Network Virtualization

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

A talk given at CS770 Networking Research Seminar at Washington University in Saint Louis, November 1, 2011

Audio/Video recordings of this lecture are available at:

http://www.cse.wustl.edu/~jain/talks/net_v.htm

Overview

1. TRILL: Transparent Interconnection of Lots of Links
2. OTV: Overlay Transport Virtualization
3. VXLAN: Virtual Extensible LAN

Virtualization Trend

- Virtual Memory ⇒ L1, L2, L3, ... ⇒ Recursive
- Virtual Desktop ⇒ Virtual Server ⇒ Virtual Datacenter

  Thin Client ⇒ VMs ⇒ Cloud
- Networks consist of:

  Hosts - L2 Links - L2 Bridges - L2 Networks - L3 Links - L3 Routers - L3 Networks – L4 Transports – L5 Applications
- Each of these can be virtualized
- This presentation is limited to L2 Network (LAN) virtualization

Why Virtualize?

- Ease of Management ⇒ Centralization
- Sharing ⇒ Carrier Hotels = Sharing buildings
- Cost Savings
- Isolation ⇒ Protection
- Dynamics: Replication, load balancing
- Mobility for fault tolerance

LAN Virtualization Technologies

- Problem: LANs were not designed for:
  1. Long distances
  2. Dynamic on-demand connectivity
  3. Very large number of nodes
  4. Multiple tenants

- Solutions:

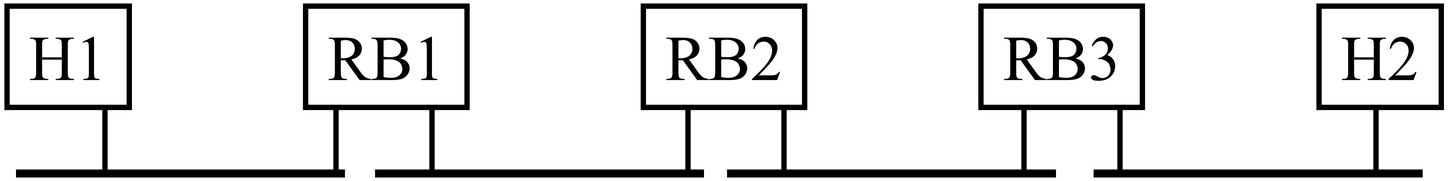
1. TRILL
2. Overlay Transport Virtualization
3. VXLAN

TRILL

- Transparent Interconnection of Lots of Links
- Allows an entire campus to be a single extended LAN
- IETF TRILL working group based on Radia Perlman's Infocom 2004 paper
- Problem:
    - LANs allow free mobility inside the LAN but
      Spanning tree is inefficient for a large campus LAN
      Many of the links are disabled
      Multipath is not allowed.
      Small changes in network ⇒ large changes in spanning tree
    - Subnets provide efficient utilization of links but mobility is a problem because IP addresses change from one subnet to next and break transport connections

TRILL (Cont)

- Solution: Take the best of both worlds
  Use MAC addresses and IP routing
  RBridges use IS-IS to route MAC frames but learn addresses.
- RBridges run IS-IS to compute pair-wise optimal paths for unicast and distribution trees for multicast
- L2 frames are encapsulated and sent to destination RBridge
  Header contains a hop-limit to avoid looping



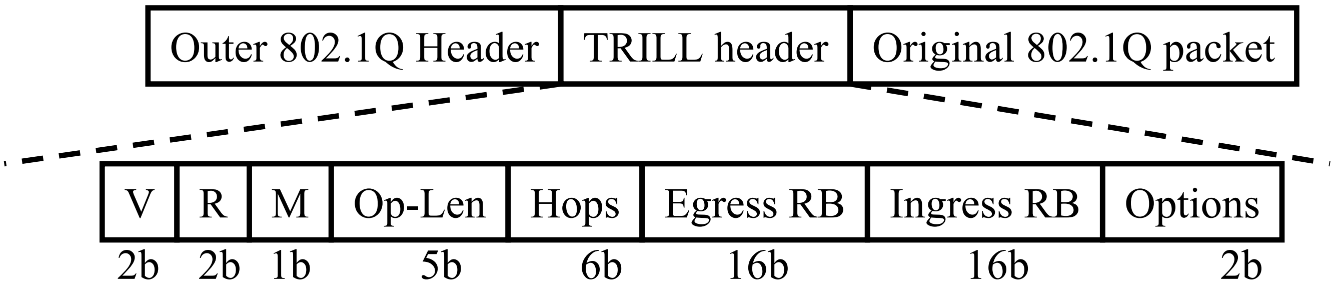
Ref: Ref: R. Perlman, "RBridges: Transparent Routing," Infocom 2004

TRILL (Cont)

- Each RBridge gets a 2B IS-IS ID which is unique within the campus using a “nickname” protocol
- Each VLAN on the link has one (and only one) designated RBridge using IS-IS election protocol
- RBridge learn source MAC addresses by snooping and announce their MAC table to other RBridges

Ref: Ref: R. Perlman, "RBridges: Transparent Routing," Infocom 2004

TRILL Encapsulation



Outer 802.1Q Header	TRILL header	Original 802.1Q packet

V	R	M	Op-Len	Hops	Egress RB	Ingress RB	Options
2b	2b	1b	5b	6b	16b	16b	2b

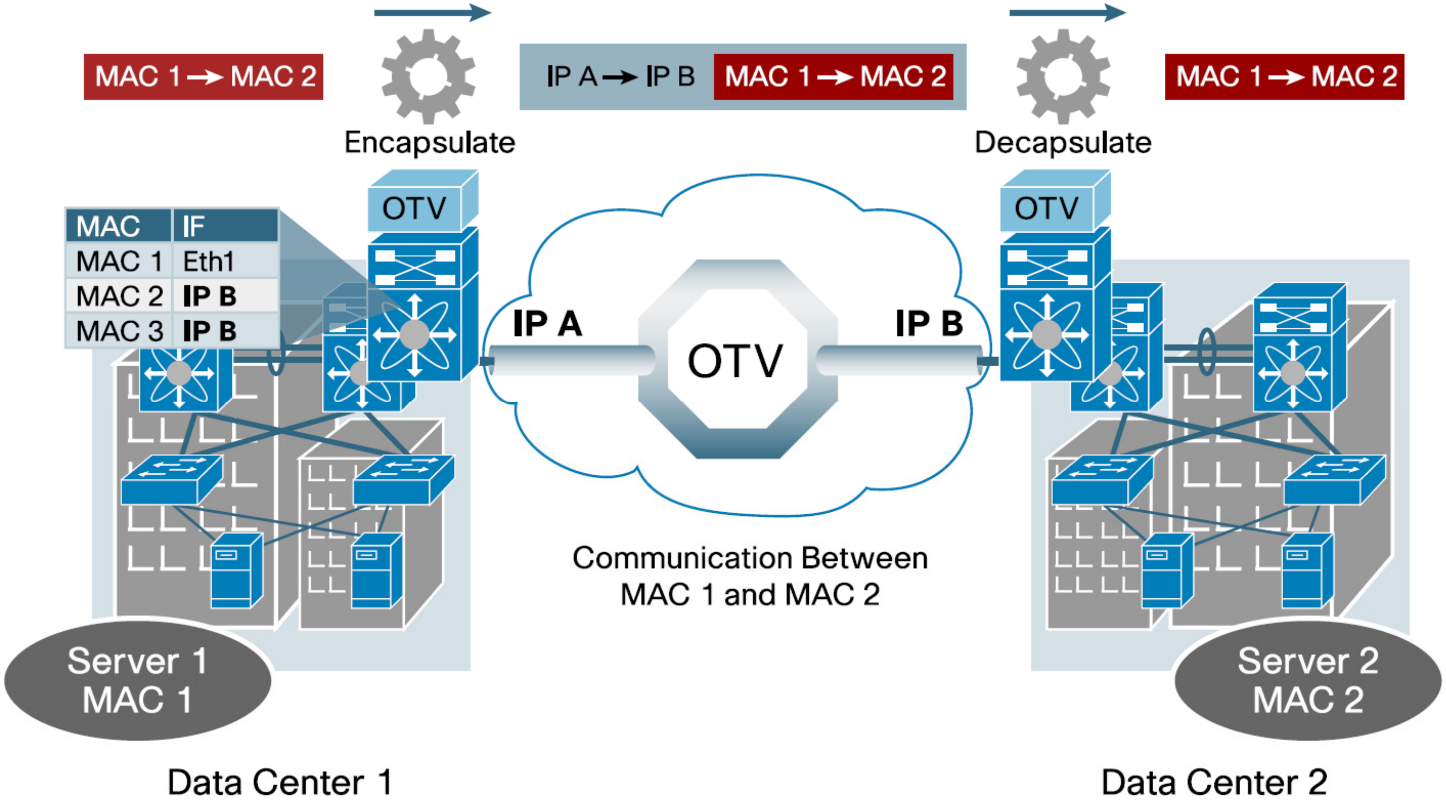
- Version, Reserved, Multi-destination, Options length, Hops
- For outer headers both PPP and Ethernet headers are allowed.
- Outer VLAN ID is the VLAN used for TRILL
  Outer VLAN priority is copied from inner VLAN tag

TRILL: Summary

- TRILL allows a large campus to be a single IP subnet
- Packets are encapsulated and routed using IS-IS routing in L2

Overlay Transport Virtualization (OTV)

- Cisco technology for LAN Extension over IP



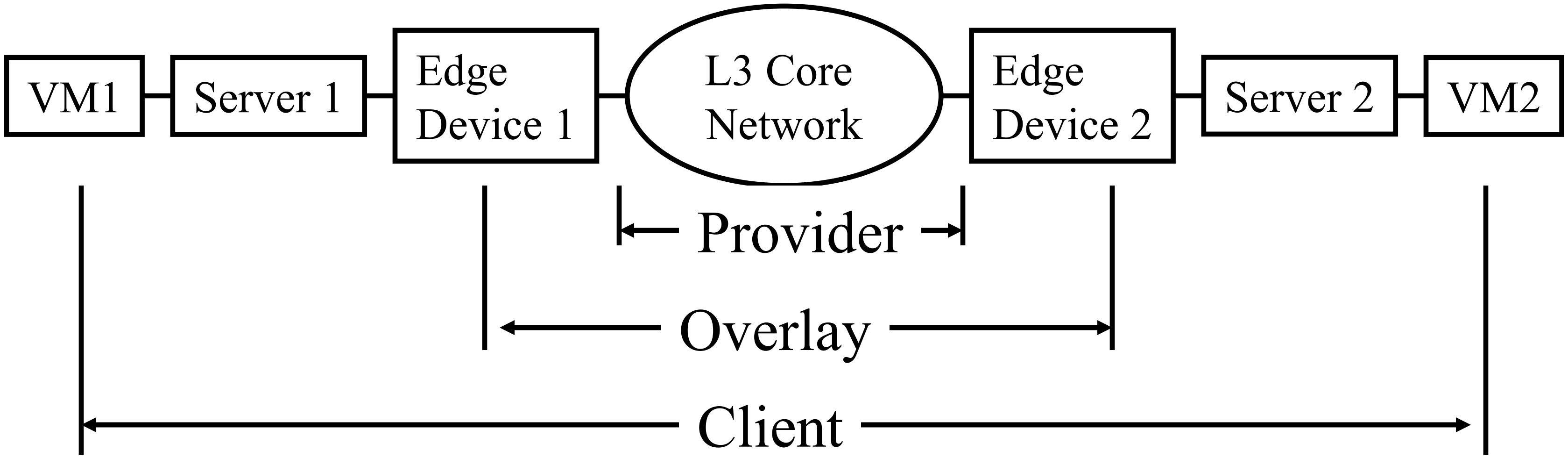
MAC	IF
MAC 1	Eth1
MAC 2	**IP B**
MAC 3	**IP B**

Ref: [Cisco-OTV] Cisco, “Enhance Business Continuance with Application Mobility Across Data Centers,” http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-591960.pdf

OTV Features

- Allows a single LAN to span multiple data centers connected via IP over a WAN
- MAC in IP: 802.3 packets are encapsulated and transmitted over to the destination data center
- Edge switches maintain a list of all MAC addresses in all data centers
- Provides fault tolerance ⇒ Applications migrate from down data center to another
- Allows load balancing by moving VMs to datacenters close to the client or "follow the sun"

OTV Control Plane



- Edge devices join a provider multicast group in the core
- Overlay control plane provides discovery of edge devices and exchange of MAC address reachability data
- Overlay sets up adjacency between only edge devices in the same VPN
- One Edge device can support multiple overlay VPNs
- Each VPN may have multiple VLANs. VLAN IDs among VPNs may overlap but are treated as disntict

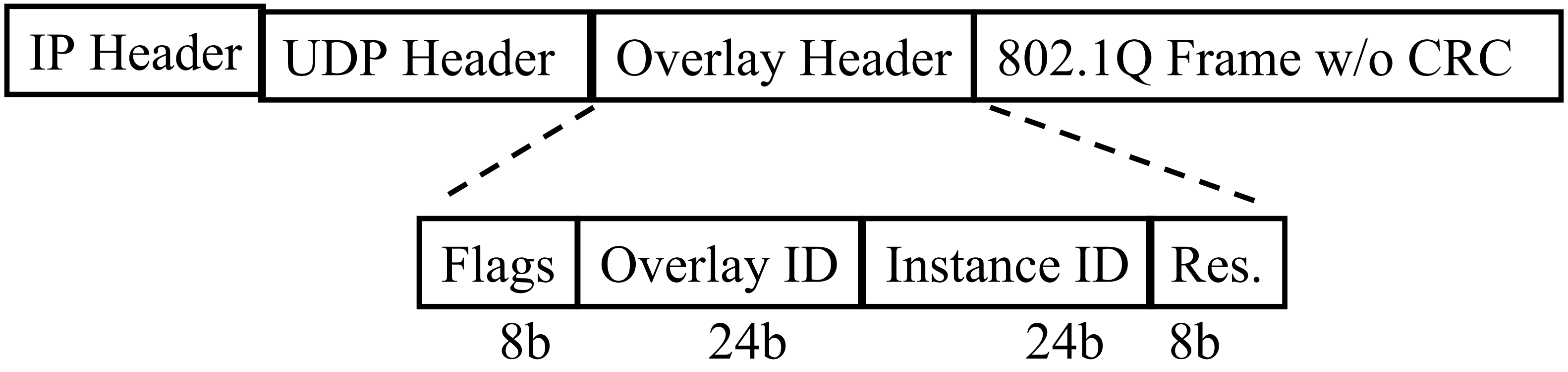
OTV Control Plane (Cont)

- Edge device is distinct from provider routers and so it does not participate in core routing exchange
- Edge devices participate in an overlay routing exchange
- Edge devices route packets based on MAC addresses
  ⇒ "MAC Router"
- Edge devices are IP hosts in provider network, MAC routers in overlay network, and Bridges in client network.
- Edge devices participate in Spanning Tree Protocol on the internal interface. There is no STP on the external interface.
- Unknown and Spanning tree messages do not cross a data center ⇒ Limits broadcast storms

OTV Control Plane (Cont)

- A site may be multi-homed. An authoritative edge device per VLAN is elected to take the frames on/off the overlay network.
- IS-IS protocol is used as the overlay control protocol
- Multicasts are sent via IP multicast trees
- Uses equal cost multi path

OTV Data Plane



IP Header | UDP Header | Overlay Header | 802.1Q Frame w/o CRC

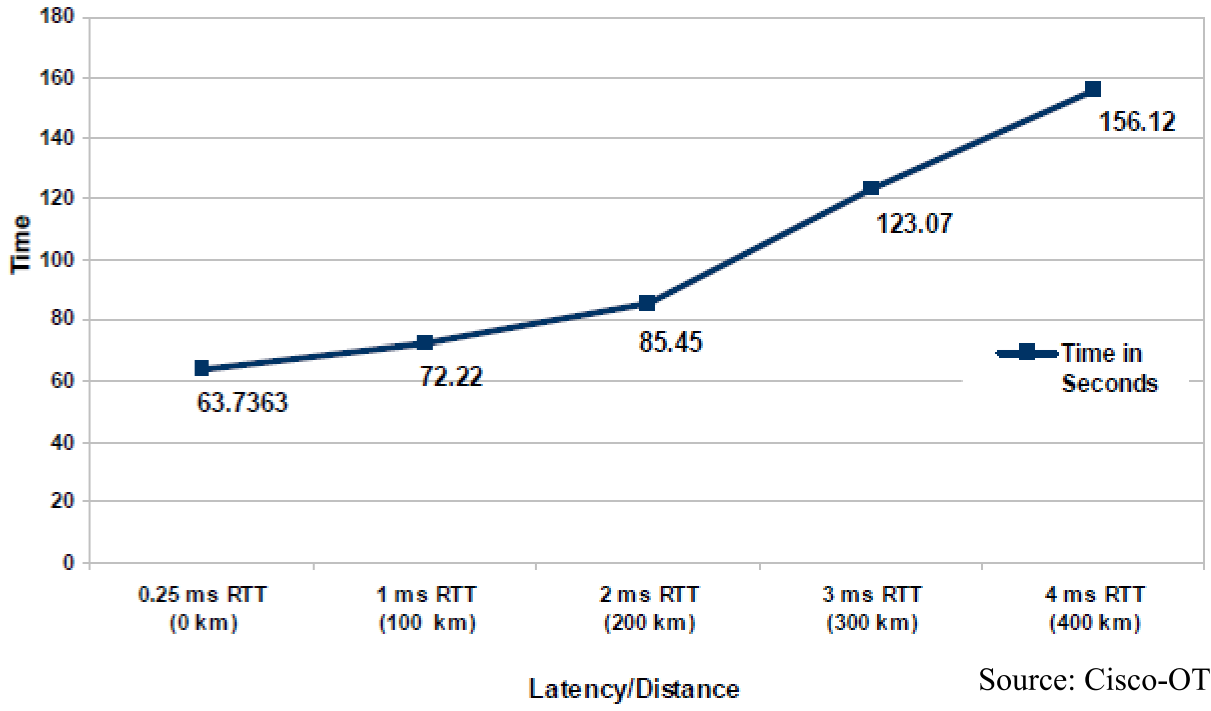
Flags | Overlay ID | Instance ID | Res.

8b 24b 24b 8b

- L2 802.1Q frame encapsulated in UDP inside IPv4/IPv6 UDP destination Port = 8472 ⇒ Overlay Transport Protocol
- Don’t fragment bit is set to 1 ⇒ Core network should be able to support encapsulated Ethernet packets.
- 4-byte CRC is removed and 8 byte overlay header is added.
- I-Flag bit ⇒ Destination edge device should use forwarding table for that particular instance

Delay Performance

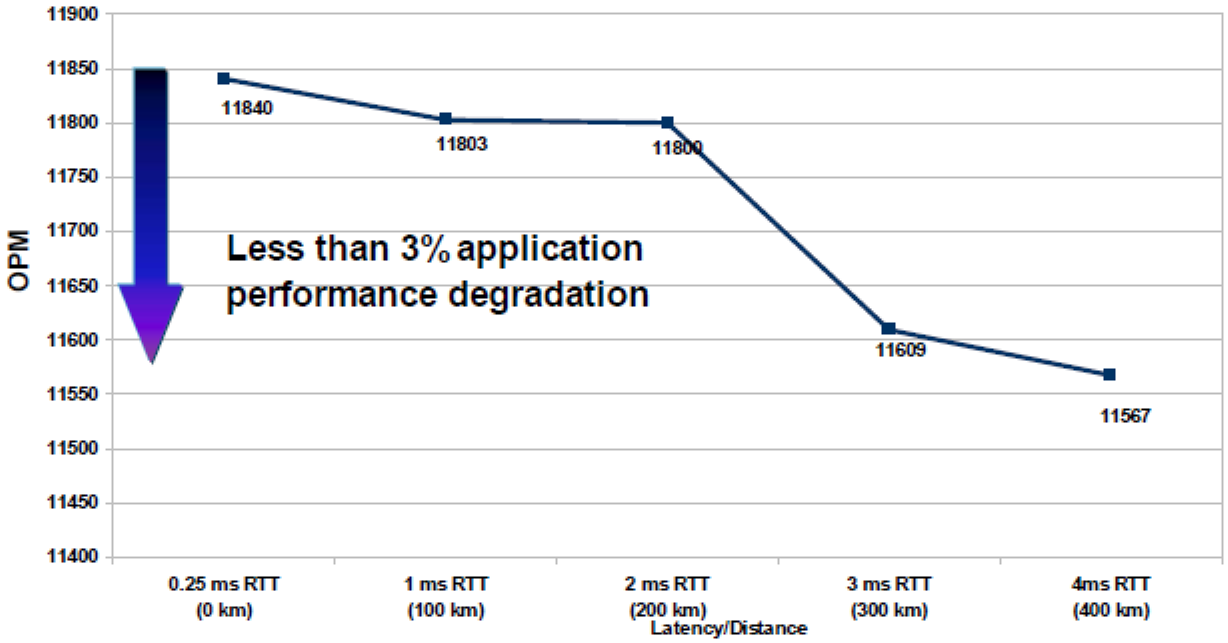
- Duration of intra- and Inter- datacenter VM migration



Source: Cisco-OTV

Throughput Performance

- Orders per minute (OPM)
- Less than 3% difference over 400 km



Source: Cisco-OTV

OTV: Summary

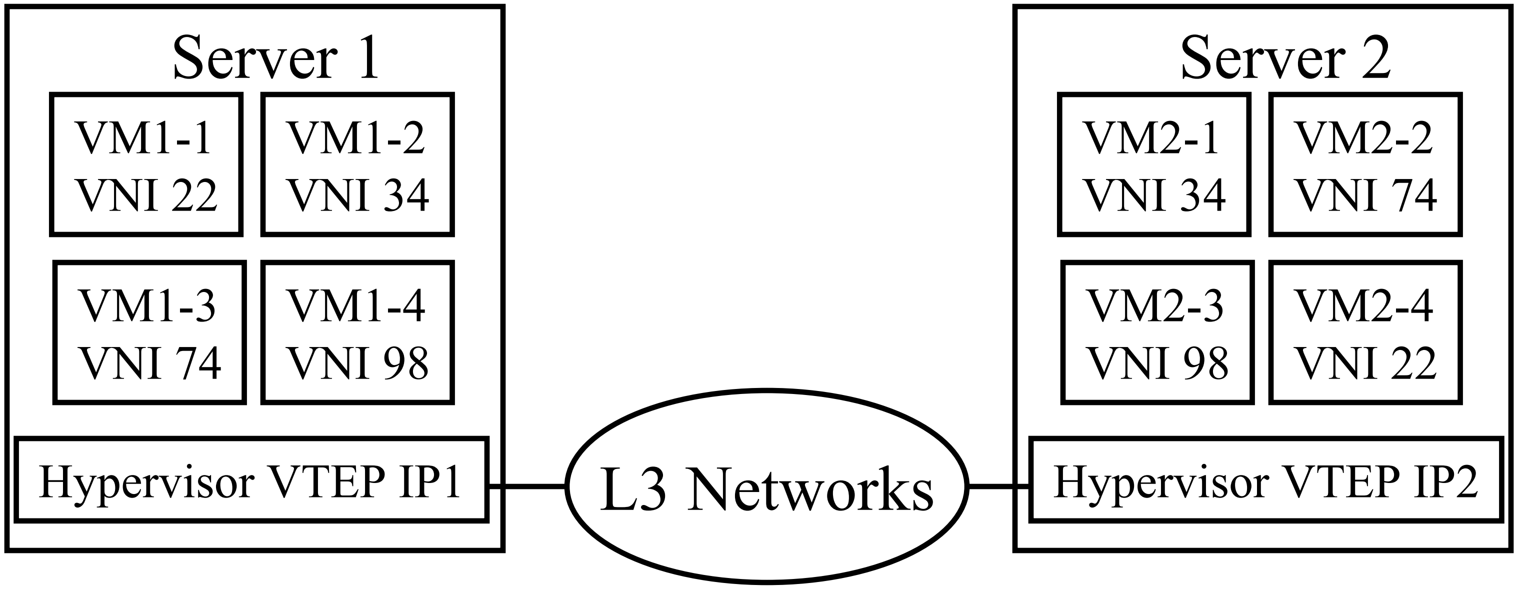
- OTV allows a single LAN to span multiple datacenters located far apart
- Encapsulates L2 frames and sends using L3

VXLAN

- Virtual Extensible Local Area Networks
- Developed by VMware
- Supported by many companies for standardization in IETF
- Allows overlay networks within virtualized datacenters (public or private clouds) accommodating multiple tenants
- Problems:
  - VMs have increased the need for MAC addresses and VLANs
  - 4096 VLANs are not sufficient
  - Multiple tenants need their own networking domains with their own control over VLAN IDs
  - Spanning tree is inefficient with this large number.
    ⇒ Too many links are disabled
  - Better throughput with IP equal cost multipath (ECMP)

Ref: VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks, draft-mahalingam-dutt-dcops-vxlan-00, 2011-08-27

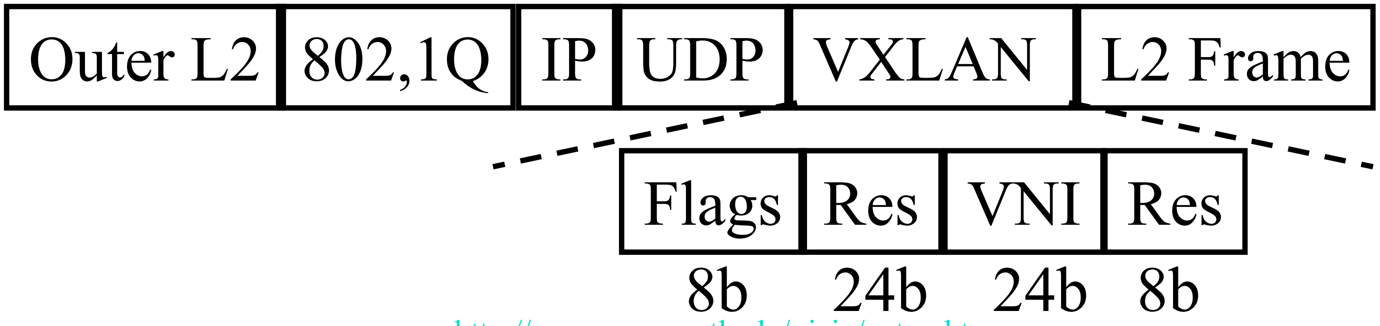
VXLAN Deployment Example



- 4 VXLAN segments

VXLAN Architecture

- VXLAN allows many L2 overlays over an L3 network
- Each L2 overlay is called "VXLAN Segment"
  24b segment VXLAN Net ID (VNI)
  ⇒ 16M segments within the same administrative domain
- VMs can communicate with other VMs in the same segment
- Segments may have overlapping MAC addresses and VLANs but L2 traffic never crosses
- Uses tunneling to overlay Segments over L3
  Tunnels end points (VTEP) in hypervisors
- VTEP encapsulates L2 frames and sends to dest VTEP via IP:



Outer L2 | 802,1Q | IP | UDP | VXLAN | L2 Frame

Flags | Res | VNI | Res

8b | 24b | 24b | 8b

VXLAN (Cont)

- Outer VLAN tag is optional. Used to isolate VXLAN traffic on the LAN
- Destination VTEP learns inner-Src-MAC-to-outer-src-IP mapping ⇒ Avoids unknown dest flooding for returning responses
- Source VM ARPs to find Destination VM's MAC address. This packet is encapsulated and sent via IP multicast. Dest VM sends a standard IP unicast ARP response.
- IGMP is used to prune multicast trees
- Multicast is used for carrying unknown dest, broadcast/multicast L2 frames.
- I flag is set if VNI field is valid
- UDP source port is a hash of the inner MAC header
  ⇒ Allows good load balancing using Equal Cost Multi Path

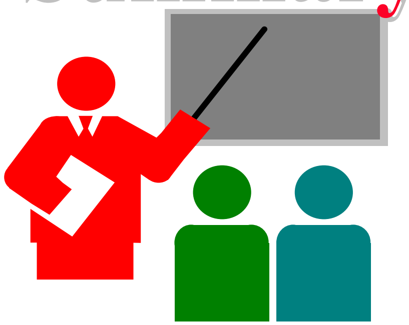
VXLAN (Cont)

- Inner VLAN tags are discarded. Outer tags are sufficient.
- A VXLAN gateway switch can forward traffic to/from non-VXLAN networks. Encapsulates or decapsulates the packets.

VXLAN: Summary

- ❑ VXLAN solves the problem of multiple tenants in a cloud environment.
- ❑ A server may have VMs belonging to different tenants
- ❑ Allows each tenant to have their own VLANs that connect their VMs

Summary



1. Ethernet is being extended to cover multiple tenants in multiple data centers and large campuses
2. Most of these efforts encapsulate Ethernet frames and transport them using layer 3 protocols
3. TRILL is mostly for large campuses
4. OTV allows LANs covering multiple datacenters
5. VXLAN allows multiple tenants on the same server using their own VLANs
6. Networks are being “flattened” (L2 end-to-end)