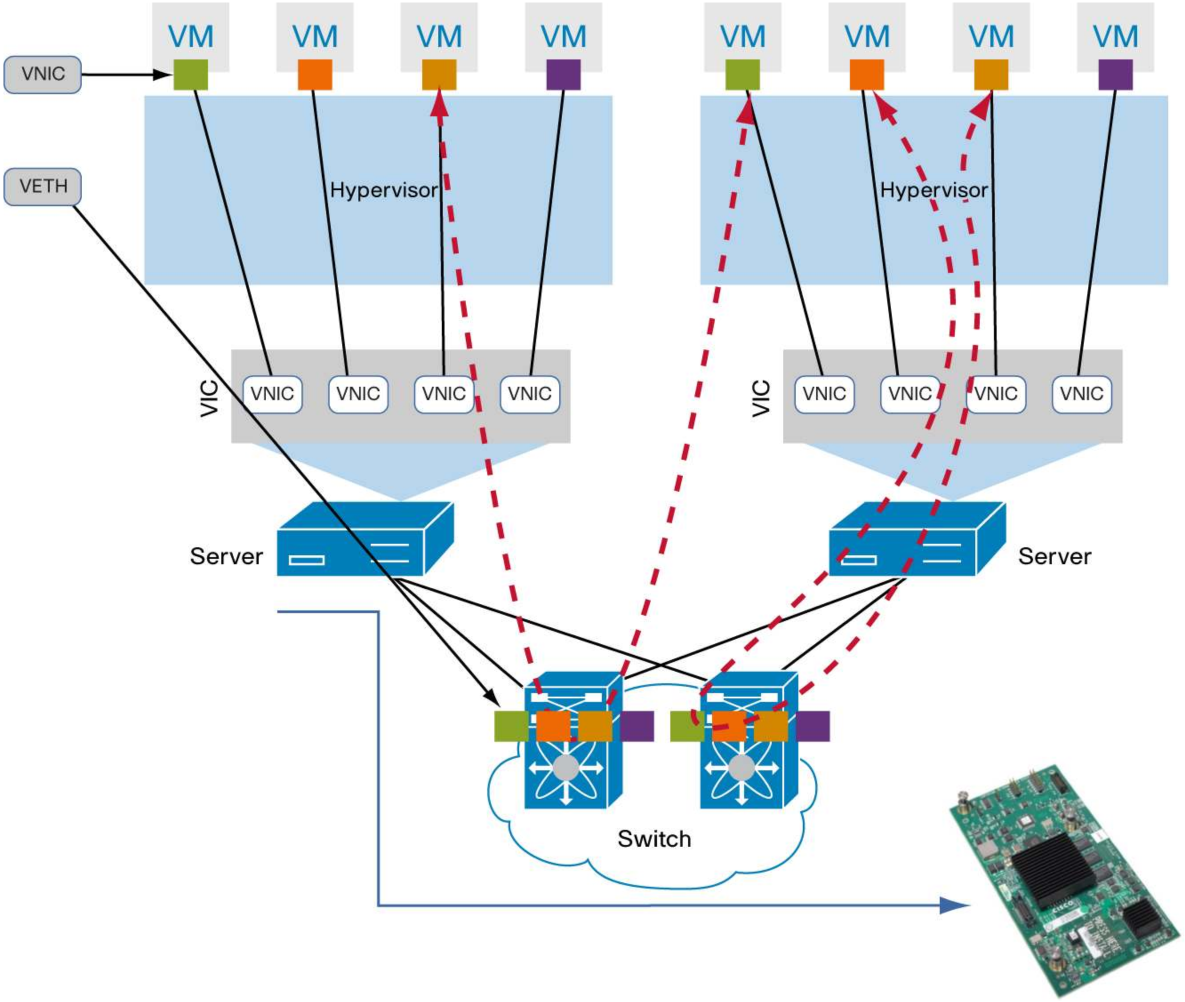
- **Virtual Ethernet module (VEM)-data plane:** This lightweight software component runs inside the hypervisor. It enables advanced networking and security features, performs switching between directly attached virtual machines, provides uplink capabilities to the rest of the network, and effectively replaces the virtual switch (vSwitch). A VEM is embedded with each hypervisor.
- **Virtual supervisor module (VSM)-control plane:** This standalone, external, physical or virtual appliance is responsible for the configuration, management, monitoring, and diagnostics of the overall system (that is, the combination of the VSM itself and all the VEMs that it controls) as well as integration with VMware vCenter. A single VSM can manage up to 64 VEMs. VSMs can be deployed in an active-standby model, helping ensure high availability.

With the Cisco Nexus 1000V, traffic between virtual machines is switched locally at each instance of a VEM. Each VEM is also responsible for interconnecting the local virtual machines with the rest of the network through the upstream access-layer network. The VSM is responsible for running the control plane protocols and configuring the state of each VEM accordingly, but it never takes part in the actual forwarding of packets.

Cisco UCS M81KR and P81E Virtual Interface Cards

The Cisco UCS M81KR VIC is a virtualization-optimized, dual-ported 10 Gigabit Ethernet converged network adapter (CNA) in a mezzanine form factor that enables near-bare-metal I/O performance, ease of management, and network visibility benefits for joint VMware and Cisco customers (Figure 6). The same VIC card is available in a stand-alone PCIe form factor, the P81E VIC.

Figure 6. Cisco UCS M81KR VIC



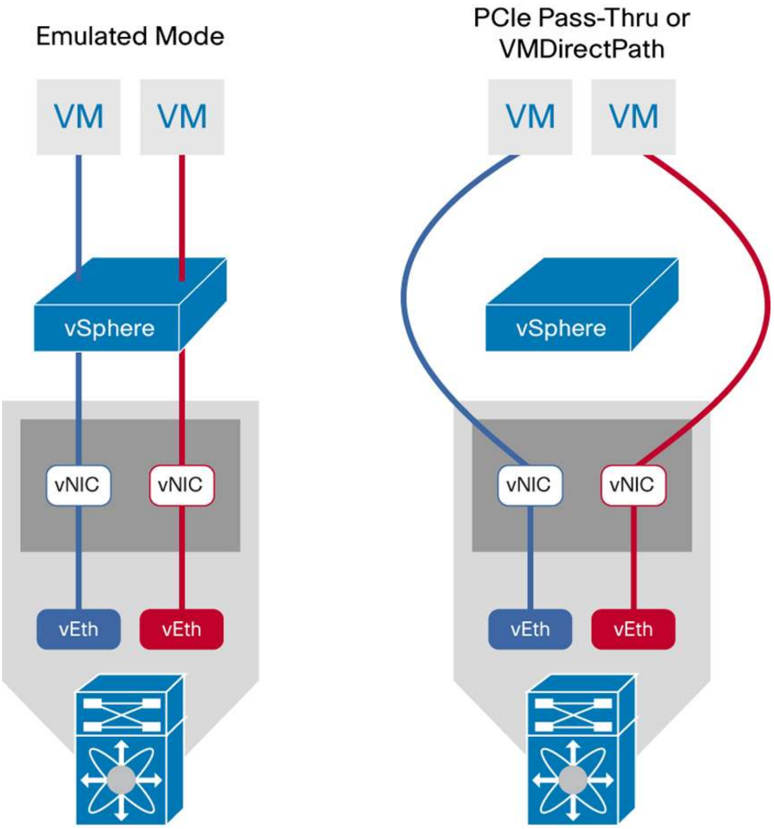
Two main innovations available with the VIC are:

- **Port-extension-like functions with Cisco VM-FEX (Virtual Machine Fabric Extender):** The VIC is the first implementation of VM-FEX technology from Cisco. VM-FEX eliminates the vSwitch within the hypervisor by providing individual virtual machine virtual ports on the physical network switch. Virtual machine I/O is sent directly to the upstream physical network switch, which takes full responsibility for virtual machine switching and policy enforcement. This approach leads to consistent treatment for all network traffic, virtual or physical. VM-FEX consolidates virtual and physical switching layers into a single layer and reduces the number of network management points by an order of magnitude.

- **Hypervisor pass-through:** The VIC uses VMware VMDirectPath technology to significantly improve throughput and latency of virtual machine I/O. VMware VMDirectPath allows direct assignment of PCIe devices to virtual machines. Virtual machine I/O bypasses the hypervisor layer and is placed directly on the PCIe device associated with the virtual machine. The Cisco VIC is the industry's first implementation of this technology. VMware VMDirectPath supports VMware vMotion, thereby enabling dynamic workload management.

The Cisco VIC enables VM-FEX functions in both VMware and Red Hat virtualization environments.

Figure 7. Modes of VM-FEX



Conclusion

- The Cisco Nexus 1000V is a standard IEEE 802.1Q bridge and does not require any proprietary tag.
- Both port extension and multichannel capabilities require new tags.
- Both port extension and multichannel capabilities require modification of the NIC and the controlling bridge hardware.
- Port extension can also be used outside virtualized environments with strong benefits, including simplified management and reduction in the number of devices. An example of such a non-virtualized case is a distributed access layer design using the Cisco Nexus 5000 and 2000 Series.
- The Cisco Nexus 1000V has shipped more than 1 million ports since its introduction.
- The Cisco UCS M81KR VIC has shipped hundreds of thousands of virtual ports since its introduction in Q1CY10.

- The Cisco Nexus 5000 and 2000 Series distributed access architecture has shipped more than 2 million ports since introduction.
- All standards related to virtual networking are in the draft stage at this time.
- Cisco's virtualization offerings continue to evolve and are expanding with the addition of virtual network services and advanced experimental features.

For More Information

Follow these links:

- http://www.cisco.com/en/US/solutions/ns340/ns517/ns224/ns836/ns1124/vm-fex.html
- IEEE 802.1Qbg documentation: http://www.ieee802.org/1/pages/802.1bg.html
- IEEE 802.1BR documentation: http://www.ieee802.org/1/pages/802.1BR.html
- Cisco Virtual Machine Networking Technologies: http://www.cisco.com/en/US/netsol/ns894/index.html
- Cisco Nexus 1000V: http://www.cisco.com/en/US/products/ps9902/
- Cisco Nexus 5000 and 2000 Series distributed access architecture: http://www.cisco.com.az/en/US/prod/collateral/switches/ps9441/ps10110/solution_overview_c22-588237.html
- Cisco UCS M81KR VIC:
    - http://www.cisco.com/en/US/products/ps10331/index.html
    - http://www.cisco.com/en/US/prod/collateral/ps10265/ps10276/solution_overview_c22-555987_ns894_Networking_Solution_Solution_Overview.html
    - http://www.cisco.com/en/US/prod/collateral/ps10265/ps10280/data_sheet_c78-525049.html

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C11-620065-02 08/11