Data Center Networks: Virtual Bridging

Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

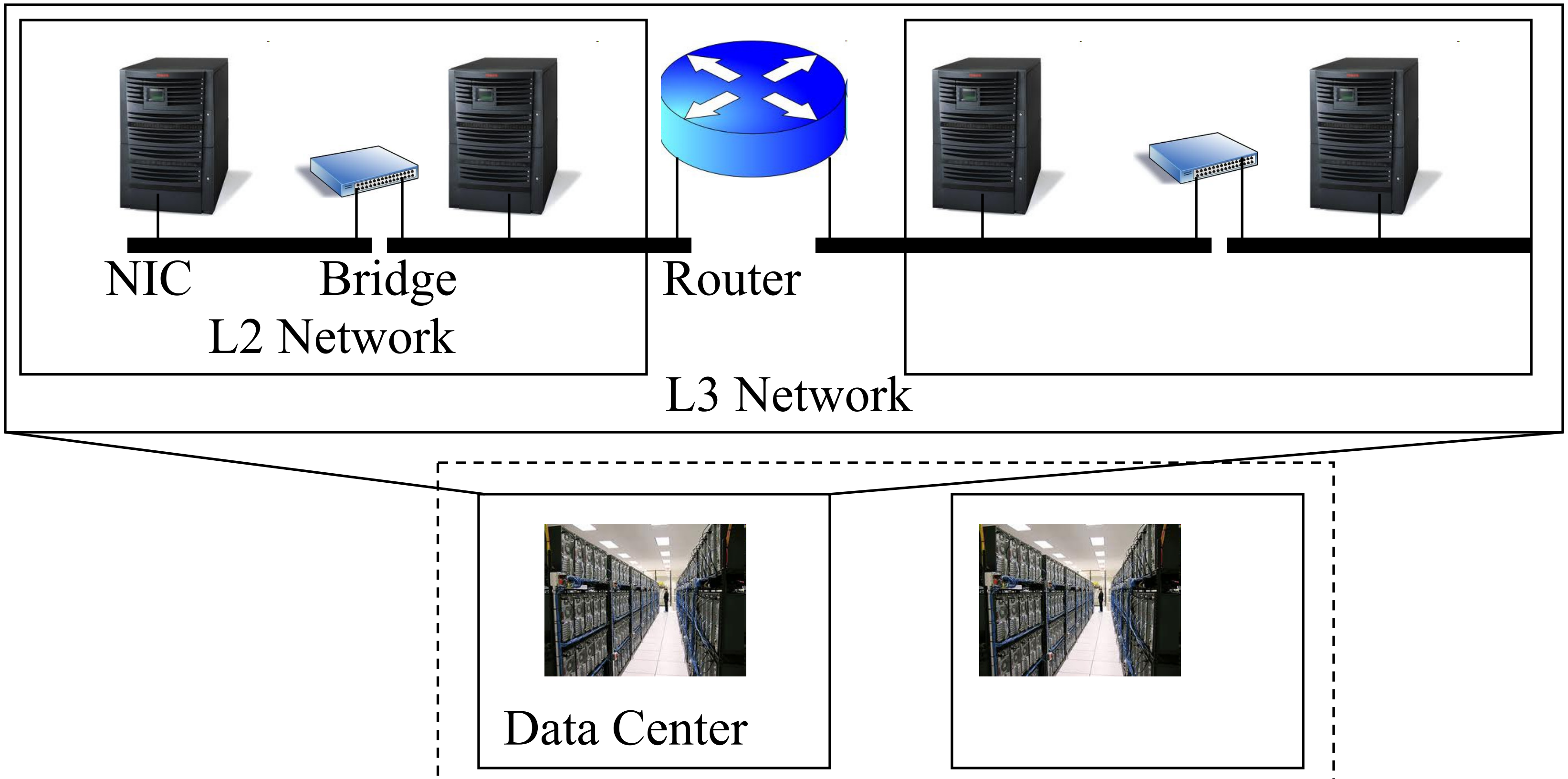
These slides and audio/video recordings of this class lecture are at:
http://www.cse.wustl.edu/~jain/cse570-15/

1. Virtual Bridges to connect virtual machines
2. IEEE Virtual Edge Bridging Standard
3. Single Root I/O Virtualization (SR-IOV)
4. Aggregating Bridges and Links: VSS and vPC
5. Bridges with massive number of ports: VBE

Network Virtualization

1. Network virtualization allows tenants to form an overlay network in a multi-tenant network such that tenant can control:
    1. Connectivity layer: Tenant network can be L2 while the provider is L3 and vice versa
    2. Addresses: MAC addresses and IP addresses
    3. Network Partitions: VLANs and Subnets
    4. Node Location: Move nodes freely
2. Network virtualization allows providers to serve a large number of tenants without worrying about:
    1. Internal addresses used in client networks
    2. Number of client nodes
    3. Location of individual client nodes
    4. Number and values of client partitions (VLANs and Subnets)
3. Network could be a single physical interface, a single physical machine, a data center, a metro, … or the global Internet.
4. Provider could be a system owner, an enterprise, a cloud provider, or a carrier.

Levels of Network Virtualization



- Networks consist of: **Host Interface** - L2 Links - **L2 Bridges** - **L2 Networks** - L3 Links - L3 Routers - L3 Networks – **Data Centers** – **Global Internet.**
- Each of these needs to be virtualized

Network Virtualization Techniques

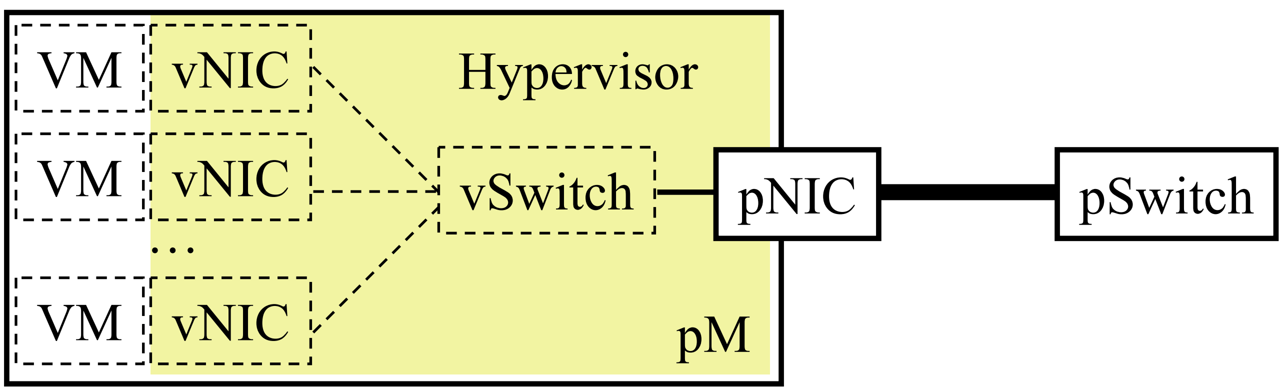
Entity	Partitioning	Aggregation/Extension/Interconnection**
NIC	SR-IOV	MR-IOV
Switch	VEB, VEPA	VSS, VBE, DVS, FEX
L2 Link	VLANs	LACP, Virtual PortChannels
L2 Network using L2	VLAN	PB (Q-in-Q), PBB (MAC-in-MAC), PBB-TE, Access-EPL, EVPL, EVP-Tree, EVPLAN
L2 Network using L3	NVO3, VXLAN, NVGRE, STT	MPLS, VPLS, A-VPLS, H-VPLS, PWoMPLS, PWoGRE, OTV, TRILL, LISP, L2TPv3, EVPN, PBB-EVPN
Router	VDCs, VRF	VRRP, HSRP
L3 Network using L1		GMPLS, SONET
L3 Network using L3*	MPLS, GRE, PW, IPSec	MPLS, T-MPLS, MPLS-TP, GRE, PW, IPSec
Application	ADCs	Load Balancers

*All L2/L3 technologies for L2 Network partitioning and aggregation can also be used for L3 network partitioning and aggregation, respectively, by simply putting L3 packets in L2 payloads.

**The aggregation technologies can also be seen as partitioning technologies from the provider point of view.

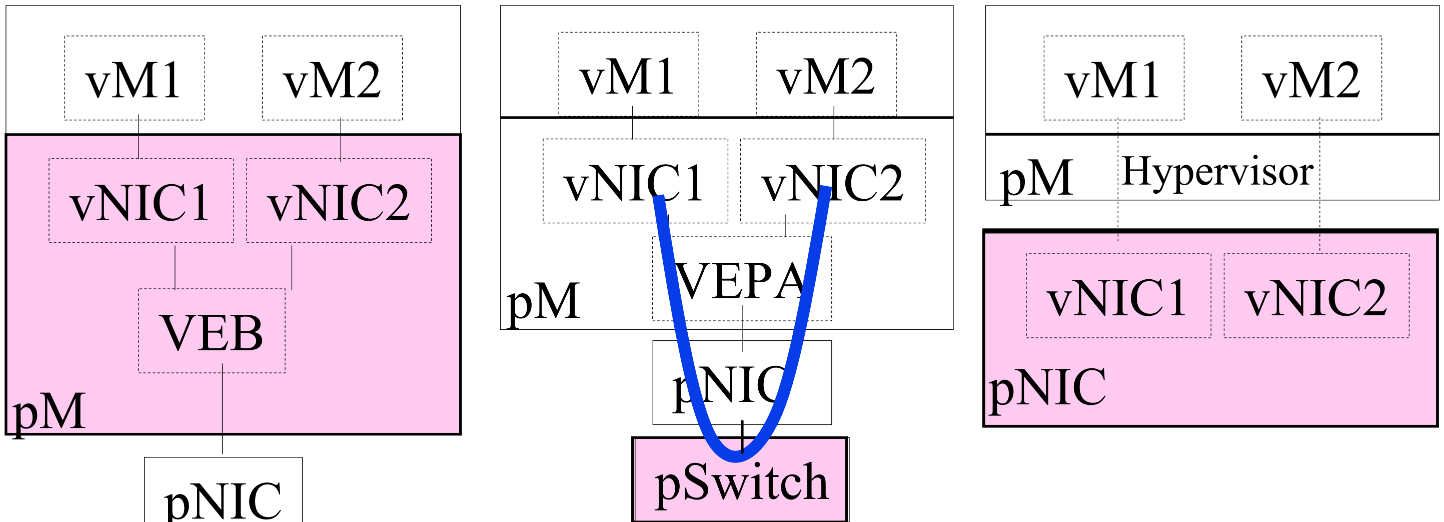
vSwitch

- **Problem**: Multiple VMs on a server need to use one physical network interface card (pNIC)
- **Solution**: Hypervisor creates multiple vNICs connected via a virtual switch (vSwitch)
- pNIC is controlled by hypervisor and not by any individual VM
- **Notation**: From now on prefixes p and v refer to physical and virtual, respectively. For VMs only, we use upper case V.



Ref: G. Santana, "Datacenter Virtualization Fundamentals," Cisco Press, 2014, ISBN: 1587143240

Virtual Bridging

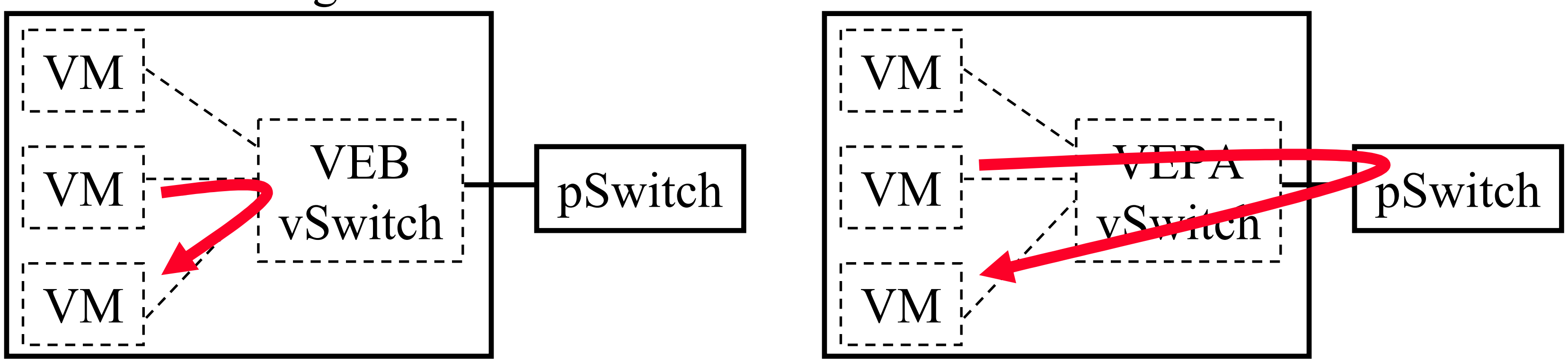


Where should most of the tenant isolation take place?

1. VM vendors: S/W NICs in Hypervisor w Virtual Edge Bridge (**VEB**)(overhead, not ext manageable, not all features)
2. Switch Vendors: Switch provides virtual channels for inter-VM Communications using virtual Ethernet port aggregator (**VEPA**): **802.1Qbg** (s/w upgrade)
3. NIC Vendors: NIC provides virtual ports using Single-Route I/O virtualization (**SR-IOV**) on PCI bus

Virtual Edge Bridge

- IEEE 802.1Qbg-2012 standard for vSwitch
- Two modes for vSwitches to handle *local* VM-to-VM traffic:
  - **Virtual Edge Bridge (VEB):** Switch internally.
  - **Virtual Ethernet Port Aggregator (VEPA):** Switch externally
- VEB
  - could be in a hypervisor or network interface card
  - may learn or may be configured with the MAC addresses
  - VEB may participate in spanning tree or may be configured\
  - Advantage: No need for the external switch in some cases



Virtual Ethernet Port Aggregator (VEPA)

- VEPA simply relays all traffic to an external bridge
- External bridge forwards the traffic. Called "*Hairpin Mode*."
  Returns local VM traffic back to VEPA
  Note: Legacy bridges do not allow traffic to be sent back to the incoming port within the same VLAN
- **VEPA Advantages**:
  - Visibility: External bridge can see VM to VM traffic.
  - Policy Enforcement: Better. E.g., firewall
  - Performance: Simpler vSwitch ⇒ Less load on CPU
  - Management: Easier
- Both VEB and VEPA can be implemented on the same NIC in the same server and can be cascaded.

Ref: HP, "Facts about the IEEE 802.1Qbg proposal," Feb 2011, 6pp.,
http://h20000.www2.hp.com/bc/docs/support/SupportManual/c02877995/c02877995.pdf

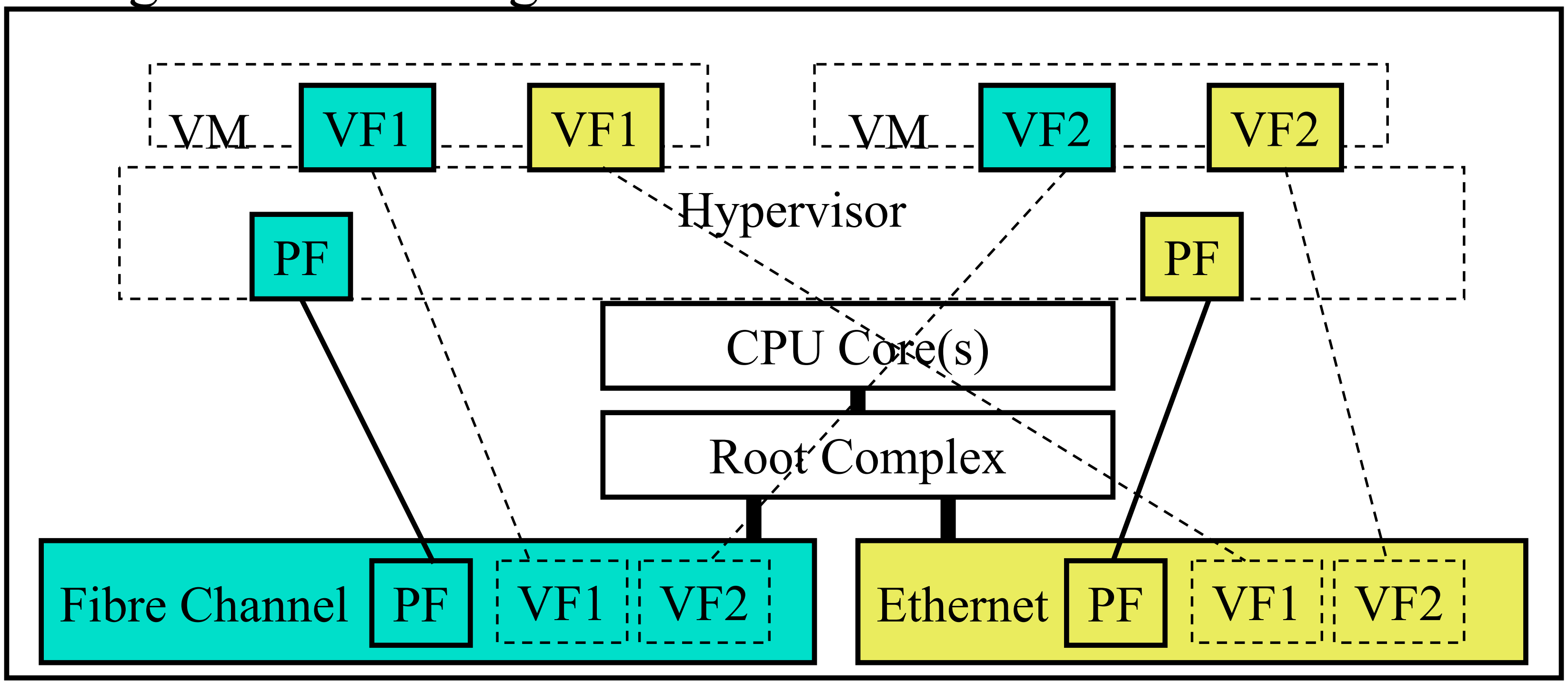
PCIe

- Peripheral Component Interconnect (PCI)
  Used in computers for I/O – storage, video, network cards
- Designed by PCI Special Interest Group (PCI-SIG)
- **PCI Express (PCIe)**: Serial point-to-point interconnect with multiple lanes, 4 pins per lane. X1=1 Lane, x32=32 lanes
  2 GB/s/lane.
- **Root complex** is the head of connection to CPU
- **Physical Function (PF)**: Ethernet, Fibre Channel, Video, …
- A PCIe card can provide multiple **virtual functions (VFs)** of the same type as PF, e.g., one 10Gbps pNIC = 2× 5Gbps vNICs

Ref: R. Emerick, "PCI Express IO Virtualization Overview," SNIA Education, 2012, http://www.snia.org/sites/default/files/RonEmerick_PCI_Express_IO_Virtualization.pdf (Excellent)

Single Root I/O Virtualization (SR-IOV)

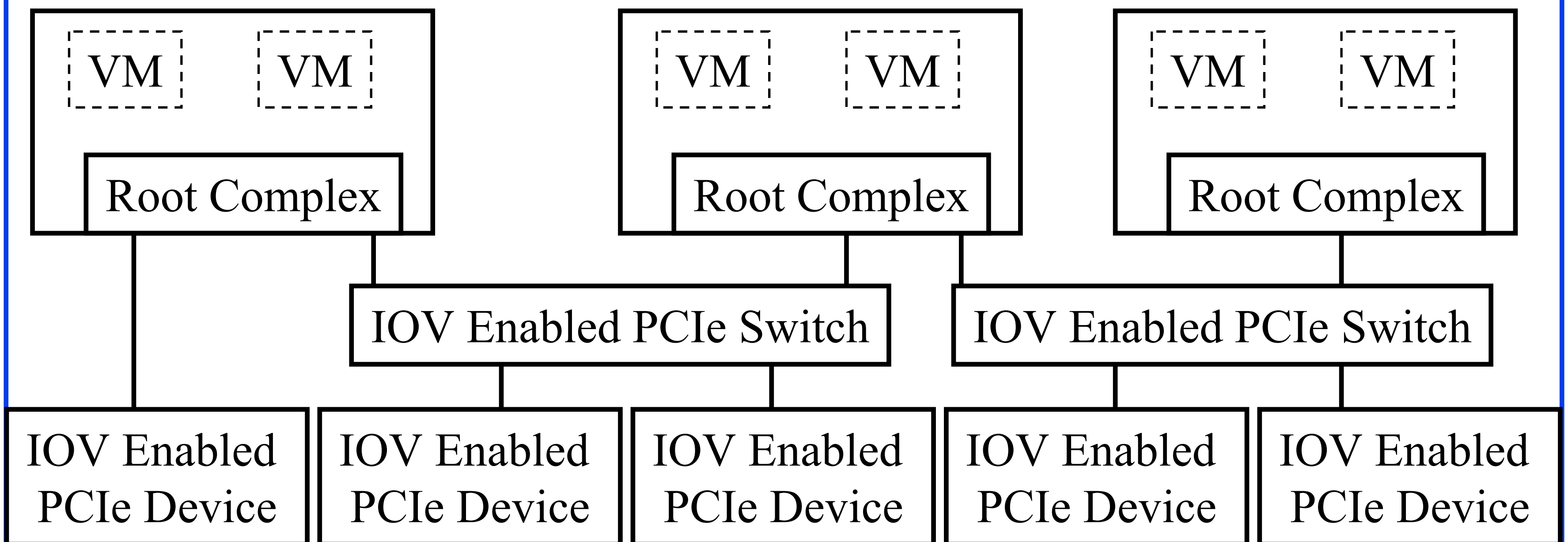
- After configuration by hypervisor, VFs allow direct VM access without hypervisor overhead
- Single Root ⇒ Single hardware domain ⇒ In one Server



Ref: Intel, "PCI-SIG SR-IOV Primer," Jan 2011, http://www.intel.com/content/dam/doc/application-note/pci-sig-sr-iov-primer-sr-iov-technology-paper.pdf

Multi-Root IOV

- Multiple external PCIe devices accessible via a switch
  - Move PCIe adapter out of the server into a switching fabric
  - Allows adapters to serve many physical servers
  - Used with rack mounted or blade servers
- Fewer adapters ⇒ Less cooling. No adapters ⇒ Thinner servers



Combining Bridges

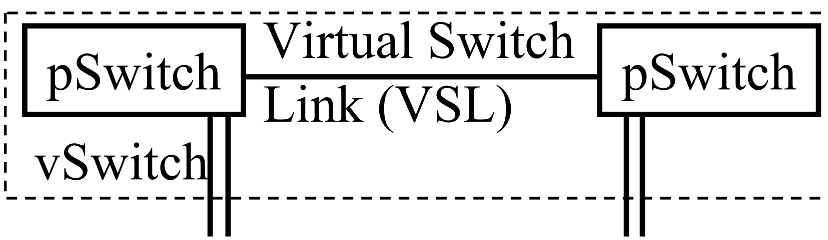
- **Problem**:
  - Number of VMs is growing very fast
  - Need switches with very large number of ports
  - Easy to manage one bridge than 100 10-port bridges
  - How to make very large switches ~1000 ports?
- **Solutions**: Multiple pswitches to form a single switch
  1. Distributed Virtual Switch (DVS)
  2. Virtual Switching System (VSS)
  3. Virtual PortChannels (vPC)
  4. Fabric Extension (FEX)
  5. Virtual Bridge Port Extension (VBE)

Distributed Virtual Switch (DVS)

- VMware idea to solve the scalability issue
- A centralized DVS controller manages vSwitches on many physical hosts
- DVS decouples the control and data plane of the switch so that each VM has a virtual data plane (virtual Ethernet module or VEM) managed by a centralized control plane (virtual Switch Module or VSM)
- Appears like a single distributed virtual switch
- Allows simultaneous creation of port groups on multiple pMs
- Provides an API so that other networking vendors can manage vSwitches and vNICs

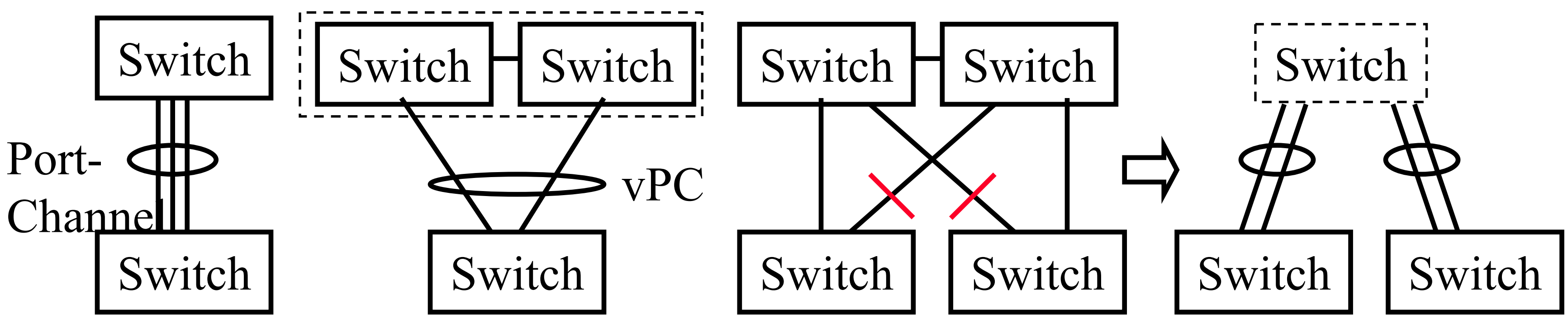
Virtual Switch System (VSS)

- Allows two physical switches to appear as one
- Although VSS is a Cisco proprietary name, several vendors implement similar technologies. E.g., Virtual Switch Bonding by Enterasys.
- Implemented in Firmware ⇒ No degradation in performance
- Only one control plane is active.
  Data-place capacity is doubled.
- Both switches are kept in sync to enable inter-chassis stateful switchover and non-stop forwarding in case of failure



Virtual PortChannel (vPC)

- **PortChannel**: Cisco name for aggregated link
- **Virtual PortChannel**: A link formed by aggregating links to multiple physical switches acting as a virtual switch
- The combined switch is called "**vPC Domain**"
- Each member of the vPC domain is called "**vPC peer**".
- vPC peer link is used to synchronize state and to forward traffic between the peers. No address learning on the peer link.
- All learned address tables are kept synchronized among peers. One peer learns an address ⇒ Sends it to every one else.

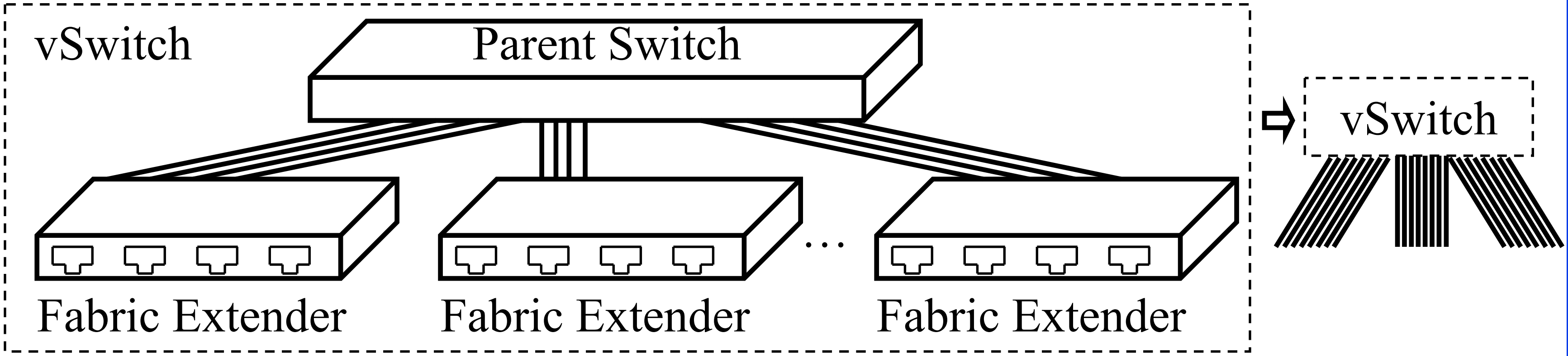


Virtual Port Channel (vPC)

- Allows aggregation of links going to different switches
  ⇒ STP does not block links ⇒ All capacity used
- Unlike VSS, maintains two independent control planes
- Independent control plane ⇒ In-service upgrade
  Software in one of the two switches can be upgraded without service interruption
- Falls back to STP ⇒ Used only in small domains
- vPC is Cisco proprietary. But other vendors have similar technologies. E.g., Split Multi-link Trunking (SMLT) by Nortel or "Multi-Chassis Link Aggregation (MC-LAG)" by Alcatel-Lucent. There is no standard.

Fabric Extenders

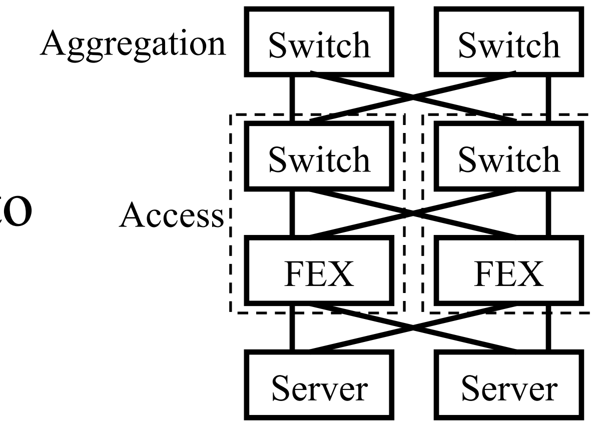
- Fabric extenders (FEX) consists of ports that are managed by a remote parent switch
- 12 Fabric extenders, each with 48 host ports, connected to a parent switch via 4-16 10 Gbps interfaces to a parent switch provide a virtual switch with 576 host ports
  $\Rightarrow$ **Chassis Virtualization**
- All software updates/management, forwarding/control plane is managed centrally by the parent switch.
- A FEX can have an active and a standby parent.

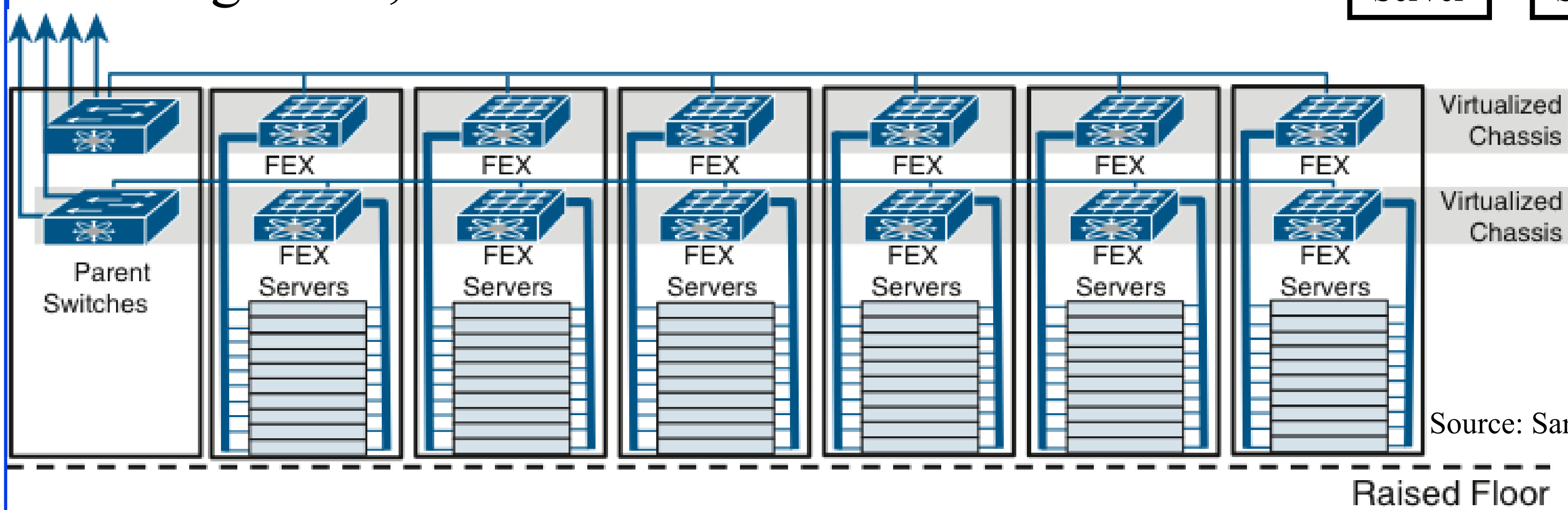


Ref: P. Beck, et al., "IBM and Cisco: Together for a World Class Data Center," IBM Red Book, 2013, 654 pp., ISBN: 0-7384-3842-1, http://www.redbooks.ibm.com/redbooks/pdfs/sg248105.pdf

FEX Topology Example

- All hosts are dual homed to FEX
  ⇒ Two FEX per rack
- Both FEX are dual homed to two parents
  ⇒ Two virtual access switches
- Virtual Access switches are dual homed to aggregation switches.
- Using vPCs, all links can be active.

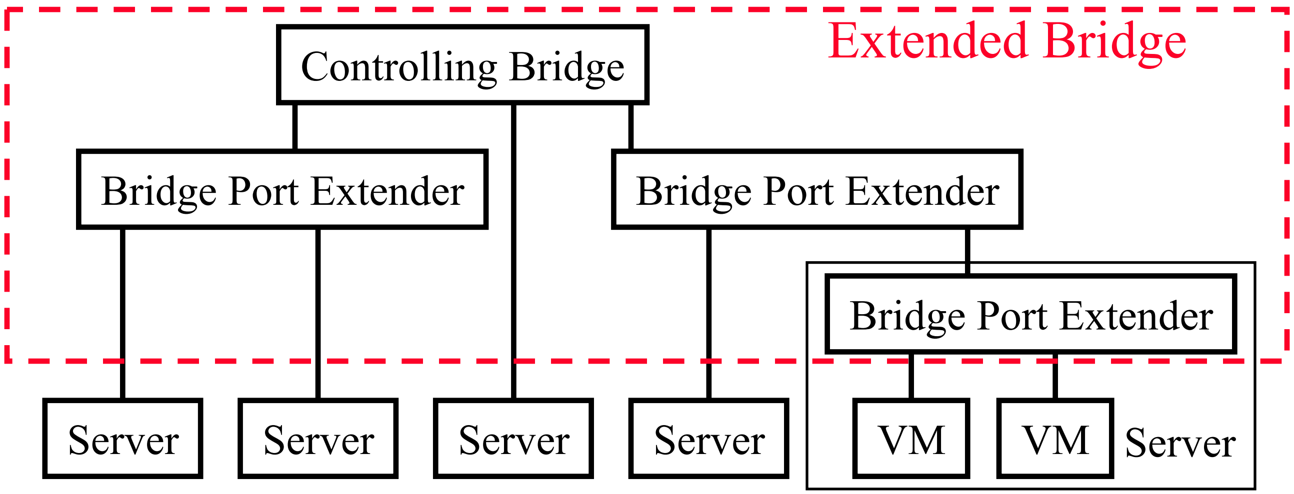


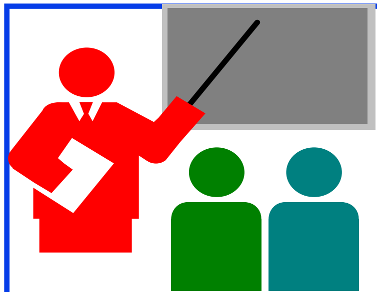


Source: Santana 2014

Virtual Bridge Port Extension (VBE)

- IEEE 802.1BR-2012 standard for fabric extender functions
- Specifies how to form an extended bridge consisting of a controlling bridge and Bridge Port Extenders
- Extenders can be cascaded.
- Some extenders may be in a vSwitch in a server hypervisor.
- All traffic is relayed by the controlling bridge
  $\Rightarrow$ Extended bridge is a bridge.





Summary

1. Network virtualization includes virtualization of NICs, Bridges, Routers, and L2 networks.
2. Virtual Edge Bridge (VEB) vSwitches switch internally while Virtual Ethernet Port Aggregator (VEPA) vSwitches switch externally.
3. SR-IOV technology allows multiple virtual NICs via PCI and avoids the need for internal vSwitch.
4. VSS allows multiple switches to appear as one logical switch vPortChannels allow links to multiple switches appear as one.
5. Fabric Extension and Virtual Bridge Extension (VBE) allows creating switches with a large number of ports using port extenders (which may be vSwitches)

Reading List

- HP, "Facts about the IEEE 802.1Qbg proposal," Feb 2011, 6pp., http://h20000.www2.hp.com/bc/docs/support/SupportManual/c02877995/c02877995.pdf
- Juniper, "Standardizing Data Center Server-Network Edge Virtualization," Oct 2010, http://www.juniper.net/us/en/local/pdf/whitepapers/standardizing-datacenter-server-network.pdf
- G. Santana, "Datacenter Virtualization Fundamentals," Cisco Press, 2014, ISBN: 1587143240 (Safari Book)
- P. Thaler, et al., "IEEE 802 Tutorial: Edge Virtual Bridging," Nov 2009, 54 slides, http://www.docstoc.com/docs/88675018/Edge-Virtual-Bridging
- H. Shah, "Management Standards for Edge Virtual Bridging (EVB) and Network Port Profiles," Nov 2010, http://www.ieee802.org/1/files/public/docs2011/bg-shah-dmtf-evbportprofile-overview-0311.pdf

Reading List (Cont)

- Intel, "PCI-SIG SR-IOV Primer," Jan 2011, http://www.intel.com/content/dam/doc/application-note/pci-sig-sr-iov-primer-sr-iov-technology-paper.pdf
- P. Beck, et al., "IBM and Cisco: Together for a World Class Data Center," IBM Red Book, 2013, 654 pp., ISBN: 0-7384-3842-1, http://www.redbooks.ibm.com/redbooks/pdfs/sg248105.pdf
- R. Emerick, "PCI Express IO Virtualization Overview," SNIA Education, 2012, http://www.snia.org/sites/default/files/RonEmerick_PCI_Express_IO_Virtualization.pdf (Excellent)
- R. Sharma, et al., "VSI Discovery and Configuration," Jan 2010, http://www.ieee802.org/1/files/public/docs2010/bg-sharma-evb-VSI-discovery-0110-v01.pdf

Wikipedia Links

- http://en.wikipedia.org/wiki/Address_Resolution_Protocol
- http://en.wikipedia.org/wiki/EtherChannel
- http://en.wikipedia.org/wiki/IEEE_802.1aq
- http://en.wikipedia.org/wiki/Link_aggregation
- http://en.wikipedia.org/wiki/MC-LAG
- http://en.wikipedia.org/wiki/Network_virtualization
- http://en.wikipedia.org/wiki/PCI_Express
- http://en.wikipedia.org/wiki/Port_Aggregation_Protocol
- http://en.wikipedia.org/wiki/Reverse_Address_Resolution_Protocol
- http://en.wikipedia.org/wiki/Root_complex
- http://en.wikipedia.org/wiki/Virtual_Routing_and_Forwarding

Acronyms

- A-VPLS Advanced Virtual Private LAN Service
- Access-EPL Access Ethernet Private Line
- Access-EVPL Access Ethernet Virtual Private Line
- ADC Application Delivery Controllers
- API Application Programming Interface
- ARP Address Resolution Protocol
- BPE Bridge Port Extension
- CDCP S-Channel Discovery and Configuration Protocol
- CPU Central Processing Unit
- DMTF Distributed Management Task Force
- DVS Distributed Virtual Switching
- ECP Edge Control Protocol
- EDCP Edge Discovery and Configuration Protocol
- EPL Ethernet Private Line
- EVB Edge Virtual Bridging
- EVP-Tree Ethernet Virtual Private Tree

Acronyms (Cont)

- EVPL — Ethernet Virtual Private Line
- EVPLAN — Ethernet Virtual Private Local Area Network
- EVPN — Ethernet Virtual Private Network
- FEX — Fabric Extender
- GB — Giga Byte
- GMPLS — Generalized Multi-Protocol Label Switching
- GRE — Generic Routing Encapsulation
- H-VPLS — Hierarchical Virtual Private LAN Service
- HSRP — Hot Standby Router Protocol
- IO — Input/Output
- IOV — Input/Output Virtualization
- IP — Internet Protocol
- IPoMPLSoE — IP over MPLS over Ethernet
- IPSec — Internet Protocol Security
- L2TPv3 — Layer 2 Tunneling Protocol Version 3
- LAG — Link Aggregation

Acronyms (Cont)

LISP	Locator ID Split Protocol
MAC	Media Access Control
MPLS-TP	Multiprotocol Label Switching Transport
MPLS	Multi-Protocol Label Switching
MR-IOV	Multi-Root I/O Virtualization
NIC	Network Interface Card
NVGRE	Network Virtualization using GRE
NVO3	Network Virtualization Over L3
OTV	Overlay Transport Virtualization
OVF	Open Virtual Disk Format
PB	Provider Bridge
PBB-EVPN	Provider Backbone Bridging with Ethernet VPN
PBB-TE	Provider Backbone Bridge with Traffic Engineering
PBB	Provider Backbone Bridge
PCI-SIG	Peripheral Component Interconnect Special Interest Group
PCI	Peripheral Component Interconnect

Acronyms (Cont)

- PCIe Peripheral Component Interconnect Express
- PF Physical Function
- pM Physical Machine
- pNIC Physical Network Interface Card
- pSwitch Physical Switch
- PW Pseudo Wire
- PWoGRE Pseudo Wire Over Generic Routing Encapsulation
- PWoMPLS Pseudo Wire over Multi-Protocol Label Switching
- SMLT Split Multi-link Trunking
- SNIA Storage Networking Industry Association
- SR-IOV Single Root I/O Virtualization
- STP Spanning Tree Protocol
- STT Stateless Transport Tunneling
- TP Transport Profile
- T-MPLS Transport Multiprotocol Label Switching
- TRILL Transparent Interconnection of Lots of Link

Acronyms (Cont)

- VBE Virtual Bridge Extension
- VDC Virtual Device Context
- VDP VSI Discovery and Configuration Protocol
- VEB Virtual Edge Bridge
- VEM Virtual Ethernet Module
- VEPA Virtual Ethernet Port Aggregator
- VF Virtual Function
- VIP Virtual IP
- VLAN Virtual Local Area Network
- VM Virtual Machine
- vNIC Virtual Network Interface Card
- vPC Virtual PathChannel
- VPLS Virtual Private LAN Service
- VPN Virtual Private Network
- vPort Virtual Port
- VRF Virtual Routing and Forwarding

Acronyms (Cont)

- VRRP Virtual Routing Redundancy Protocol
- VSI Virtual Station Interface
- VSL Virtual Switch Link
- VSS Virtual Switch System
- VXLAN Virtual eXtensible Local Area Network