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/cse570-18/
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

- Each of these needs to be virtualized
## Network Virtualization Techniques

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

*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.

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.
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

Washington University in St. Louis http://www.cse.wustl.edu/~jain/cse570-18/ ©2018 Raj Jain
Single Root I/O Virtualization (SR-IOV)

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


Washington University in St. Louis http://www.cse.wustl.edu/~jain/cse570-18/ ©2018 Raj Jain
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-plane 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 → **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.


Washington University in St. Louis http://www.cse.wustl.edu/~jain/cse570-18/ ©2018 Raj Jain
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.
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
  ⇒ Extended bridge is a bridge.

Controlling Bridge

Bridge Port Extender

Server
Server
Server

VM
VM
Server

Extended 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

Reading List (Cont)


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

CSE567M: Computer Systems Analysis (Spring 2013),
https://www.youtube.com/playlist?list=PLjGG94etKypJEKjNAa1n_1X0bWWNyZcof

CSE473S: Introduction to Computer Networks (Fall 2011),
https://www.youtube.com/playlist?list=PLjGG94etKypJWOSPMh8Azcg5y5e_10TiDw

Wireless and Mobile Networking (Spring 2016),
https://www.youtube.com/playlist?list=PLjGG94etKypKeb0nzyN9tSs_HCd5c4wXF

CSE571S: Network Security (Fall 2011),
https://www.youtube.com/playlist?list=PLjGG94etKypKvzfVtutHcPFJXumyyg93u

Video Podcasts of Prof. Raj Jain's Lectures,
https://www.youtube.com/user/ProfRajJain/playlists