Virtual Private Networks in Data Centers

Overview

- Traditional vs. Data Center Ethernet
- Review of LANs and Virtual LANs
- Spanning Tree Protocol
- Data Center Bridging and LAN Extension
- Network Virtualization in Multi-tenant Data Centers

Traditional vs. Data Center Ethernet

Office	**Data Center**
Distance: up to 200 m	No limit
Scale: Few MAC addresses 4096 VLANs	Millions of MAC addresses Millions of VLANs
Protection: Spanning Tree	Rapid Spanning Tree not enough
Path defined by spanning tree	Deactivation of multiple links is wasteful

Names, IDs, Locators

Name: Alice Silva

ID: 123 456 78 (Identity Card Number)

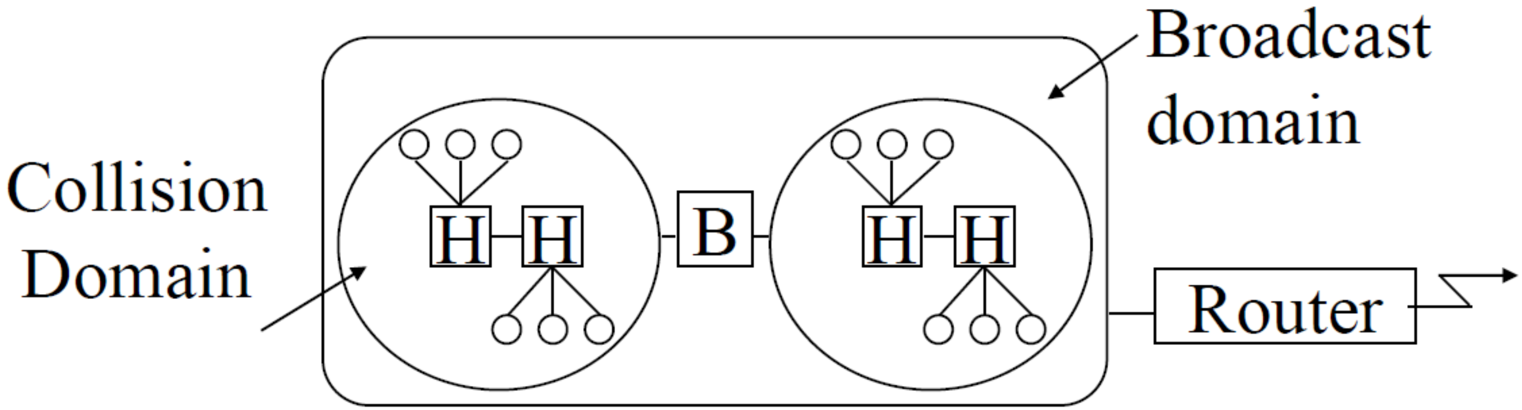
Locator:

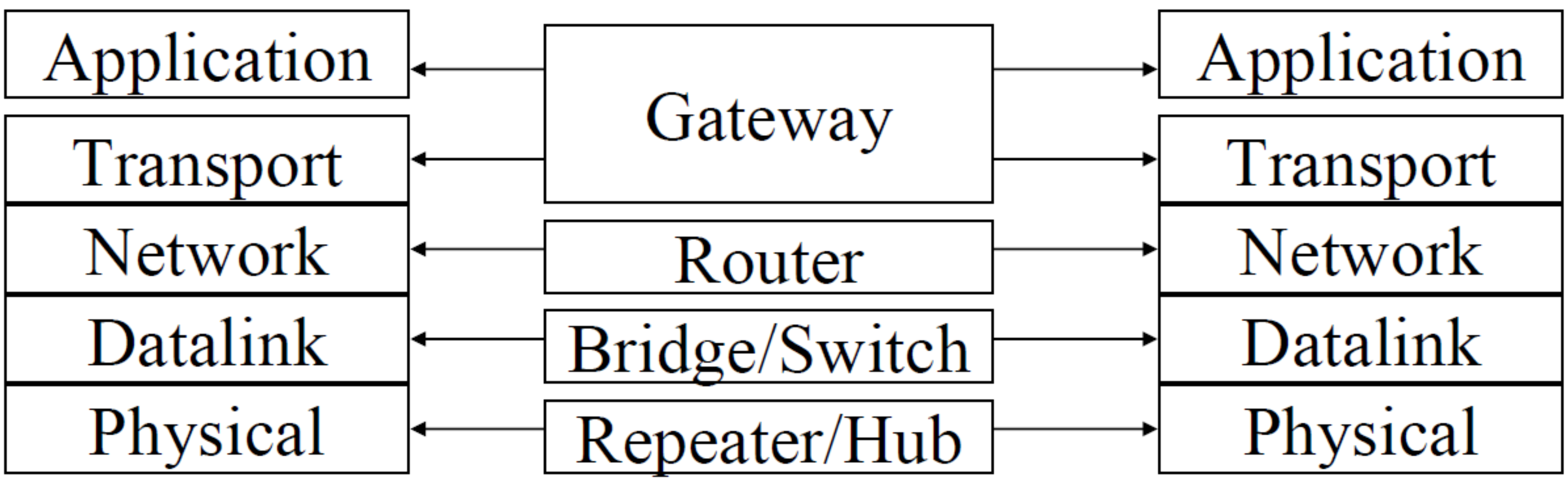
Rua Dr. Roberto Frias

4200-465 Porto, Portugal

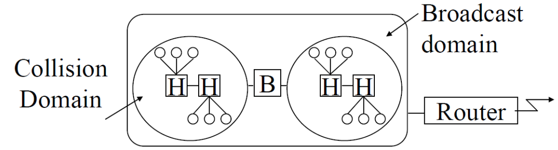
- Locator changes as you move, ID and Names remain the same
- Examples:
  - **Names**: Company names, DNS names (fe.up.pt)
  - **IDs**: Cell phone numbers, <u>Ethernet addresses</u>, Skype ID
  - **Locators**: Wired phone numbers, IP addresses

Interconnection Devices





Interconnection Devices



Application	Gateway	Application
Transport		Transport
Network	Router	Network
Datalink	Bridge/Switch	Datalink
Physical	Repeater/Hub	Physical

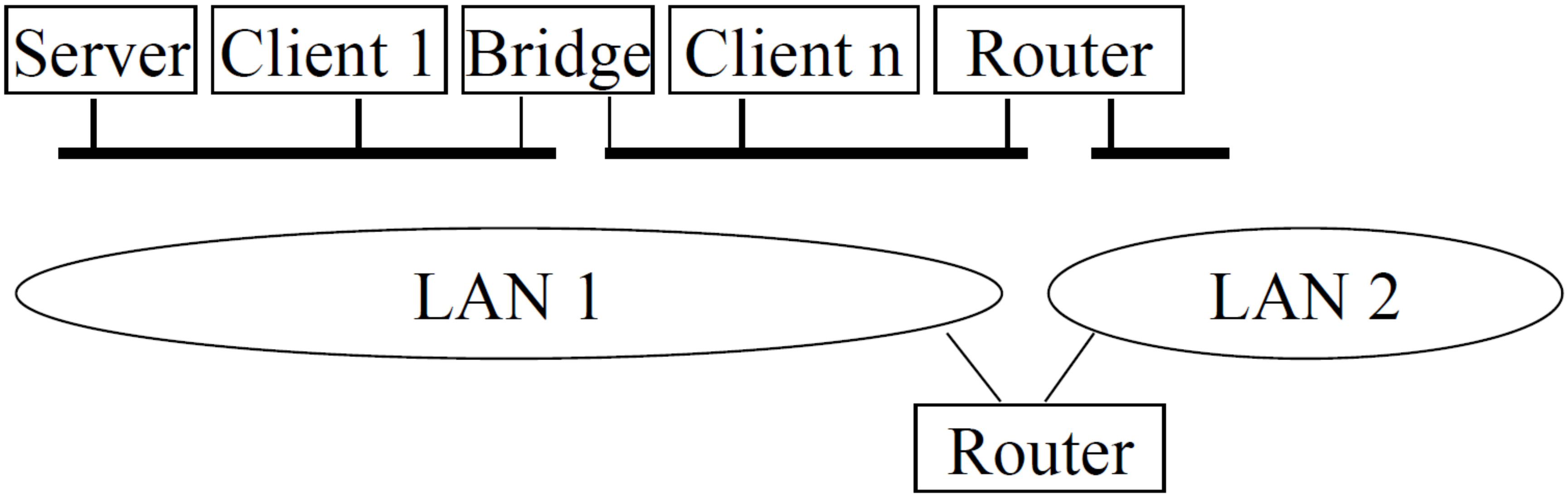
- **Repeater**: PHY device that restores data and collision signals
- **Hub**: Multiport repeater
- **Bridge**: Datalink layer device connecting two or more collision domains. MAC multicasts/broadcasts propagated throughout LAN
- **Router**: Network layer device (IP, IPX, AppleTalk) → isolates broadcast domains
- **Switch**: Multiport bridge with parallel paths

These are the functions → packaging varies

Ethernet Speeds

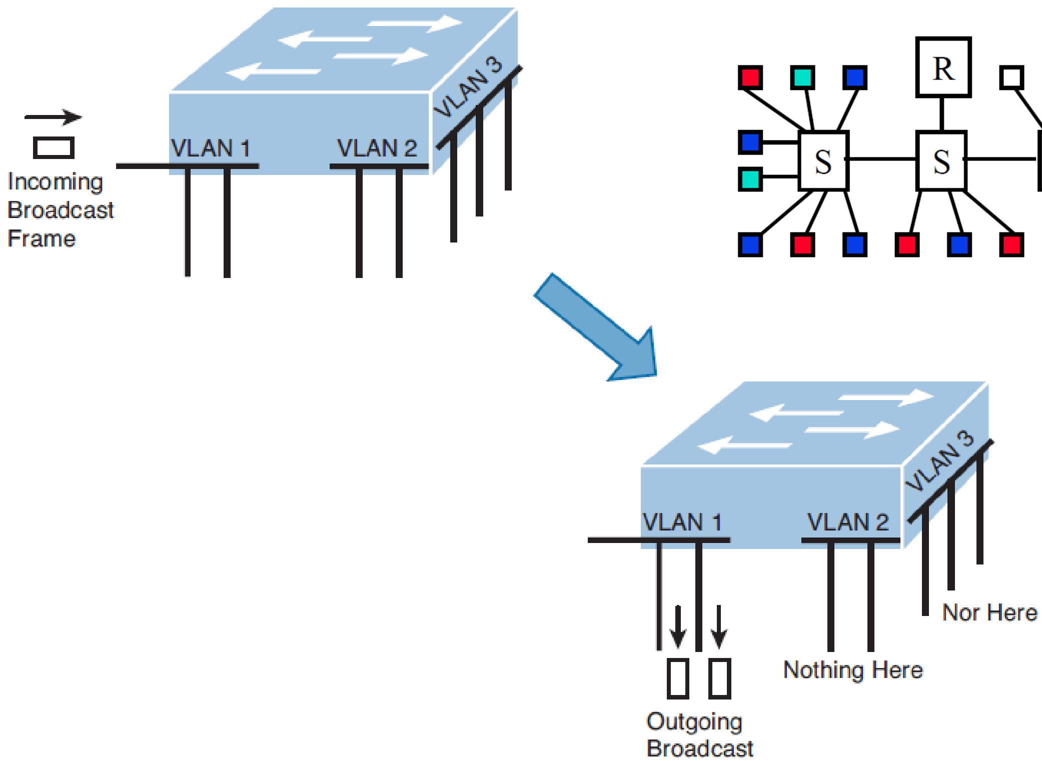
- IEEE 802.3ba-2010 (40G/100G) standard
- 10Mbps, 100 Mbps, 1 Gbps versions have both CSMA/CD and Full-duplex versions
- No CSMA/CD in 10G and up
- No CSMA/CD in practice now even at home or at 10 Mbps
- 1 Gbps in residential, enterprise offices
- 1 Gbps in Data centers, moving to 10 Gbps and 40 Gbps
- 100 Gbps in some carrier core networks

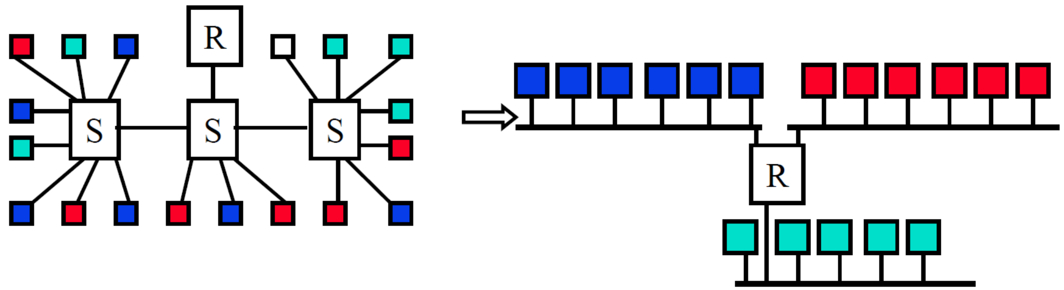
Local Area Network (LAN)



- LAN = Single broadcast domain = Subnet
- No routing between members of a LAN
- Routing required between LANs

Virtual Local Area Network (VLAN)

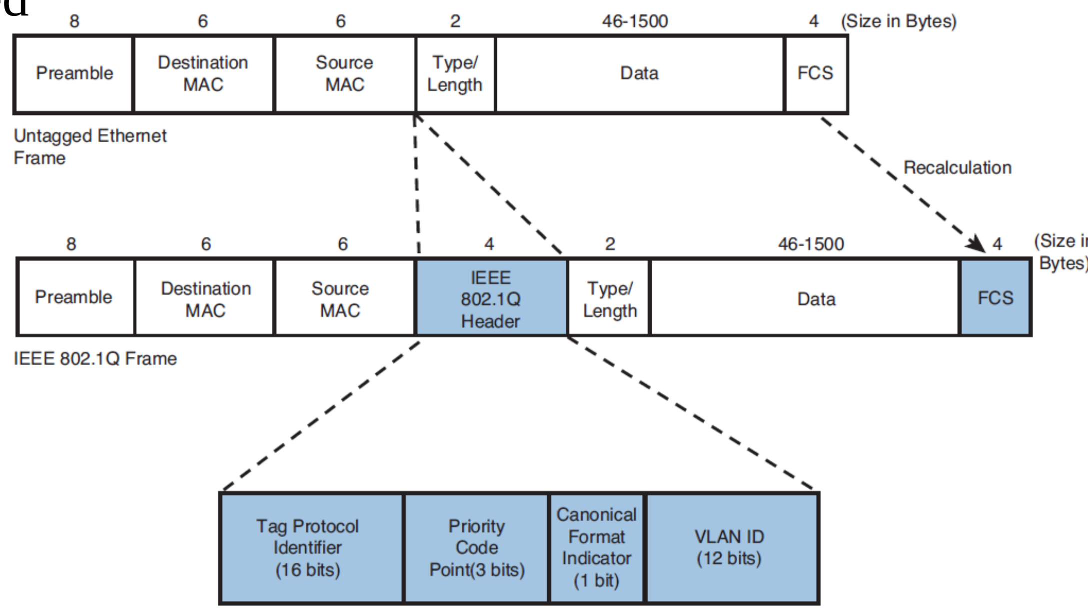




- Virtual LAN = Broadcasts and multicast goes only to the nodes in the virtual LAN
- LAN membership defined by the network manager → <u>Virtual</u>

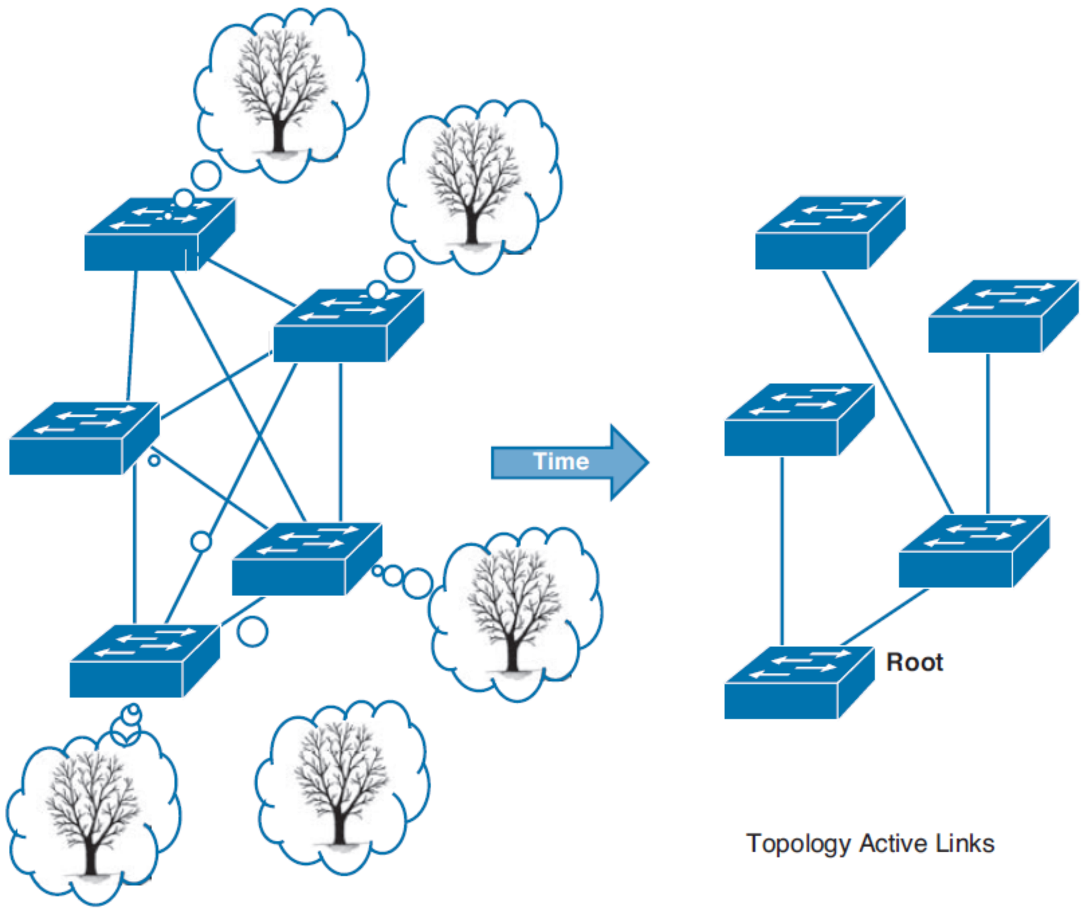
IEEE 802.1Q-2011 Tag

- Tag Protocol Identifier (TPI) → used to distinguish from untagged frames
- Priority Code Point (PCP): 3 bits = 8 priorities 0..7 (High)
- CFI: 0 → Standard Ethernet
- VLAN ID → 4094 VLANs (0 and 4095 reserved)
- Switches forward based on MAC address + VLAN ID
  - Unknown addresses → flooded



Spanning Tree Protocol

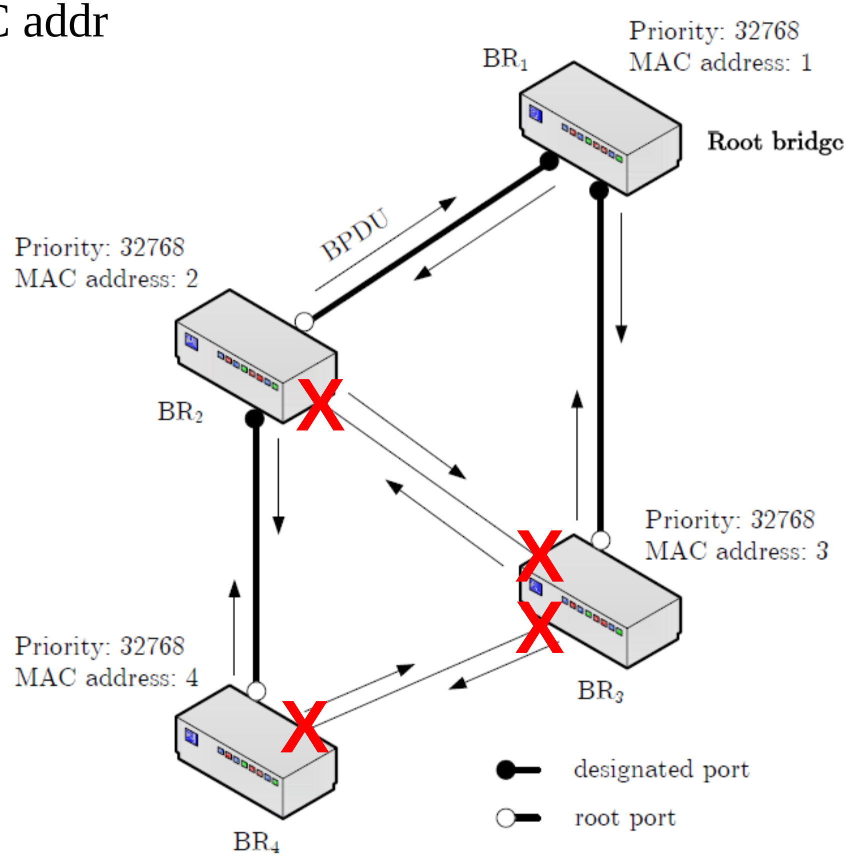
Helps form a tree out of a mesh topology



Source: G. Santana, “Data Center Virtualization Fundamentals”, Cisco Press, 2014, ISBN:1587143240

Spanning Tree Protocol – How it works?

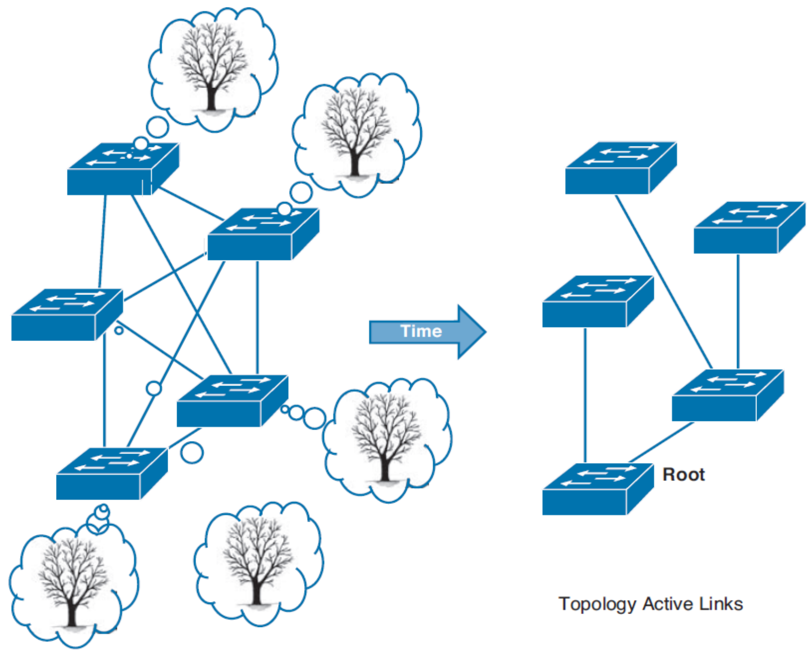
- All bridges multicast to “All bridges”
  - My ID → 64-bit ID = 16-bit priority + MAC addr
  - Root ID
  - Cost to root (inversely proportional to link bandwidth)
- Initially, all bridges are roots but eventually converge to one root as they find out <u>lowest Bridge ID</u>
- Bridges update their info using Dijkstra’s algorithm and rebroadcast BPDUs
- On each LAN segment, the <u>bridge with minimum cost to the root becomes the Designated bridge</u>
- All ports of all non-designated bridges are blocked



<u>Root port</u> → local port enabling lowest cost to elected root bridge

Spanning Tree Protocol Limitations

- Topology change can result in 1 min of traffic loss with STP
  - <u>All TCP connections break</u>
  - Rapid Spanning Tree Protocol (RSTP) → speed up convergence time
- Still, one tree for all VLANs (common spanning tree)



Source: G. Santana, “Data Center Virtualization Fundamentals”, Cisco Press, 2014, ISBN:1587143240

Data Center Bridging – Why?

- Enable data center traffic over Ethernet
- Many applications built with that assumption in mind
- <u>Ethernet’s use of IDs as addresses</u> makes it very easy to move systems in the data center
  - Keep traffic on the same Ethernet LAN
- VLANs allow traffic segregation from different tenants over the same physical network

But spanning tree is wasteful of resources and slow

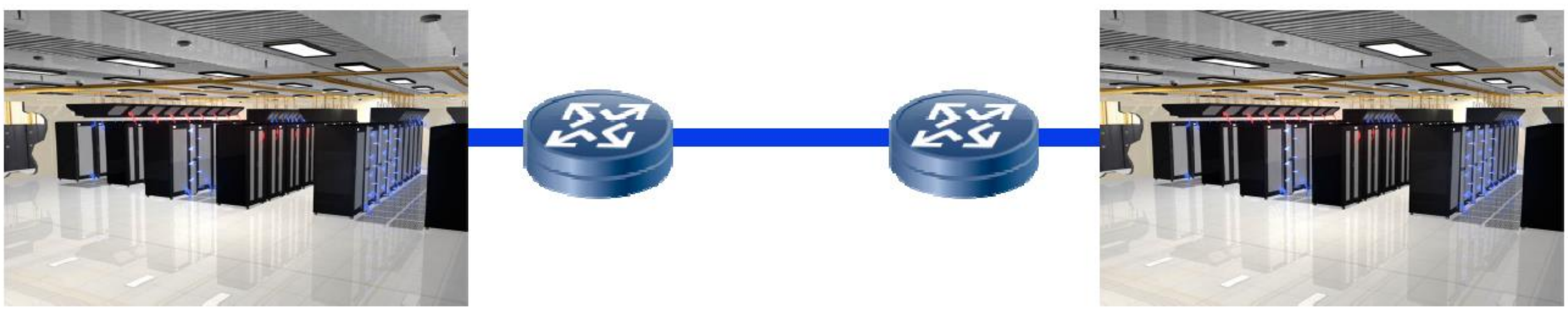
New solutions needed …

Geographic Clusters of Data Centers

- Multiple data centers are used to improve availability
- **Cold-Standby**: Data backed up on tapes and stored off-site. In case of disaster, application and data loaded in standby. Manual switchover
  - Significant downtime (1970-1990)
- **Hot-Standby**: Two servers in different geographically close data centers exchange state and data continuously. Upon failure, the application automatically switches to standby. Automatic switchover
  - Reduced downtime (1990-2005)
  - Only 50% of resources are used under normal operation
- **Active-Active:** All resources are used. Virtual machines and data can be quickly moved between sites, when needed.

Data Center Interconnection (DCI)

- Allows distant data centers to be connected in one L2 domain
  - Distributed applications
  - Disaster recovery
  - Maintenance/Migration
  - High-Availability
- Active and standby can share the same virtual IP for switchover
- Multicast can be used to send state to multiple destinations



Challenges of LAN Extension

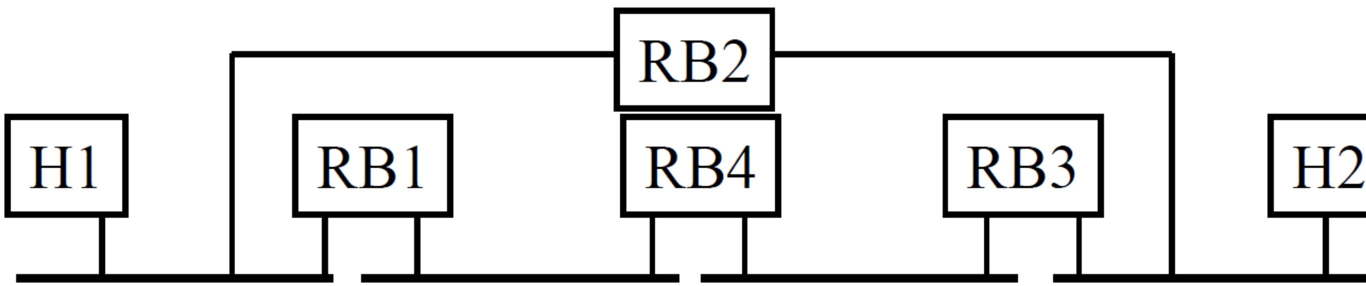
- **Broadcast storms**: Unknown and broadcast frames may create excessive flood
- **Loops**: Easy to form loops in a large network
- **STP Issues**: High spanning tree diameter (leaf-to-leaf): More than 7 (limit imposed by STP)
- Root can become bottleneck and a single point of failure
  - When STP instance is extended to multiple sites, only one of site will contain the root switch
  - If root fails → all VLANs within that instance will be affected
- Multiple paths remain unused
  - multiple DCI links between sites will not be used
- **Security**: Data on LAN extension must be encrypted

Enhancements to Spanning Tree Protocol

- MSTP (Multiple Spanning Tree)
  - Each tree serves a group of VLANs
  - Bridge port could be in forwarding state for some VLANs and blocked state for others
- Shortest Path Bridging (IEEE 802.1Q-2018)
  - Allows all links to be used → Better CapEx
  - Uses MAC-in-MAC encapsulation and IP routing
  - IS-IS link state protocol (similar to OSPF) is used to build shortest path trees for each node to every other node within the SPB domain
  - Equal-cost multi-path (ECMP) used to distribute load
    - Allowed by other major routing protocols such as OSPF and BGP

Enhancements to Spanning Tree Protocol

- TRILL
    - Transparent Interconnection of Lots of Links
    - Allows a large campus to be a single extended LAN
    - Use MAC addresses and IP routing
    - Zero Configuration: RBridges discover their connectivity and learn MAC addresses automatically
    - VLANs supported
    - Legacy bridges with spanning tree in the same extended LAN
    - Packets encapsulated and routed using IS-IS routing



Source: R. Perlman et al., “Routing Bridges (RBridges): Base Protocol Specification”, IETF RFC 6325, Jul. 2011.

Problem

- Need to support thousands of tenants and several thousand tenant networks
- 4096 (or 4K) network limitation imposed by the 12-bit VLAN field is not sufficient for supporting large multitenant data centers

Network Virtualization in Multi-tenant Data Centers

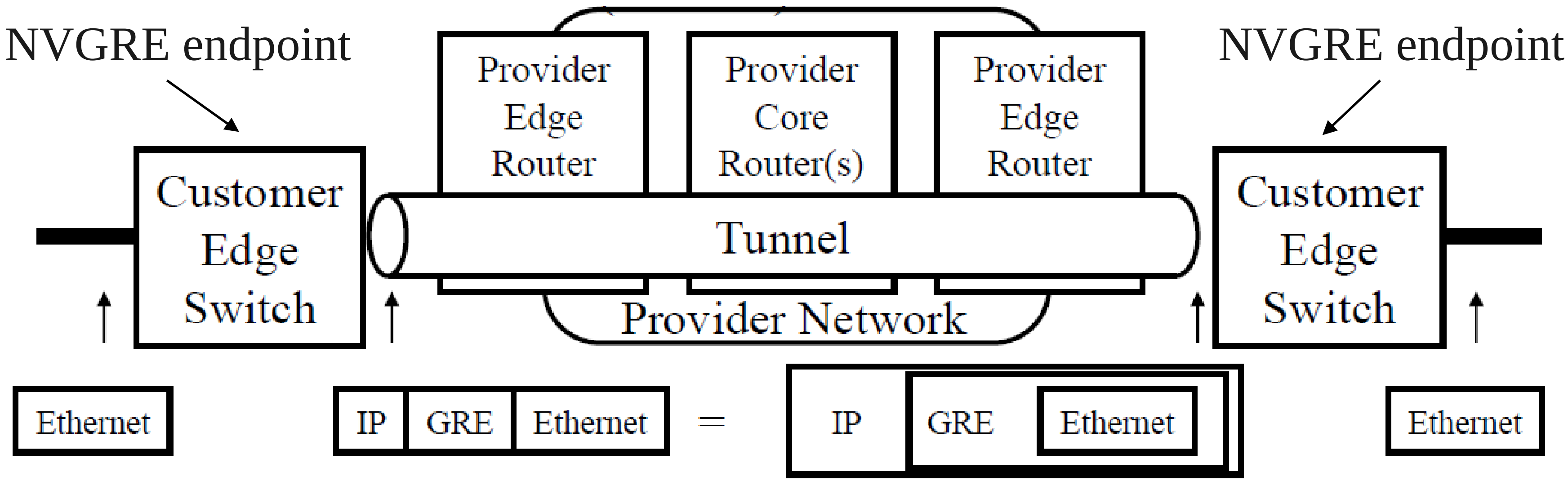
- NVGRE
- VXLAN
- STT
- Geneve

Network Virtualization using GRE (NVGRE)

- GRE
  - Generic Routing Encapsulation – RFC 2784
  - Generic →X over Y for any X and Y network protocol
  - Creates private point-to-point connection like a VPN
- NVGRE
  - Ethernet over GRE over IP (point-to-point)
  - Virtual Layer 2 topologies on top of a physical Layer 3 network
  - Unique 24-bit Tenant Network Identifier (TNI) used as lower 24-bits of GRE key field → 2^{24} tenants (more than 16 million)
  - Unique IP multicast address is used for BUM (Broadcast, Unknown, Multicast) traffic on each Virtual Subnet ID (VSID)

Source: P. Garg, Y. Wang, "NVGRE: Network Virtualization Using Generic Routing Encapsulation", IETF RFC 7637, Sep. 2015.

NVGRE – How it works

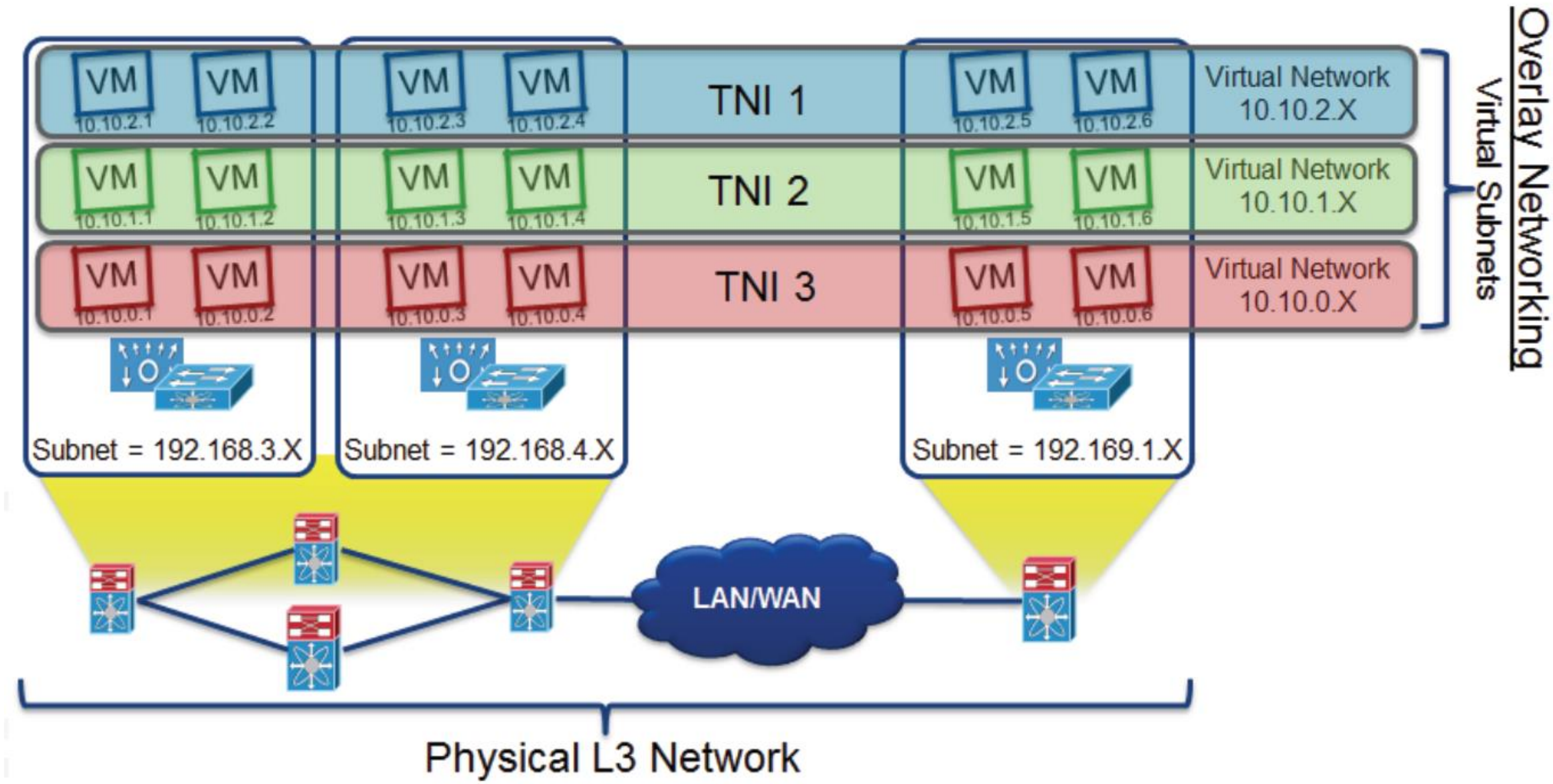


Outer MAC	Outer IP Header (PA)	GRE Header (Includes 24 Bit TNI)	Inner MAC	Inner IP Header (CA)	TCP Header	TCP user data

$\underbrace{\text{Inner MAC} \quad \text{Inner IP Header (CA)} \quad \text{TCP Header} \quad \text{TCP user data}}_{\text{Original Ethernet Frame}}$

CA – Customer Address
PA – Provider Address

NVGRE – How it works



Source: Emulex, NVGRE Overlay Networks: Enabling Network Scalability for a Cloud Infrastructure, White Paper, 2012.

Virtual Extensible Local Area Network (VXLAN)

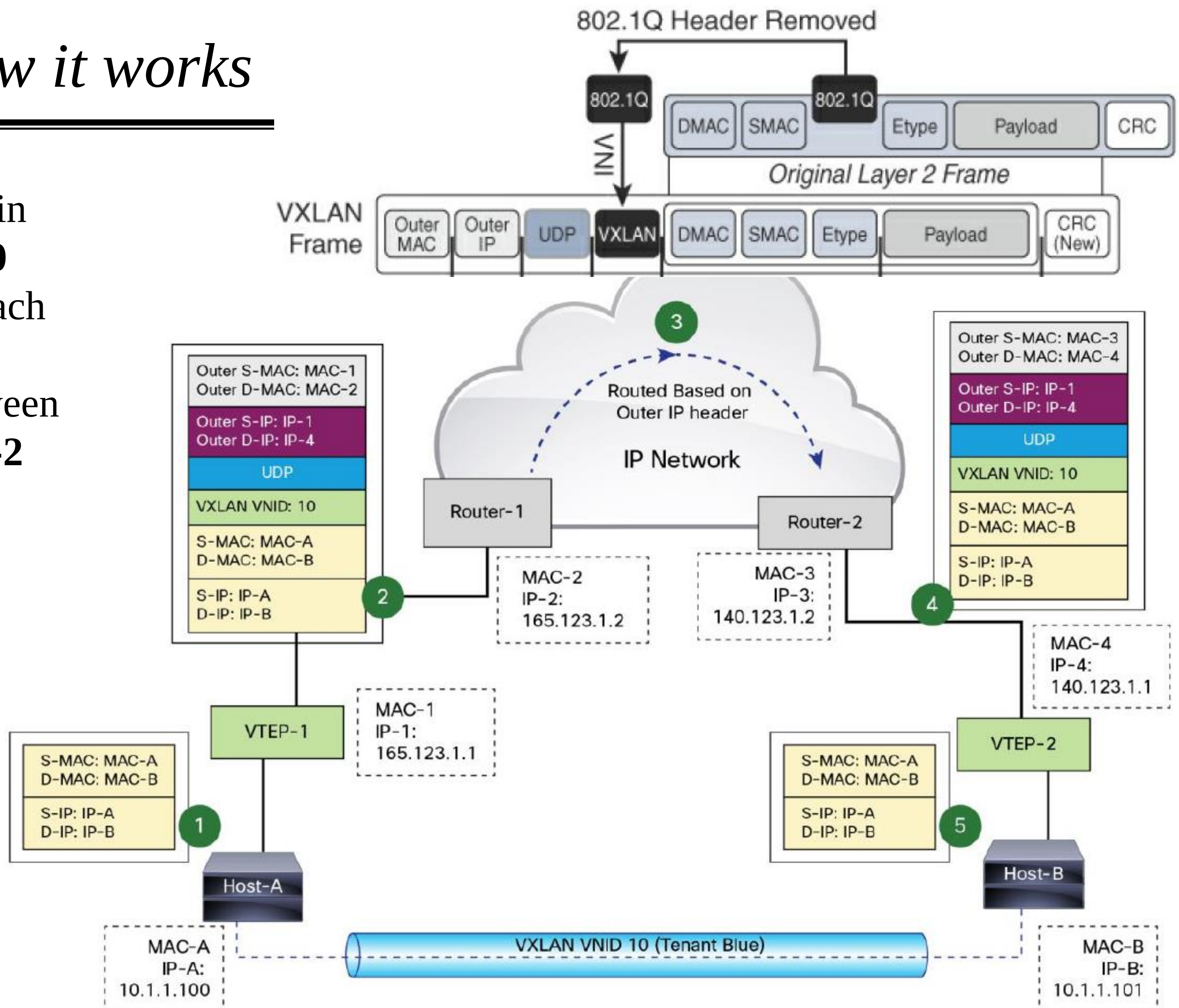
- Creates Virtual L2 overlay (called VXLAN) over L3 networks
  - 2^{24} VXLAN Network Identifiers (VNIs)
- Only VMs in the same VXLAN can communicate
- vSwitches serve as VTEP (VXLAN Tunnel End Point)
  - Encapsulate L2 frames in UDP over IP and send to the destination VTEP(s)
  - VTEPs can be <u>end hosts</u> or <u>network switches</u> or <u>routers</u>
- VMs belonging to different VXLAN segments may have overlapping MAC addresses and VLANs
  - L2 traffic never crosses a VNI
- Each **VXLAN segment** is mapped to an **IP multicast group** in the transport IP network

Source: P. Garg, Y. Wang, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", IETF RFC 7348, Aug. 2014.

VXLAN – How it works

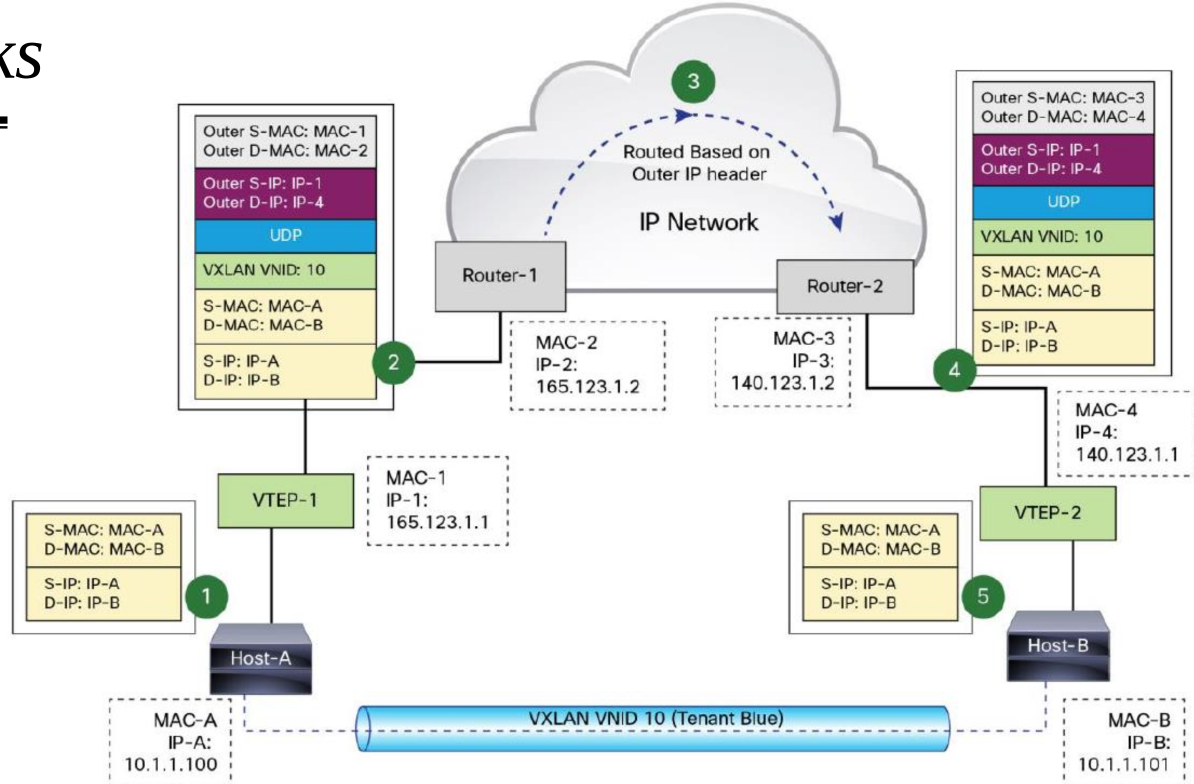
Host-A and **Host-B** in **VXLAN segment 10** communicate with each other through the VXLAN tunnel between **VTEP-1** and **VTEP-2**

(It is assumed address learning has been done on both sides, and corresponding MAC-to-VTEP mappings exist on both VTEPs)



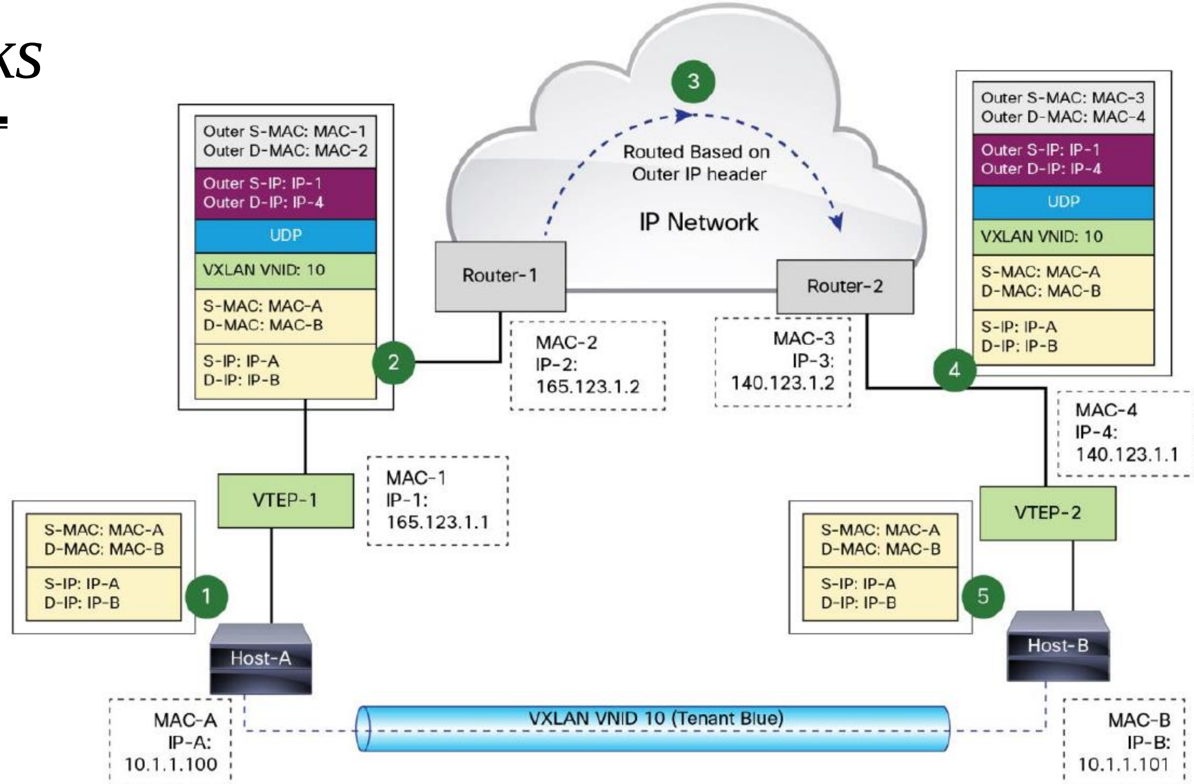
Source: Cisco, “VXLAN Overview: Cisco Nexus 9000 Series Switches”, White Paper, 2015.

VXLAN – How it works



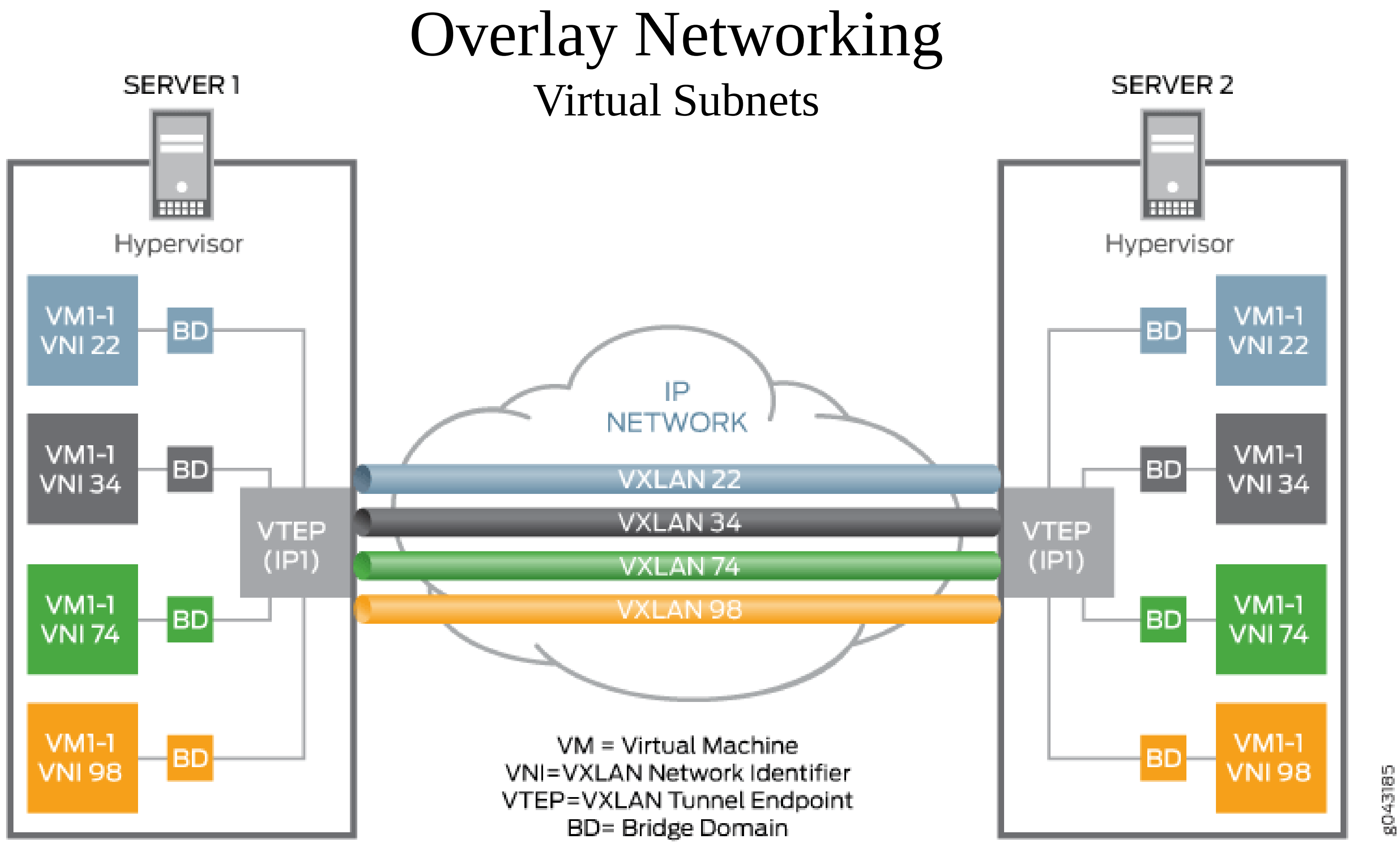
- Host-A forms Ethernet frame with MAC-B address and sends Ethernet frame to VTEP-1
- VTEP-1 has mapping of MAC-B to VTEP-2 in its mapping table
  - Performs VXLAN encapsulation on the packets by adding VXLAN, UDP, and outer IP header
  - Outer IP address → source IP address = VTEP-1; destination IP address = VTEP-2
  - VTEP-1 then performs an IP address lookup (ARP) for the IP address of VTEP-2 to resolve the next hop
  - Subsequently, uses MAC address of the next-hop device to further encapsulate the packets in an Ethernet frame to send to the next-hop device

VXLAN – How it works



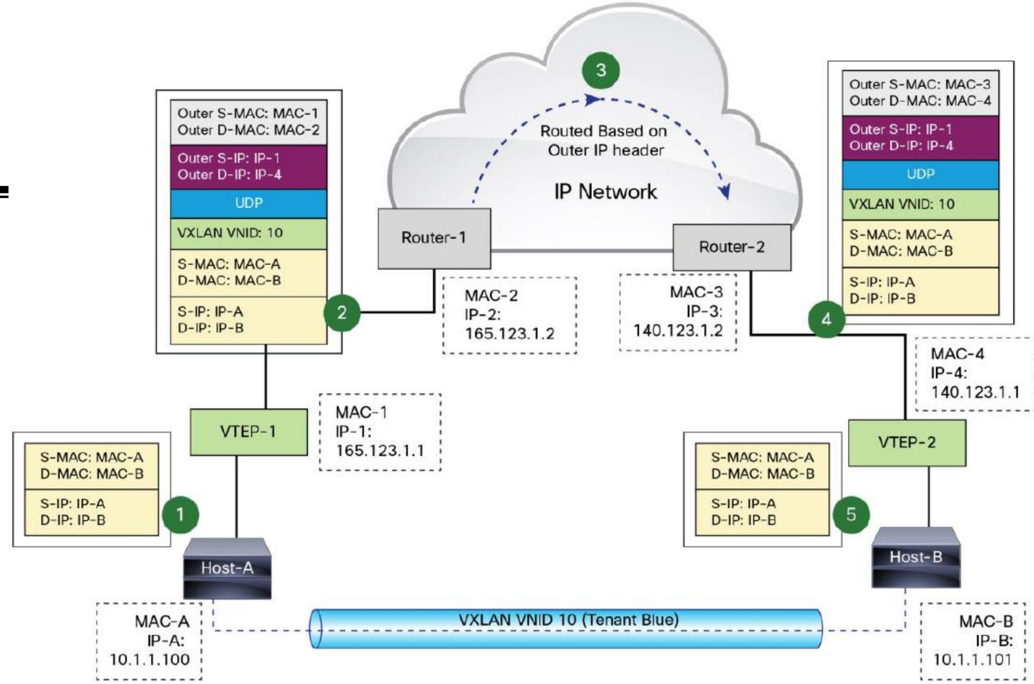
- Packets are routed toward VTEP-2 through the transport IP network
  - Based on outer IP header, which has the IP address of VTEP-2 as the destination address
- When VTEP-2 receives the packets, it strips off the outer Ethernet, IP, UDP, and VXLAN headers, and forwards the packets to Host-B, based on the original destination MAC address in the Ethernet frame

VXLAN – How it works



Source: https://www.juniper.net/documentation/en_US/junos/topics/concept/vxlan-evpn-integration-overview.html [Accessed: 10th May 2021]

VXLAN – How it works



- Source VM ARPs to find Destination VM's MAC address
    - All L2 multicasts/unknown are sent via IP multicast
    - Destination VM sends ARP response over IP unicast
- Destination VTEP learns inner-src-MAC to outer-src-IP mapping
    - Avoids unknown destination flooding for returning responses
- UDP source port is a hash of the inner MAC header
    - Allows load balancing using Equal Cost Multi Path using L3-L4 header hashing
- VMs are unaware that they are operating on VLAN or VXLAN

Stateless Transport Tunneling Protocol (STT)

- Ethernet over TCP-Like over IP tunnels
- Allows transmission of large frames → up to **64 kB**
  - Large frames segmented at the entrance of the tunnel according to the MTU of the physical network and reassembled at other endpoint of tunnel
  - Most other overlay protocols use UDP and disallow fragmentation
- TCP-Like: Stateless TCP
  - Header identical to TCP (same protocol number 6) but **no 3-way handshake**, no connections, no windows, no retransmissions, no congestion state
  - Stateless Transport (recognized by standard port number)
- Internet draft expired
  - Of historical interest only

Source: B. Davie, J. Gross, "A Stateless Transport Tunneling Protocol for Network Virtualization", IETF Internet Draft, draft-davie-stt-08, Apr. 2016.

Generic Network Virtualization Encapsulation (Geneve)

- Best of NVGRE, VXLAN, and STT
- **Generic**: Can virtualize any (L2/L3/…) protocol over UDP/IP
- **Tunnel Endpoints**: Process Geneve headers and control packets
- **Transit Device**: do not need to process Geneve headers or control packets

Source: J. Gross, I. Ganga, T. Sridhar, “Geneve: Generic Network Virtualization Encapsulation”, IETF RFC 8926, Nov. 2020.

Network Virtualization in Multi-tenant Data Centers

- Herein **focus was mostly on data plane**
- Control plane can follow different approaches
  - e.g., SDN, EVPN (BGP and more)
  - Data plane vs. control plane address learning
  - Join/Leave IP multicast groups in the underlay network

Summary

- Ethernet’s use of IDs as addresses makes it very easy to move VMs in the data center
    - Keep traffic on the same Ethernet
- Spanning tree is wasteful of resources and slow
- VLANs allow different non-trusting entities to share an Ethernet network → Yet, limited to 4k tenants
- Network Virtualization in Multi-tenant Data Centers
    - **NVGRE**, **VXLAN**, and **Geneve** solve the problem of multiple tenants with overlapping MAC addresses, VLANs, and IP addresses
    - Enable more than 16 million tenants
    - No changes to VMs → Hypervisors responsible for all details