Virtual networking technologies at the server-network edge

Technology brief

Introduction

Hypervisors add a new layer of software and virtual networking that dramatically affects data center servers and their associated network connectivity.

In the past, network administrators managed the external network infrastructure and occasionally managed the server NICs. Server administrators managed the server, the applications running on the server, and usually the server NICs. Hypervisors push the boundary of network infrastructure into the physical server by their use of virtual switches (commonly referred to as soft switches or vSwitches). This blurs the line between the domains of the server administrator and of the network administrator. Server administrators typically configure the vSwitches but can't see or change the external network configurations. Network administrators can't configure or debug the vSwitches.

Other challenges arising from hypervisors include performance loss and management complexity of integrating software-based vSwitches into your existing network management.

Industry leaders are proposing two fundamentally different approaches to deal with server-network edge challenges and to provide more management insight into networking traffic in a virtual machine:

- Edge Virtual Bridging (EVB) with Virtual Ethernet Port Aggregator (VEPA) technology
- Port extension technology

The EVB approach uses industry-standard technologies at the server-network edge. It promotes network management and network service provisioning as close to the edge as possible. The industry-standard approach ensures that new technologies will work within your existing environments and organizational roles. The goal of HP is to enable a simple migration to advanced technologies at the server-network edge without requiring an entire overhaul strategy for your data center.

The port extension approach reflects all network traffic onto a central controlling bridge. This gives network administrators full access and control but at the cost of bandwidth and latency.

The IEEE standards supporting networking in virtual machine (VM) environments are in the final draft stages. It is not clear the extent to which hardware and hypervisor vendors will support these standards. This uncertainty means that whether you are a server administrator or network administrator, you may need to consider numerous factors when choosing new server and network technologies.

Virtual Ethernet Bridges

A Virtual Ethernet Bridge (VEB) is a virtual Ethernet switch that you implement in a virtualized server environment. It is anything that mimics a traditional external Layer 2 (L2) switch or bridge for connecting VMs. VEBs can communicate between VMs on a single physical server, or they can connect VMs to the external network.

The most common implementations of VEBs are software-based vSwitches built into hypervisors. But vendors can use the PCI Single Root I/O Virtualization (SR-IOV) standard to build hardware-based VEBs in NICs.

Software-based VEBs–Virtual Switches

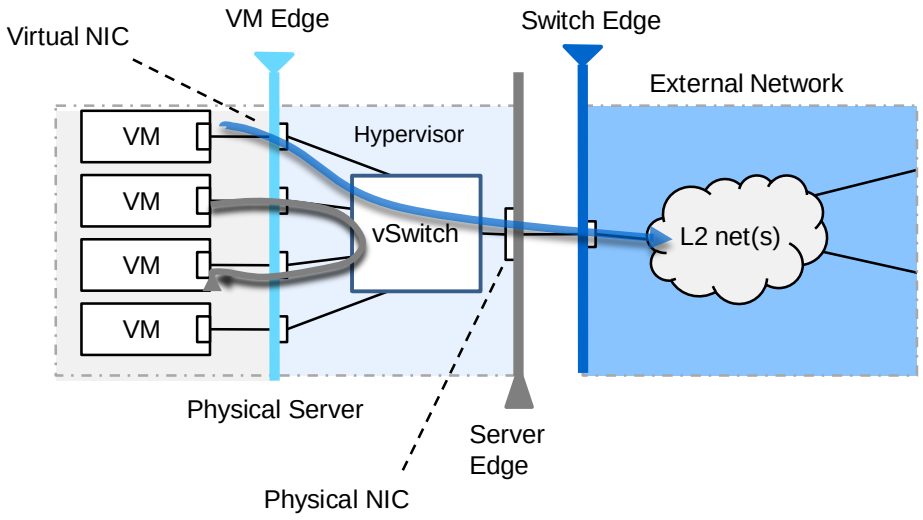
In a virtualized server, the hypervisor abstracts and shares physical NICs among multiple virtual machines, creating virtual NICs for each virtual machine. For the vSwitch, the physical NIC acts as the uplink to the external network. The hypervisor implements one or more software-based virtual switches that connect the virtual NICs to the physical NICs.

Data traffic received by a physical NIC passes to a vSwitch. The vSwitch uses its hypervisor-based configuration information to forward the traffic to the correct VMs.

When a VM transmits traffic from its virtual NIC, a vSwitch forwards the traffic in one of two ways (see Figure 1):

- If the destination is external to the physical server or to a different vSwitch, the vSwitch forwards the traffic to the physical NIC.
- If the destination is internal to the physical server on the same vSwitch, the vSwitch forwards the traffic directly back to another VM.

Figure 1: In a VEB implemented as a vSwitch, traffic can be "switched" locally inside the Hypervisor vSwitch (gray line) or sent directly to the external network via the physical NIC (blue line).



Using a software-based vSwitch has a number of advantages:

- **Good performance between VMs.** A vSwitch typically uses only L2 switching and can forward internal VM-to-VM traffic directly. Bandwidth is restricted only by available CPU cycles, memory bus bandwidth, or limits configured by the user in the hypervisor.
- **Deployment without an external switch.** Administrators can provide an internal network with no external connectivity. For example, you can run a local network between a web server and a firewall application running on separate VMs within the same physical server.
- **Support for a wide variety of external network environments.** A vSwitch is compliant with standards and can work with any external network infrastructure.

A vSwitch also has several disadvantages:

- **Consumption of valuable CPU and memory bandwidth.** The more traffic, the greater the number of CPU and memory cycles required to move traffic through the vSwitch. This reduces the ability of the vSwitch to support larger numbers of VMs in a physical server.
- **Lack of network-based visibility.** The vSwitch has a limited feature set. It doesn't provide local traffic visibility nor does it have capabilities for enterprise data monitoring, security, or network management. This can affect network policies in the data center for accounting, security, and reliability.

- **Lack of network policy enforcement.** Modern external switches have many advanced features such as port security, quality of service (QoS), and access control lists (ACL). But a vSwitch often does not have, or have limited support for, such features. Even if a vSwitch supports advanced features, its management process is often inconsistent or incompatible with that of external networks. This limits your ability to create end-to-end network policies within a data center.
- **Lack of management scalability.** When you increase the number of VMs in a data center, the number of vSwitches also expands. You must manage standard vSwitches individually. VMware has introduced distributed virtual switches that allow you to manage up to 64 vSwitches as a single device, but this only addresses the management scalability problem. It does not resolve the lack of management visibility outside the virtualized server.

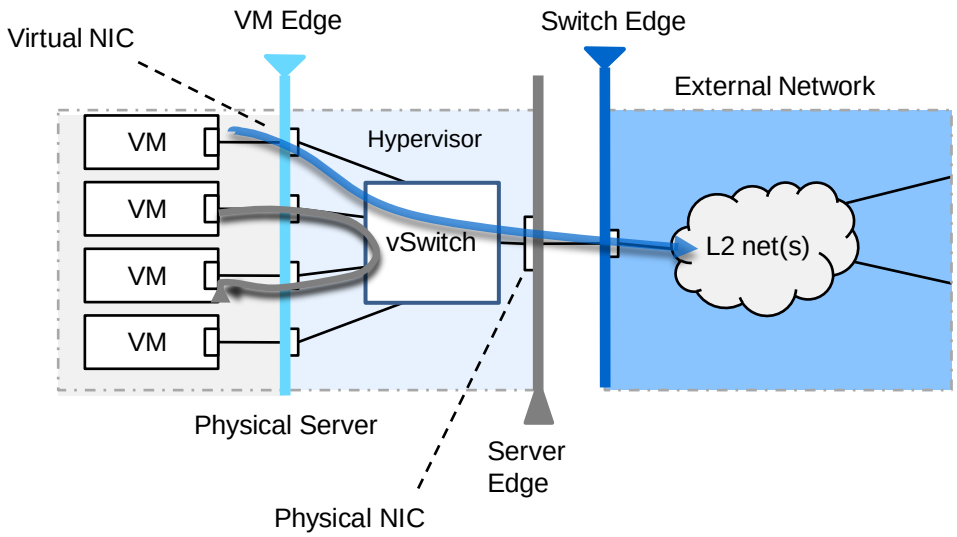
Hardware VEBs—SR-IOV enabled NICs

SR-IOV technology lets vendors deploy a VEB in the NIC hardware. Moving VEB functionality into the hardware reduces the performance issues associated with vSwitches.

The SR-IOV standard provides native I/O virtualization for shared PCIe devices on a single physical server. Typically, SR-IOV-enabled NICs support 1 to16 physical functions (full-featured PCI functions) and 32 to 256 virtual functions (lightweight PCI functions focused primarily on data movement). These are today's typical numbers. The SR-IOV standard allows for thousands of virtual functions in a device, providing headroom for future capabilities.

SR-IOV-enabled NICs let the virtual NICs bypass the hypervisor vSwitch by exposing the virtual NIC functions directly to the guest OS. Thus, the NIC reduces latency between the VM to the external port significantly. The hypervisor continues to allocate resources and handle exception conditions, but it doesn't need to perform routine data processing for traffic between the VMs and the NIC. Figure 2 illustrates the traffic flow of an SR-IOV-enabled NIC.

Figure 2: In a VEB implemented as an SR-IOV NIC, traffic flows the same way as with a vSwitch. Traffic can switch locally inside the VEB (gray line) or go directly to the external network (blue line).



The benefits to deploying VEBs as hardware-based SR-IOV-enabled NICs include:

- **Reduction of CPU and memory usage compared to software-based vSwitches.** With direct I/O, vSwitches are no longer part of the data path.
- **Support of up to 256 functions in a typical low-cost NIC.** It significantly increases the number of virtual networking functions for a single physical server.

While SR-IOV brings improvements over traditional software-only vSwitches, there are still challenges with SR-IOV NICs, including:

- **Lack of network-based visibility.** SR-IOV NICs do not solve the network visibility problem. In fact, because of limited resources of cost-effective NIC silicon, the embedded VEB may have even fewer capabilities than a vSwitch.
- **Lack of network policy enforcement.** Vendors typically don't include advanced policy features because of the limited silicon resources in cost effective SR-IOV-enabled NICs.
- **Lack of management scalability.** Network administrators still must manage SR-IOV- based NICs independently from the external network infrastructure. Also, SR-IOV-enabled devices typically have one VEB per port, unlike software-based vSwitches that can operate multiple NICs and NIC ports per VEB. Thus, you may have more VEBs to manage.
- **Requirement for a paravirtualized driver.** SR-IOV requires a guest OS to have a paravirtualized driver to support the direct I/O with the PCI virtual functions. Currently, only Xen and KVM hypervisors support SR-IOV NICs. Microsoft has announced plans to support SR-IOV in its upcoming Windows Server 8 implementation of Hyper-V.

Edge Virtual Bridging

Today, neither software vSwitches nor hardware VEBs in SR-IOV devices can achieve the level of network capabilities built into enterprise-class L2 data center switches. To solve the management challenges with VEBs, HP is working with other vendors to develop Edge Virtual Bridging (EVB) in the IEEE 802.1Qbg standard. The primary goals of EVB are to combine the best of software and hardware VEBs with the best of external L2 network switches.

The IEEE 802.1Qbg standard formalizes the following concepts and protocols:

- The edge relay concept, which can be either a Virtual Edge Port Aggregator (VEPA) or a VEB
- The reflective relay operation for switches
- An optional S-channel operation
- The Virtual Switch Interface Discovery and Configuration Protocol (VDP) for configuring VM network connections to external networks

Virtual Ethernet Port Aggregator technology

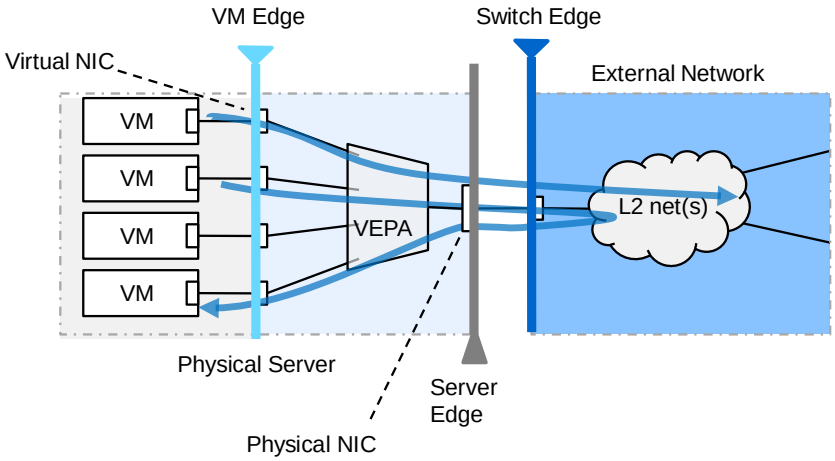
EVB is based on VEPA technology. It is a way for virtual switches to send all traffic and forwarding decisions to the adjacent physical switch. This removes the burden of VM forwarding decisions and network operations from the host CPU. It also leverages the advanced management capabilities in the access or aggregation layer switches.

Traffic between VMs within a virtualized server travels to the external switch and back through a reflective relay, or 180-degree turn (Figure 3, bottom traffic flow).

VEPA does not require new tags and involves only slight modifications to VEB operation, primarily in frame relay support. VEPA continues to use MAC addresses and standard IEEE 802.1Q VLAN tags as

the basis for frame forwarding, but changes the forwarding rules slightly according to the base EVB requirements.

Figure 3: Traffic flow with an EVB using VEPA mode goes to the adjacent network.



There are many benefits to using EVB/VEPA:

- Reduces the server's CPU and memory usage from the processing overhead related to moving I/O traffic through the vSwitch.
- Lets the adjacent switch perform the advanced management functions, so that the NIC can use low-cost circuitry.
- Moves the VM control point into the edge physical switch (top-of-rack or end-of-row switch). VEPA leverages existing investments made in data center edge switching. Administrators can manage the edge network traffic using existing network security policies and tools.
- Gives better visibility and access to external switch features from the guest OS. Network administrators can view frame processing (ACLs) and security features such as Dynamic Host Configuration Protocol guard, address resolution protocol (ARP), ARP monitoring, source port filtering, and dynamic ARP protection and inspection.
- Can be implemented in the hypervisor or with embedded hardware (for example, by using an SR-IOV NIC with hypervisor bypass). Either case requires hypervisor support.

Vendors will be able to implement software-based VEPA solutions as simple upgrades to existing vSwitches. As a proof-point, HP Labs worked with the University of California, Davis to develop a software prototype of a VEPA-enabled switch. Even without being fully optimized, the VEPA prototype was 12% more efficient than the traditional software vSwitch in environments with advanced network features enabled. See "A Case for VEPA: Virtual Ethernet Port Aggregator" at www.i-teletraffic.org/itc22/workshops/dc-caves-workshop for more information.

This prototype VEPA switch was a simple firewall solution. The performance gains will increase if a VEPA switch is offloading more complex networking services like ACLs and packet filtering. Also, you could combine VEPA with SR-IOV-enabled NICs to increase your server performance even more.

Because EVB/VEPA traffic goes deeper into the network, there is some performance reduction. VM-to-VM traffic must flow to the external switch and back—consuming twice the communication bandwidth. This only occurs for co-located VMs on the same host, in the same broadcast domain, and

in direct communication with each other. If the need for local bandwidth outweighs the need for visibility or control of network traffic, it makes sense to use VEB mode.

VEB mode

EVB is not necessarily a replacement for all VEB operating modes. The EVB standard supports VEPA-based switches and existing VEB (vSwitch) architectures simultaneously. IT architects can choose whether to manage the server-network edge traffic in the local hypervisor (VEB vSwitch) or in the adjacent physical switch (VEPA-based switch).

S-channel technology

VEPA technology alone does not satisfy all use cases. S-channel technology adds an enhanced tagging mechanism to the basic VEPA technology. S-channel benefits the following use cases:

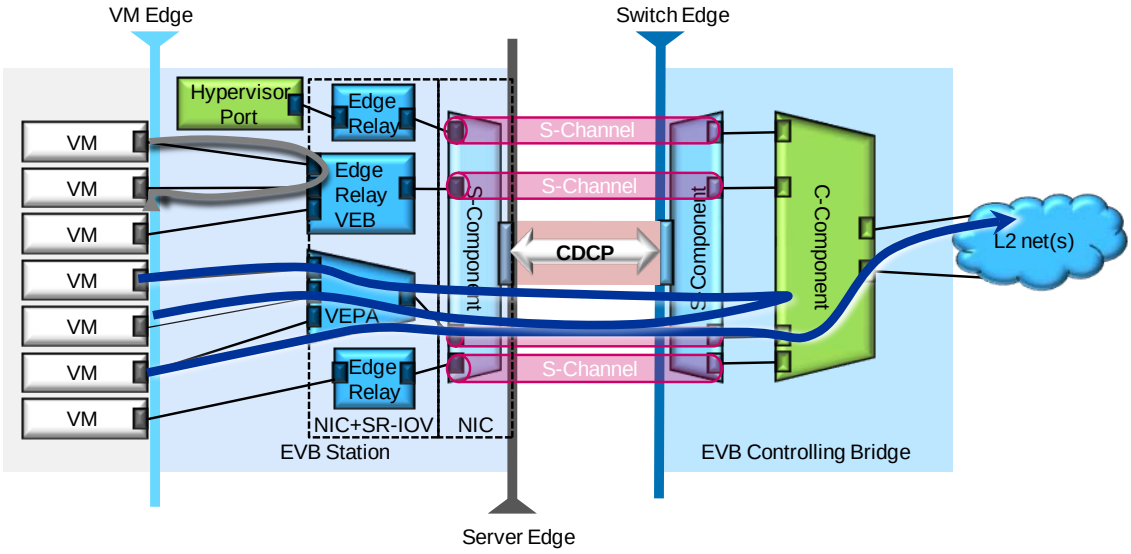
- Hypervisor functions that need direct access to the hardware NIC
- VMs that require direct access to the hardware NIC
- Sharing a physical network connection between multiple virtual switch types (VEB and VEPA) to optimize local, VM-to-VM performance.
- Directly mapping a VM that requires promiscuous mode operation.

S-channel technology uses existing Service VLAN tags (S-Tags) from the "Provider Bridge" or "Q-in-Q" standard (IEEE 802.1ad). The VLAN tags let you logically separate traffic on a physical network connection or port (like a NIC device) into multiple channels. Each logical channel operates as an independent connection to the external network.

S-channel uses the Channel Discovery and Configuration Protocol (CDCP) to provision and configure the logical channels to the NIC. It uses Link-Layer Discovery Protocol and enhances it for servers and external switches.

Figure 4 illustrates how S-channel technology uses these capabilities within a virtualized server.

Figure 4: S-channel technology supports VEB, VEPA, and direct-mapped VMs on a single NIC.



Vendors can take advantage of simple VEPA operation without supporting S-channel. S-channel merely enables more complex virtual network configurations in servers using VMs. You can assign each of the logical channels to any type of virtual switch (VEB, VEPA, or directly mapped to any virtual machine within the server). This lets IT architects match their application requirements with the design of their specific network infrastructure:

- VEB for performance of VM-to-VM traffic
- VEPA/EVB for management visibility of the VM-to-VM traffic
- Sharing physical NICs with direct mapped virtual machines
- Optimized support for promiscuous mode applications

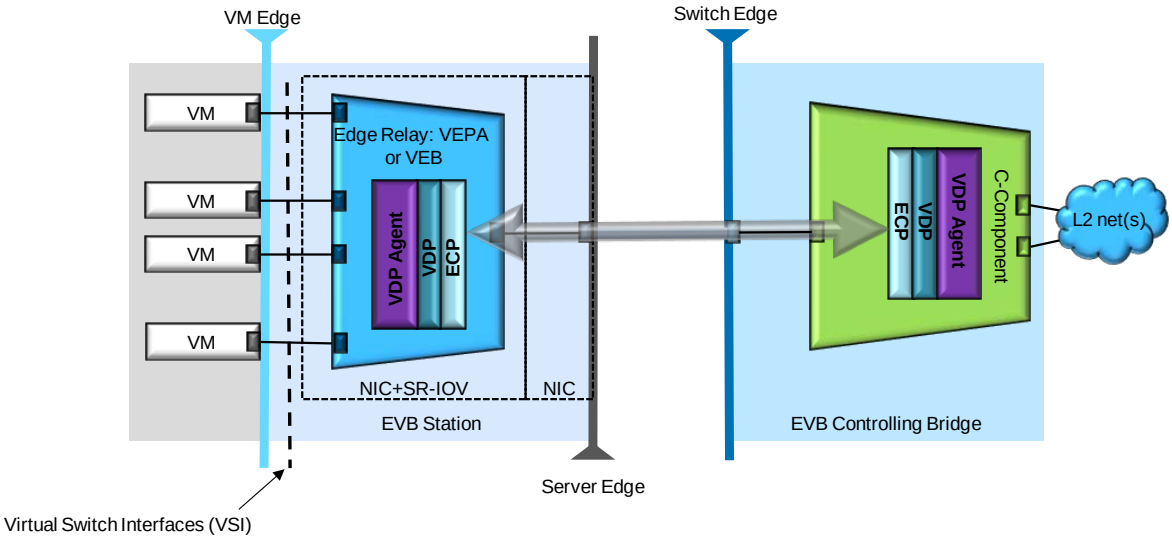
Virtual Switch Interface Discovery and Configuration Protocol

Virtual Switch Interface Discovery and Configuration Protocol (VDP) and its underlying Edge Control Protocol (ECP) provide a virtual switch interface that sends the required attributes for physical and virtual connections to the controlling bridge (Figure 5). VDP/ECP also lets the controlling bridge validate connections and provides the appropriate resources.

The VDP/ECP protocols:

- Facilitate the movement of VMs within the data center
- Support VEB and VEPA modes of edge relay
- Support S-Channel and legacy physical NIC-switch connections

Figure 5: VDP/ECP protocols operate in the context of edge relays (VEBs and VEPAs) to instantiate the virtual switch interfaces and request connection services from the adjoining controlling bridge.

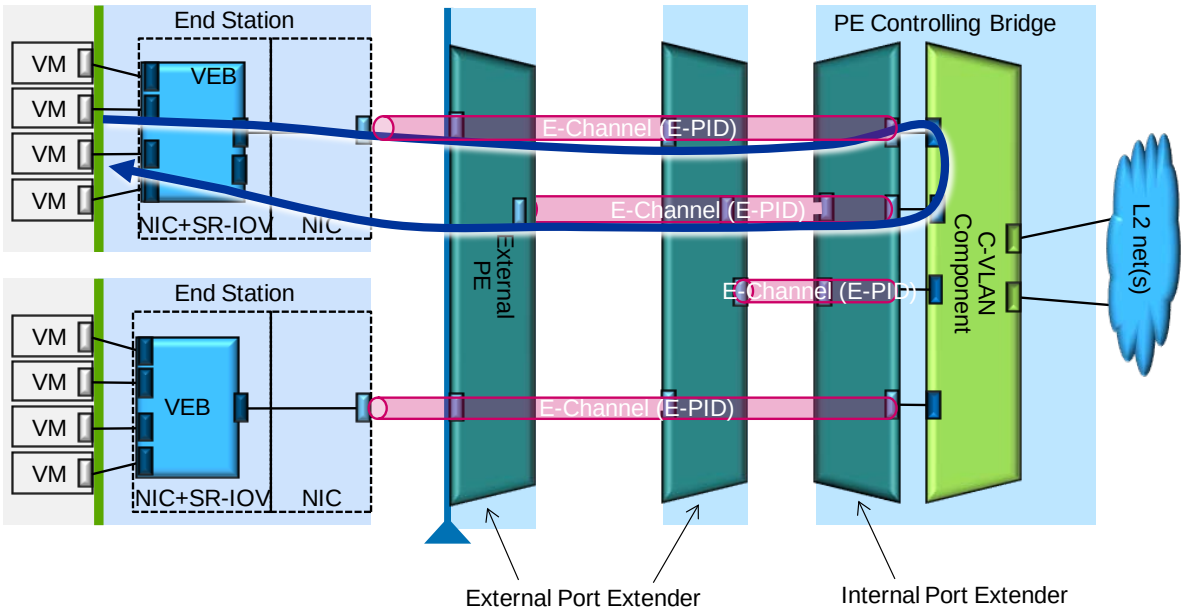


Port extension technology

Port extension technology allows network designers to extend a network switch or controlling bridge with a physical switch of limited functionality. The external network switch connects to an external port extender using logical E-channels (Figure 6). These logical channels appear as virtual ports in the

external network switch. Because the port extender has limited functionality, the external network switch manages all virtual ports and their associated traffic.

Figure 6: Port extender architectures expand the network switch with an external port extender, which can be cascaded.



Port extenders either use existing proprietary Cisco technology with VN-tags or will use the upcoming E-tag from the draft IEEE 802.1 BR Port Extension specification. The E-tag is longer than the VN-tag. It has different field definitions and different field locations but serves the same purpose. Figure 7 shows how the nomenclature changed between the Cisco VN-tag and the E-tag.

Figure 7: The proprietary VN-tag format existed before the 802.1 BR E-tag and uses a different format.

VN-Tag

VN-Tag EtherType			
d	p	Destination Virtual I/F ID *(Extended Port ID in E-Tag)*	
l	r	ver	Source Virtual I/F ID *(Ingress Extended Port ID in E-tag)*

E-Tag

E-Tag EtherType		
PCP	DEI	Ingress Extended Port ID *(Source Virtual I/F ID in VN-Tag)*
r		Extended Port ID *(Destination Virtual I/F ID in VN-Tag)*
r		

VN-Tag Specific Fields
VN-Tag EtherType
d - Direction Flag
p - Pointer Flag
l - Looped Flag
ver - Version of Tag

Common Fields
Destination Virtual I/F ID = Extended Port ID
Source Virtual I/F ID = Ingress Extended Port ID
r = Reserved Field

E-Tag Specific Fields
E-Tag EtherType
PCP - Priority Bits
DEI - Drop Eligible Indicator

Port extenders use the information in VN-tags or 802.1 BR E-tags to:

- Map the physical ports on the port extenders as virtual ports on the upstream switches
- Control how they forward frames to or from upstream switches
- Control how they replicate broadcast or multicast traffic

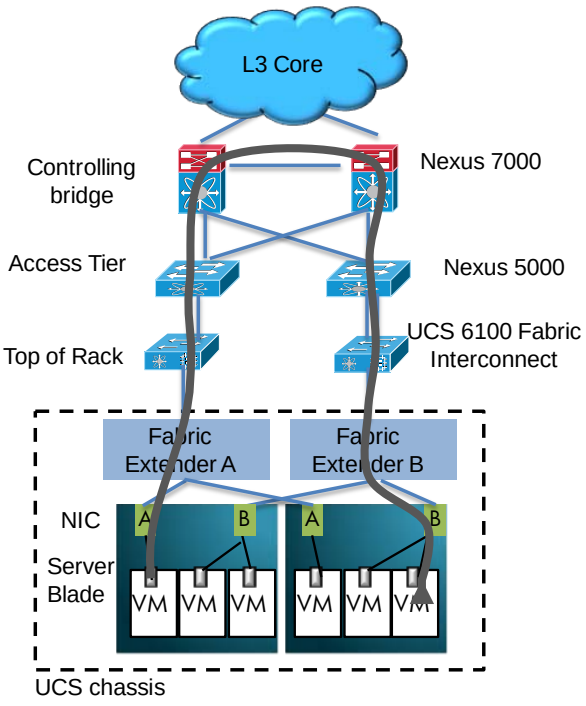
Port extension solves the problem of network management visibility into the virtual networking functions by reflecting all network traffic onto a central controlling bridge. This gives network administrators full access and control but at the cost of bandwidth and latency.

Disadvantages of port extension approach

Port extension technology adds yet another layer to the existing hierarchical network, forcing packets to go across multiple hops on the network. Products such as the Cisco Nexus Fabric Extenders and Cisco UCS Fabric Extenders (FEX) are examples of port extenders using Cisco's proprietary VN-tags. Let's look at the Cisco UCS architecture.

Cisco UCS uses a Fabric Interconnect and recommends that you configure it in "End Host Mode." This means VM-to-VM traffic must travel from the NIC A to FEX A to the Fabric Interconnect, to an upstream switch, back to Fabric Interconnect B, to FEX B and finally back to NIC B (Figure 8). When the architecture is already oversubscribed, this adds even more congestion to the network and aggravates the oversubscription problem.

Figure 8: Port extension technology adds one or more extra hops to the typical three-tier architecture and can magnify congestion problems. This diagram shows two extra hops.



As data centers support more clustered, virtualized, and cloud-based applications requiring high performance across hundreds or thousands of physical and virtual servers, port extension technology just seems to add cost and complexity.

Remember that the pre-standard VN-tags and the IEEE 802.1-BR standard E-tags use different formats. If you adopt VN-tag solutions in your data center, you will have to develop transition strategies when future hardware changes to the IEEE 802.1-BR E-tag format.

Status of IEEE standards and industry adoption

The IEEE is working to ratify both the 802.1Qbg EVB standard and the 802.1-BR Port Extension standard by mid-2012. IEEE has withdrawn the previous 802.1 Qbh port extension standard.

We expect hardware vendors to adopt EVB quickly because implementing EVB requires few hardware changes. NIC, CNA, and switch vendors are already planning to support EVB in the next generation of devices.

Official IEEE port extension adoption might be slower, because existing VN-tag formats in shipping hardware today differ from the proposed IEEE E-tag format.

As of this writing, it's unclear which hypervisor vendors will support either EVB or port extension technologies. If you are interested in using either of these technologies, contact your hypervisor vendor to understand their strategic direction.

Conclusion

Virtual networking technologies are becoming increasingly important and complex because of the growing use of virtual machines, distributed applications, and cloud infrastructures.

Current vSwitches are software entities, created and managed inside hypervisors. As such, you cannot manage hypervisor vSwitches in the same ways nor maintain the same level of visibility into the virtual network that you have today with physical network switches.

Moving the network virtualization into hardware (SR-IOV NICs) relieves the performance problems associated with vSwitches but does not help the management and visibility issues.

HP is promoting VEPA technology to solve the network management and visibility problems caused by virtual networking. Using VEPA technology shifts most of the network processing activities close to the server-network edge, just inside the dedicated network fabric. This lets the lowest level of network switches—the access-layer switches—manage the virtual network traffic. You can also combine VEPA with SR-IOV-enabled NICs to gain even more performance improvements than the individual technologies.

Other companies such as Cisco are promoting port extender technology to solve the server-edge problem. Port extender technology shifts the network processing further up into the aggregation or core level switches. The two approaches are fundamentally different in how they address the challenges with virtual networking.

Finally, because hypervisor support for either approach is still emerging, IT managers should carefully evaluate their choices as they move forward with virtualization and cloud computing. HP is committed to working with partners and customers to develop virtual network technologies that address a broad range of customer requirements at the server-network edge.

For more information

Visit the URLs listed below if you need additional information.

Resource description	Web address
HP Industry Standard Server technology communications white papers	www.hp.com/servers/technology
HP Virtual Connect technology	www.hp.com/go/virtualconnect
IEEE 802.1 Work Group website	www.ieee802.org/1/
IEEE Edge Virtual Bridging standard 802.1Qbg	www.ieee802.org/1/pages/802.1bg.html
IEEE Bridge Port Extension standard	www.ieee802.org/1/pages/802.1br.html

Send comments about this paper to TechCom@HP.com


Follow us on Twitter: http://twitter.com/ISSGeekatHP

TC0000758, November 2011