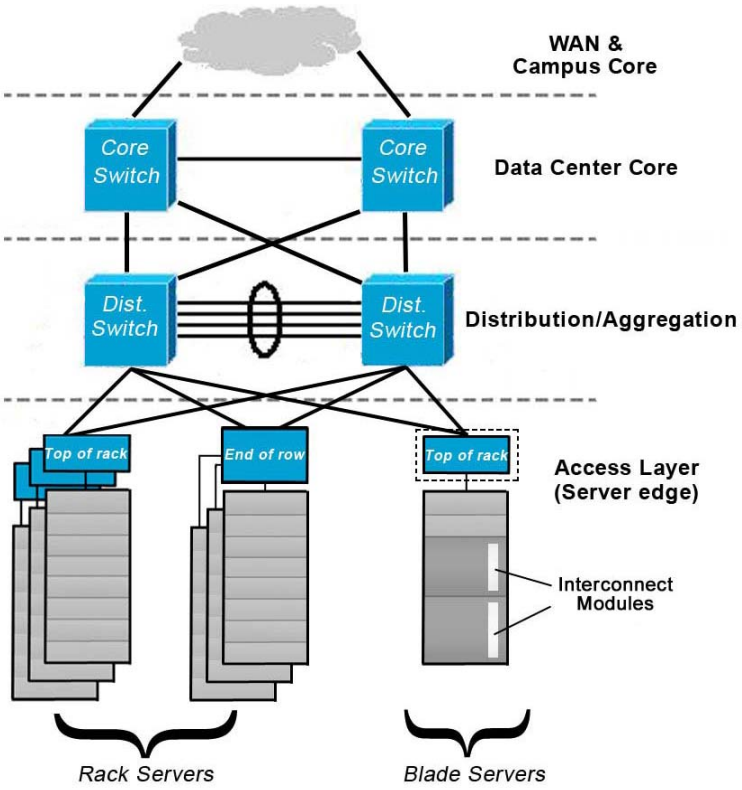
Effects of virtualization and cloud computing on data center networks

Technology brief

Introduction

Most data centers today have a three- or four-tier hierarchical networking structure. It consists of access layer switches, aggregation switches, and core switches (Figure 1). Three-tier networking architectures were designed around client-server applications and single-purpose application servers. Client-server applications caused traffic to flow primarily in North/South (N/S) patterns: from a server up to the data center core, to the campus core where it moves out to the campus-wide network or internet. These large core switches usually contain the vast majority of the intelligence in the network.

Figure 1: A typical data center structure today uses three layers: access, aggregation, and core.



The dotted line around the Blade server/Top of rack (ToR) switch indicates an optional layer, depending on whether the interconnect modules replace the ToR or add a tier.

This network architecture, however, is becoming problematic for the data center. Today's application environments are more distributed, often with multiple tiers, and oriented toward service delivery. These application architecture changes have resulted in:

- Greater traffic volume on the Ethernet network, including storage traffic such as FCoE and iSCSI
- More storage traffic as applications use distributed file systems and increase the amount of synchronization and replication data across the network
- Greater traffic flow between peer servers such as server-to-server or virtual machine-to-virtual machine—that is, East/West (E/W) rather than primarily N/S traffic flows.

This paper won't answer all your questions about the future of data center networks, but it will identify some industry trends and present some possible solutions. We suggest that you consider adopting technologies that

- Reduce network hierarchy
- Optimize E/W traffic flows
- Simplify operations at the server-network edge with intelligent management capabilities that align with the needs of all operational groups in the data center

Changing business applications

The growth of server virtualization (virtual machines or VMs), virtualized desktop infrastructures, cloud-computing models, federated or distributed applications, and mobile access devices are all causing shifts in networking traffic patterns toward more E/W traffic flow (Table 1). Industry sources attribute up to 80 percent of network traffic for these next generation applications coming from E/W traffic flows.

We expect web applications to deliver integrated, context-specific information and services. And, we expect it right now—low-latency, high performance connections are critical. At the same time, cloud computing and service-oriented applications are introducing more stringent service-level and security demands.

Table 1: New software applications are driving changes to networking infrastructure.

Yesterday's applications	**Applications for 2011 and beyond**
Single application on single-purpose server	Multiple applications operating on VMs within a single physical server.
Client-server architecture	Distributed computing applications (massively parallel compute clusters) Clusters of multiple servers in compute resource pools, requiring server mobility within a cluster and requiring resources across clusters.
Static deployment model	Cloud computing and new service delivery models Platform as a service (PaaS) Software as a service (SaaS) Infrastructure as a service (IaaS) Storage as a service (STaaS) "X" as a service (XaaS)

Server virtualization

Since the introduction of hypervisors over a decade ago, the increase in VM density (fostered by ever more powerful CPU and memory subsystems) and the significant increase in VM mobility have resulted in greater performance demands on the network subsystems at the server-network edge. Table 2 outlines how these trends are causing traffic flow patterns to shift.

Table 2: Changing realities of VMs in x86 environments.

Early deployments	2011 and beyond
Consolidate two or three VMs per physical server	10-50+ VMs per server
Network bandwidth could be easily shared by a few VMs	Higher network bandwidth requirements driven by: Greater VM density per physical server More powerful CPU-memory systems VM mobility
VM operates statically. Moving VM workloads is a rare event, such as when: - Decommissioning servers - Moving to higher performing server	Dynamic workload placement is common. Moving VM workloads occurs for events such as: - Time-based VM creation for handling peak workloads - Time-based workload shifting for optimizing power and performance - Automated tools for facilitating workload placement
VM workload remains inside a data center	Disaster recovery requires VM workloads to move across physical locations (for example, to a disaster recovery data center)

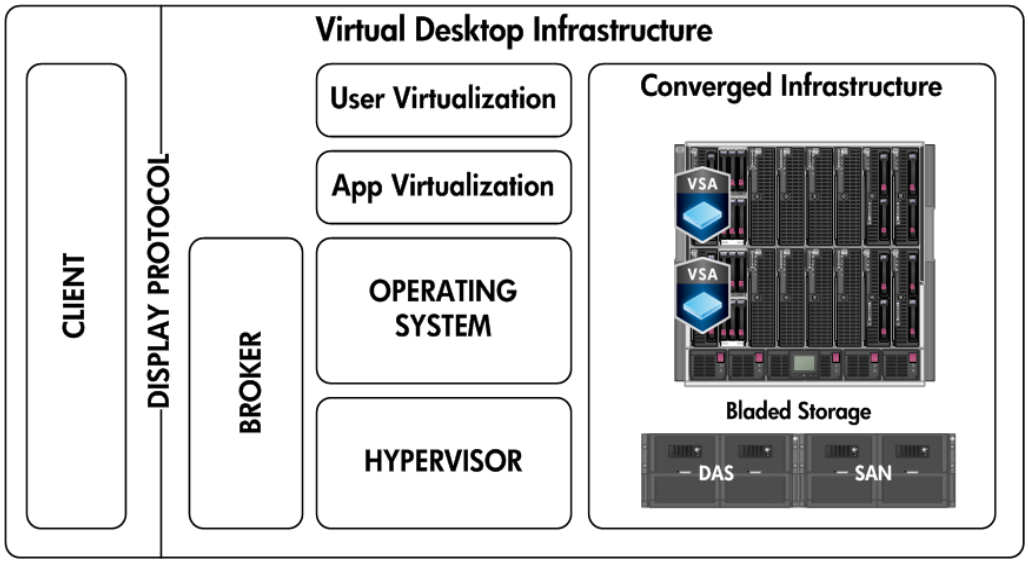
Moving workloads dynamically requires VMs to stay within a common VLAN in the same Layer 2 (L2) network. If you want to move a VM outside its L2 domain, you have to use manual processes such as assigning and updating the IP addresses for the failed-over services and updating DNS entries correctly. To provide maximum VM flexibility, many enterprises are evaluating ways to enlarge their L2 networks.

New capabilities such as Virtual eXtensible LAN (VXLAN) and Network Virtualization using Generic Routing Encapsulation (NVGRE) logically extend an L2 network across L3 networks. However, even with this potential to move VMs across a L3 network, local traffic will still have higher performance and lower latency if it stays within a large L2 network.

Client virtualization

A specialized type of VM is the client virtualization technology such as virtual desktop infrastructure (VDI). VDI creates a client desktop as a VM. The VDI instance, however, is more than a simple VM. It includes the real-time compilation of the end user's data, personal settings, and application settings with a core OS instance and a shared generic profile. You can either install the end-user applications locally as a fully packaged instance or stream them from outside the VM. Applications and user personality are injected into the core desktop VM, and a brokering mechanism manages connecting the end users to the VM (Figure 2).

Figure 2: Architectural overview of VDI with blade servers, storage, and HP Virtual SAN Appliance (VSA) included for the infrastructure.



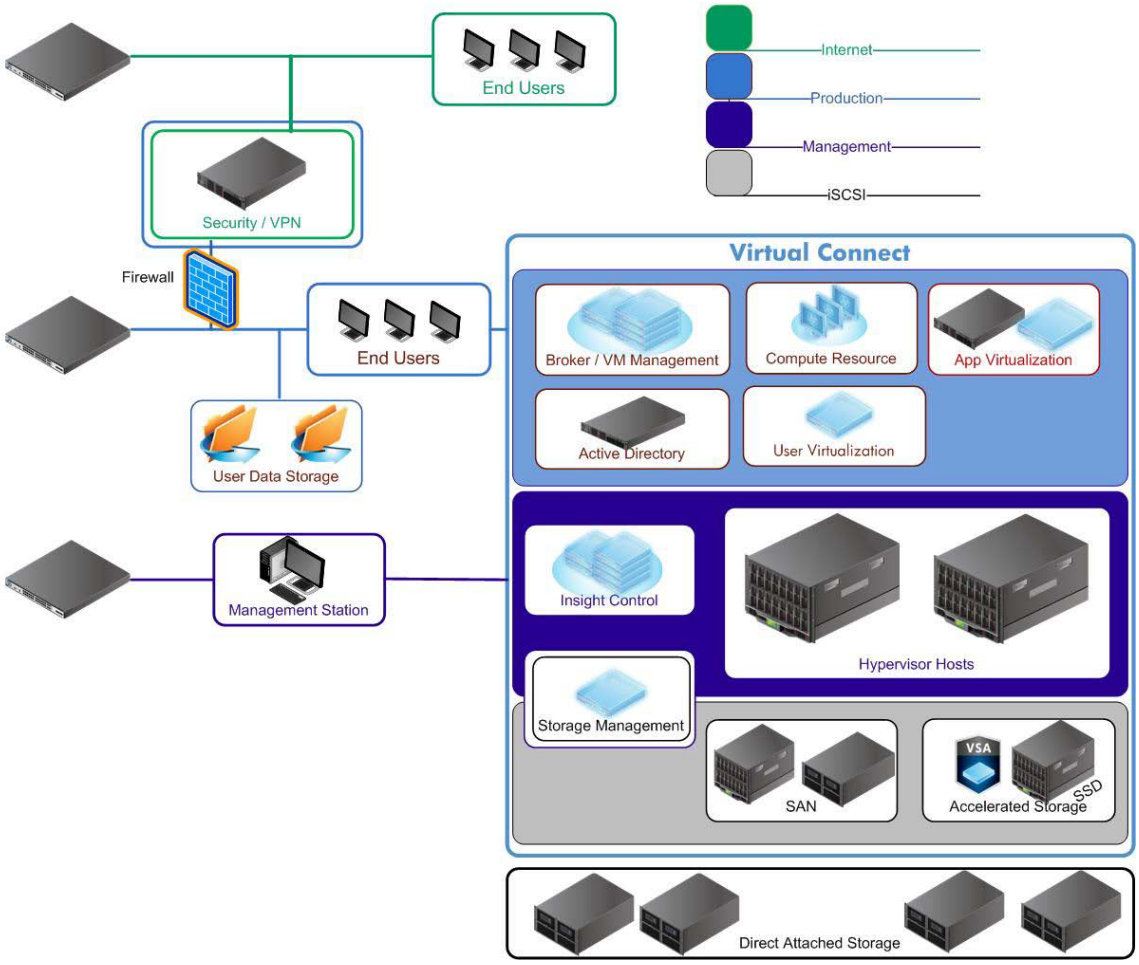
For example, a VDI could let you access a device based on a Microsoft Windows Vista OS, leave for the evening, and come back the next workday to a Microsoft Windows 7-based device—with all your data, customized desktop settings, and customized application settings intact.

A standard VDI configuration would use rack-based servers distributed across the data center with Top of Rack (ToR) switches at the network edge. Network traffic from each rack of the distributed servers (for example, Microsoft Exchange servers, Active Directory servers, user application servers, or the VDI servers) would travel to its own ToR switch before traveling to the network core and then out to the client.

But HP has designed VDI reference architectures on HP BladeSystem with Virtual Connect hardware. This keeps the majority of network traffic inside the Virtual Connect domain, as local E/W traffic that never travels out to the network core (Figure 3). Only a small, well-defined amount of traffic for the connection and management protocols exits the Virtual Connect domain. The HP design:

- Optimizes the E/W traffic
- Minimizes the need for expensive switch ports
- Lets a single infrastructure administrator manage the intra-domain traffic
- Improves performance and reliability by using mostly cable-free internal chassis connections between hosts and management services

Figure 3: When using a VDI configuration Virtual Connect technology, only a small amount of production (protocol) and management traffic exit the Virtual Connect domain, thus optimizing the E/W traffic flow.



Cloud applications

Enterprise businesses are moving beyond server virtualization and VMs to embrace cloud-computing environments. The solution stack for these enterprise cloud-computing environments often includes infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). These "X as a service," on-demand models require flexibility and the need for immediate growth. So, enterprises often build private clouds on virtualization and large L2 domains to allow live migration of VMs across as large a domain as possible. Cloud models, especially in service provider environments, also require a multi-tenancy infrastructure to provide separate and secure services to many customers simultaneously.

Some public cloud applications and service provider environments use cluster-like architectures, with massively parallel workload and data distribution characteristics. Common cluster applications include data aggregation and "big data" analytics applications like Hadoop, Needlebase, Platform Computing's Symphony software, or Vertica software. As a request comes in, a task scheduler spawns multiple jobs to multiple servers—causing a flurry of network traffic that does not go up to the network core but out to peer servers. Social networking sites use services like memcache for distributing memory objects to alleviate database load and speed up performance. Applications like Swift and services like Amazon's Simple Storage Service (S3) distribute storage across multiple nodes.

Distributing storage, distributing databases across multiple servers, sending requests to multiple servers, and accumulating the responses are all E/W traffic intensive.

For example, consider how you plan a vacation. You visit the dynamic travel website of your choice and enter your variable data (when you want to travel, where, whether you need a hotel, flight, or a car). The site pulls together the appropriate responses from multiple databases, along with related ads, and shows you the options within a matter of seconds. Not only is this process very heavy in E/W traffic flow because it pulls data from multiple servers, it is also latency sensitive. If a travel website cannot serve the data to you within a matter of seconds, you're likely to go to a competitor.

Mobile access devices

Finally, consider the effect of mobile access devices on data center traffic. There are hundreds of thousands of smartphone applications. These applications use a thin client that pulls much of the application and data from private or public clouds in a data center. It puts tremendous loads on the data center's Ethernet fabrics. These E/W traffic loads are not only bandwidth sensitive; they are also latency-sensitive. Many internet-based applications like travel websites have a limited time window for the back-end applications to retrieve requested data. If your network infrastructure cannot handle these traffic loads, you will have inadequate application responses, resulting in customers moving on to a competitor's services.

Limitations of a hierarchical networking structure

The more E/W traffic you have in a network, the more limitations you may face with a hierarchical network structure designed primarily for N/S flow. The challenges include traditional Spanning Tree Protocol (STP) limitations, oversubscription, port extension technology, and increased latency.

STP limitations

STP detects and prevents loops in L2 networks. Loops are an undesirable situation that can occur when there are multiple active paths between any pair of non-adjacent switches in the network. (Multiple paths between adjacent switches can use link aggregation technology such as 802.3AD LACP). To eliminate loops, STP allows only one active path from one switch to another. If the active path fails, STP automatically selects a backup connection and makes that the active path. Thus, STP blocks all parallel paths to a destination except the one it has selected as active, regardless of how many actual connection paths might exist in the network. Even when the network is operating normally, STP usually reduces the effective available bandwidth by 50 % or more. The process to activate new links can be time-consuming, often taking considerable time to re-converge on a new path.

As businesses move away from client- server applications to more dynamic, latency-sensitive applications, the limitations of STP-based protocols become more burdensome. As E/W traffic volume increases, so does the need to use all available bandwidth and links. STP itself has no capability to do dynamic load balancing over multiple paths. Enhancements to STP such as Rapid Spanning Tree Protocol (RSTP) and Multiple Spanning Tree Protocol (MSTP) help resolve some of these issues, but at the cost of complex manual management. It's clear the industry requires a new approach.

Oversubscription

Depending on the data center architecture you choose, oversubscription can be a problem. For example, if you use the Cisco Universal Computing System (UCS) architecture, you may have oversubscription rates of anywhere from 4:1 to 32:1 into the aggregation layer. Oversubscription can be an especially critical issue if your applications cause a lot of storage movement because of large

file block sizes. If your network is oversubscribed, your ability to do live migrations might also be compromised because of the large bandwidth capabilities required.

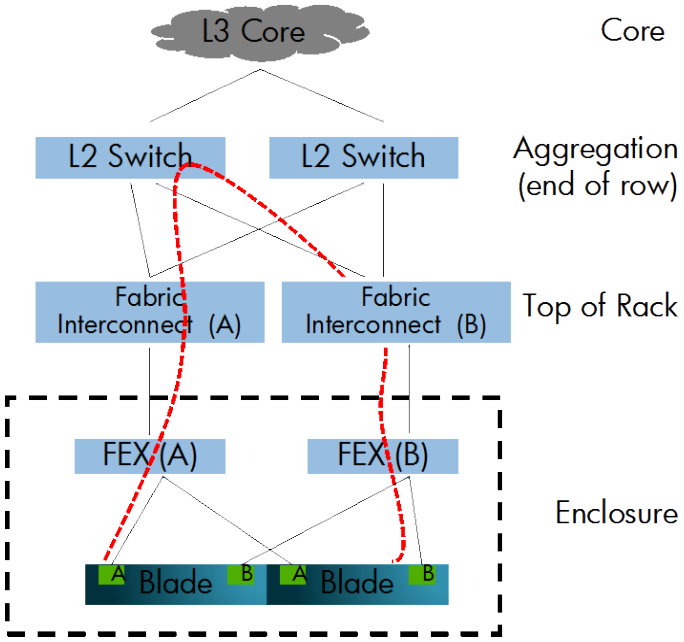
Oversubscription can also lead to requirements for controlling traffic with QoS mechanisms such as enhanced transmission selection, minimum bandwidth guarantees, and maximum rate limits. Most of these mechanisms are manual processes and require a substantial amount of management. Using QoS schemes to manage a scarce bandwidth resource increases complexity in your network and expands the associated management overhead.

Port Extension technology

IEEE is developing port extension technology as part of the draft IEEE P802.1BR, Bridge Port Extension standard. It introduces a device called a "port extender" that is essentially a physical switch with limited functionality, managed as a line card of the upstream physical switch. Products such as the Cisco Nexus Fabric Extenders (FEX) and Cisco UCS FEX are examples of port extenders.

Port extension technology extends the difficulties of the existing hierarchical network by adding yet another layer, forcing packets to go across multiple "hops" on the network. For example, Cisco recommends that you configure the Fabric interconnect in "End Host Mode" in its UCS system. Using this mode, VM-to-VM traffic in a UCS blade enclosure must travel from the NIC-A to FEX A to Fabric Interconnect A up to an upstream switch back to Fabric Interconnect B, to FEX B and finally back to NIC B (Figure 4). If the architecture is already oversubscribed, it adds even more congestion to the network and aggravates the oversubscription problem.

Figure 4: Port extension technology adds an extra "hop" to the typical three-tier architecture and can magnify congestion problems.



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.

Latency

In many cases, latency may be more of a challenge than oversubscription and raw bandwidth. As pointed out in the "Changing Business Applications" section, as businesses need to supply immediate, context-sensitive information to end users (remember that travel website on page 7?), latency and application responsiveness are driving network designs. Traditional hierarchical and newer architectures like the fabric extension technology require more network hops and increase latency. When running even moderate levels of traffic in these systems, latencies increase because of the multiple hops through the congested and oversubscribed points in the network.

Practical solutions for optimizing E/W traffic flow

This section describes some different architectures and technologies to consider when optimizing your data center structure for E/W traffic flows. They include:

- Fostering E/W traffic flow at the physical server-network edge
- Distributing management intelligence at the physical server-network edge rather than concentrating it at higher layers in your network
- Flattening your L2 network by using technologies like HP Intelligent Resilient Framework™ (IRF)
- Making your L2 network more efficient by implementing future multi-path standards

Identify your traffic bottlenecks

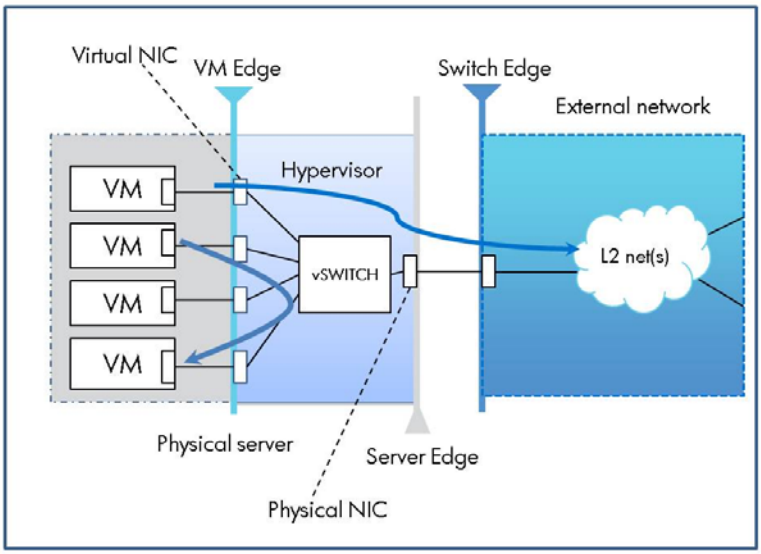
It is important to identify where the E/W traffic flows are occurring in your data center: at the physical switch-server edge between physical servers, or internal to the virtualized server (VM-to-VM). You can make tradeoffs, depending on two criteria:

- Whether you want to optimize the E/W traffic flow by providing intelligent management at the physical switch-server edge
- Whether you want to optimize for performance inside a physical server between multiple VMs (with possible degradation of network management visibility and control)

Virtual switch architectures

Today's hypervisors implement a virtualized network switch commonly known as a vSwitch. It is also known as a Virtual Ethernet Bridge (VEB). The vSwitch supports communication between VMs, the hypervisor, and external network switches. It provides efficient and low latency traffic flow between local VM-to-VM servers without the need to go to an external network switch (Figure 5).

Figure 5: vSwitches do a good job of efficiently routing internal VM-to-VM traffic.



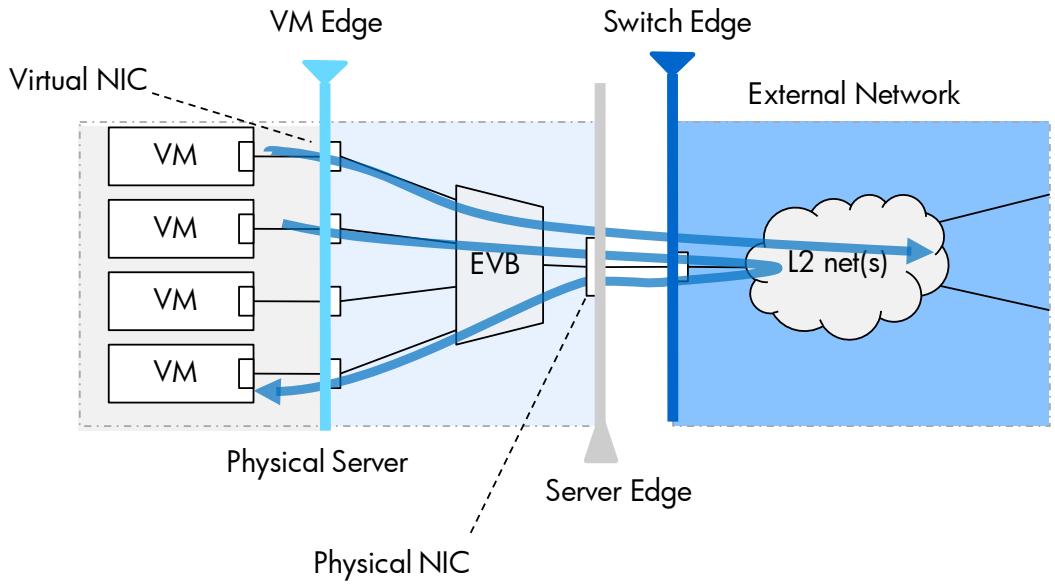
However, there are some limitations to a vSwitch:

- It moves the control point for networking infrastructure into the domain of the server administrator. This management stack is typically a component of the server-based hypervisor tool aimed at system and virtualization administrators. As such, vSwitch management generally does not integrate with existing external physical network policy and management tools. This usually means two different teams (with different processes) manage the physical network and the virtual network, even though the management tasks and functionality overlap.
- It consumes valuable CPU bandwidth. The higher the traffic load, the greater the number of CPU cycles required to move traffic through the vSwitch. This reduces the ability to support larger numbers of VMs in a physical server.
- It lacks network-based visibility. A vSwitch does not have standard network monitoring capabilities such as flow analysis, advanced statistics, and remote diagnostics of external network switches. When network outages or problems occur, identifying the root cause can be difficult in a virtualized server environment.

EVB architectures—VEPA and VEB

To solve some of these management drawbacks, HP is working with other vendors in the IEEE 802.1 Work Group to develop the Edge Virtual Bridging (EVB) standard. The EVB standard uses Virtual Ethernet Port Aggregator (VEPA) technology as its foundation. VEPA is a way for virtual switches to send all traffic and forwarding decisions to the adjacent physical switch (Figure 6). This removes the burden of VM forwarding decisions and network operations from the host CPU. And it leverages the existing management capabilities in the access-layer switches.

Figure 6: VEPA sends all traffic to the adjacent physical switch.



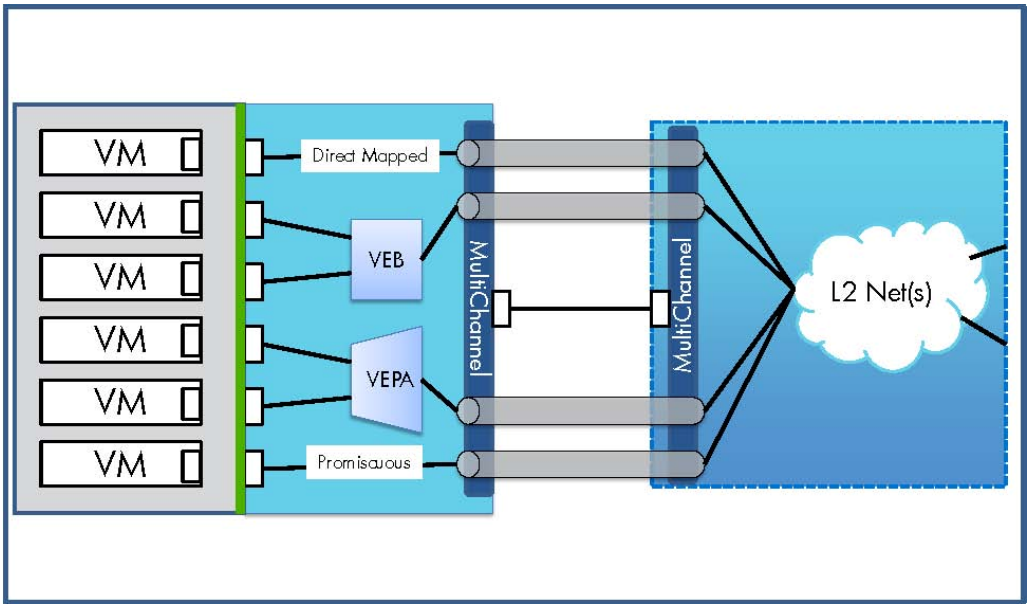
Advantages of VEPA include:

- Moves the VM connection control point into the edge physical switch (ToR or EoR). VEPA leverages existing investments made in data center edge switching. Administrators can manage the edge network traffic using existing network security policies and tools.
- Offloads the server's CPU from the overhead related to virtualization specific network processing and forwarding
- Improves security. Most ToR switches support hardware-based access control lists (TCAMs), allowing thousands of these filters to be processed without any effect on performance.
- Improves visibility. Monitoring technologies like sFlow in the edge switch can provide a full, end-to-end understanding of traffic flows.

If you plan to implement VEPA technology in the future, you can get the advantages of high-performance E/W traffic across physical servers and management visibility at the physical server-network edge. But it doesn't give you advantages with E/W traffic within a single virtualized server.

The EVB standard supports VEPA-based switches and existing vSwitch (VEB) architectures simultaneously (Figure 7). IT architects can choose whether to manage the edge traffic (frame processing, security features, networking monitoring, and so on) in the local hypervisor (vSwitch) or in the adjacent physical switch (VEPA-based switch).

Figure 7: Multichannel capabilities let vSwitch (VEB) and VEPA modes co-exist in the same server platform.



HP Virtual Connect

One of the ways to optimize the server edge for E/W traffic flow is by implementing HP Virtual Connect technology. Virtual Connect is a set of interconnect modules and embedded software for HP BladeSystem c- Class enclosures that provides server-edge and I/O virtualization. It delivers direct server-to-server connectivity within an enclosure—especially important for the latency sensitive, bandwidth-intensive applications that we've been discussing. For example, as described in the Client Virtualization section, using BladeSystem with Virtual Connect lets you design an infrastructure that can optimize network traffic without leaving the enclosure.

Using Flex-10 technology with FlexFabric adapters lets you consolidate multiple network connections—data and storage traffic—onto a single 10 Gb Ethernet pipe. This lets you reduce the number of physical adapters, cables, switches, and required ports at the server edge.

You can also use stacking links with the VC Ethernet modules to allow all server NICs in the Virtual Connect domain to have access to any VC-Ethernet uplink port. Using these module-to-module links, a single pair of uplinks serves as the network connections for the entire Virtual Connect domain. This reduces the core switch traffic, because internal communication stays inside the Virtual Connect domain. The stacking links provide high-speed connections between the enclosures that you can adjust by adding more physical links to the stack. You can increase the server-edge bandwidth by using the stacking links between Virtual Connect modules.

HP Intelligent Resilient Framework

HP Intelligent Resilient Framework (IRF) is an HP-developed switch virtualization technology that simplifies network architectures:

- Aggregates multiple physical devices to operate as a single logical device
- Virtualizes and distributes L2 and L3 functions
- Delivers high performance by using all available links

IRF lets you create large network fabrics with multiple switches at a single layer (access, distribution, or core) that operate and appear logically as a single switch. You can combine as many as eight HP networking stackable switches or up to four HP networking chassis switches to create a single, logical switch comprised of hundreds or even thousands of 1-GbE or 10-GbE switch ports. Because you now have a single switch, the routing protocols calculate routes based on a single logical domain rather than the multiple switches it represents.

IRF mimics the management behavior of a chassis switch by moving what was the management control plane in the backplane of a chassis switch to a distributed control plane across the network fabric. You only need one configuration file and one software image to manage IRF devices. Devices inserted into an IRF domain automatically update their configuration file and software, preventing you from modifying one device in the domain in isolation from the others. Connecting to the fabric through any port, console port or management port will link you to the single, redundant domain controller hosted on an elected IRF device in the domain.

Edge or aggregation switches (including Virtual Connect) that are dual-homed to IRF-enabled switches effectively "see" the associated upstream switches as a single entity, eliminating the need for slow convergence technologies such as STP. IRF-enabled switches also let you use simple link-aggregation protocols (such as 802.3AD LACP) for effective, failure tolerant active-active multi-chassis dual homing. This creates a distributed switch that is highly available, has no single point of failure, and requires no complex load balancing or failover protocols.

Unlike STP, which limits the active links to prevent loops, IRF delivers high performance by fully using all links between switches and servers. IRF also provides low latency communications, ensuring rapid network recovery (less than 50 milliseconds) from a link or device failure. This is much faster than the several seconds that an STP- or even an RSTP-network uses.

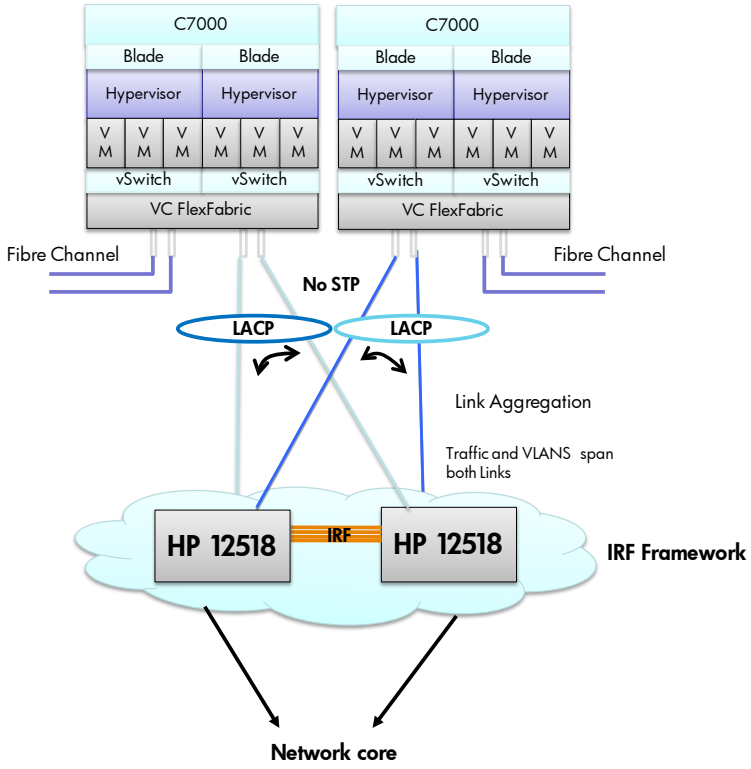
Because IRF is a switch virtualization architecture, it works with other higher layer protocols rather than competing with them. This includes protocols like Shortest Path Bridging (SPB), IETF Transparent Interconnection of Lots of Links (TRILL), and Multiprotocol label switching (MPLS). This gives you the most flexibility while maintaining resiliency and performance.

IRF runs on the high-end HP networking switches. The HP 12500, 9500, 7500, 58XX, and 55XX switches all come with HP IRF technology built in.

HP IRF with HP Virtual Connect technology

By combining IRF with Virtual Connect technology, you can take an existing three-tier network and flatten it to two tiers—the aggregation layer with IRF and the network core layer (Figure 8). You also get the advantages of reduced hardware at the server edge by using VC FlexFabric adapters and modules.

Figure 8: HP IRF technology can combine with Virtual Connect technology to flatten the L2 network.



This architecture lets you eliminate two levels of physical hardware switches by leveraging the vSwitch for the access layer, aggregating multiple vSwitches at the Virtual Connect interface, and going directly to the network core using IRF technology.

The architecture gives you:

- L2 flexibility optimized for VMs. A flattened L2 network lets you move VMs without the need for IP address changes. A flattened L2 design supports long-range vMotion using VPLS (Virtual Private LAN Service).
- Reduced cost of hardware components at the server edge. Using Virtual Connect FlexFabric lets you reduce the amount of cables, adapters, and interconnect modules by 95% compared to a blade infrastructure that uses pass-through modules.
- Reduced management complexity. Using large aggregation switches reduces the number of devices that you have to provision and monitor when setting up VLANs for VMs. Virtual Connect lets you manage network connections in a single management interface, especially if you use Virtual Connect Enterprise Manager. VCEM can manage up to 250 Virtual Connect domains comprising up to 16,000 blades.
- No need for STP or RSTP. IRF provides a loop-free network design, eliminating the need for STP along with its bandwidth and latency limitations.
- Optimized frame forwarding and packet forwarding performance. VMs communicate between each other, with their converged I/O resources served by an optimal L2 network, VLANs, and LACP. This virtually eliminates the need for TCP flow. Virtual Connect also supports standard LACP and lets you connect to HP networking upstream switches using IRF technology.

If you use the HP 12518 switch as shown in Figure 8, you can have up to 248 non-blocking 10 Gb/s ports (assuming 4 ports per switch for IRF and 4 ports per switch for uplinks). Each BladeSystem c7000 enclosure supports up to 16 servers. You only need two VC FlexFabric modules to connect all 16 servers to your LAN and SAN.

By deploying IRF in conjunction with high-performance HP server edge switches, enterprises can directly interconnect hundreds of VMs at the edge of the network, eliminating unnecessary network hops, reducing latency, and improving performance for large intra-data center workloads.

Conclusion

Virtualization, cloud computing, and federated applications bring flexibility, scalability, and cost advantages to your business. They also significantly alter how network traffic flows in your data center. Our position is to support multiple data center options rather than forcing you down a proprietary path that may limit other choices in your infrastructure. The typical hierarchical L2 network structure is limited to a single-path architecture with the Spanning Tree Protocol (STP). A multipath solution that lets traffic easily flow across multiple paths would improve performance in a L2 network with heavy E/W traffic.

Options to consider include using VEPA or VEB technology, flattening the L2 network, or making an L2 network more efficient by eliminating STP technologies.

Keep in mind that one size does not fit all, even in the same data center. Portions of your data center (for example, green field deployments of a cloud infrastructure) may require a high-performance, intelligent edge that supports lots of E/W traffic flow. Other parts of your data center may continue to operate with the traditional three-tier architecture.

For more information

Visit the URLs listed below if you need additional information.

Resource description	Web address
Cloud Computing for dummies, HP edition	www.ingrammicro.com/visitor/servicesdivision/cloudcomputingfordummies.pdf
"Reducing network complexity, boosting performance with HP IRF technology"	http://h20195.www2.hp.com/v2/GetPDF.aspx/4AA2-9402ENW.pdf
HP networking blog, "Simplifying data center architecture," 7/13/11	http://h30507.www3.hp.com/t5/HP-Networking/Simplifying-Data-Center-Network-Architectures/ba-p/95429
HP Virtual Desktop Infrastructure	http://h18000.www1.hp.com/products/servers/vdi/index.html

Send comments about this paper to TechCom@HP.com


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

TC1108853, October 2011