LAN Extension and Network Virtualization in Cloud Data Centers

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/cse570-23/

Student Questions

Q on 21, 27, 28

1. TRILL
2. NVGRE
3. VXLAN
4. NVO3
5. Geneve
6. EVPN
7. GUE

Student Questions

Geographic Clusters of Data Centers

- Multiple data centers are used to improve availability
- Cold-Standby: Data is backed up on tapes and stored off-site. In case of disaster, applications and data are loaded on standby. Manual switchover ⇒ Significant downtime. (1970-1990)
- Hot-Standby: Two servers in geographically close data centers exchange state and data continuously.
  Synchronous or Asynchronous data replication to standby.
  On a 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.

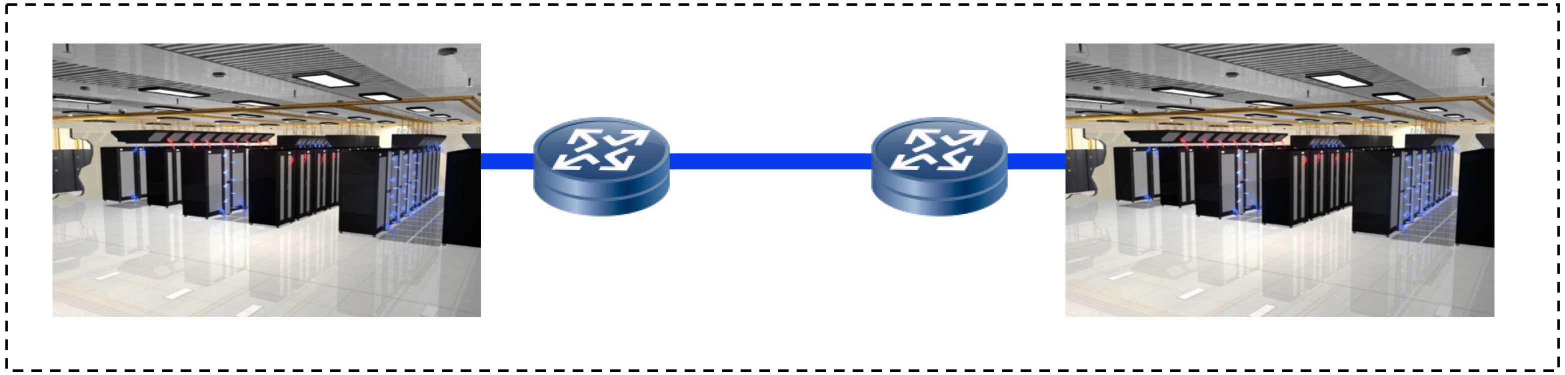
Ref: G. Santana, "Datacenter Virtualization Fundamentals," Cisco Press, 2014, ISBN: 1587143240

Student Questions

- What's the difference between active-active and hot standby? Is it always that hot standby is synchronous and cold standby is asynchronous?
  - *Cold=Standby mostly down*
  - *Hot=Standby always up*
  - *Synchronous=Continuously*
  - *Asynchronous=Periodically*
  - *Active-Active=Both serving continuously and also ready to serve other customers*

Data Center Interconnection (DCI)

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



Student Questions

- Does the distance or lack of carriers without L2 services prevent us from connecting these data centers in layer 2?

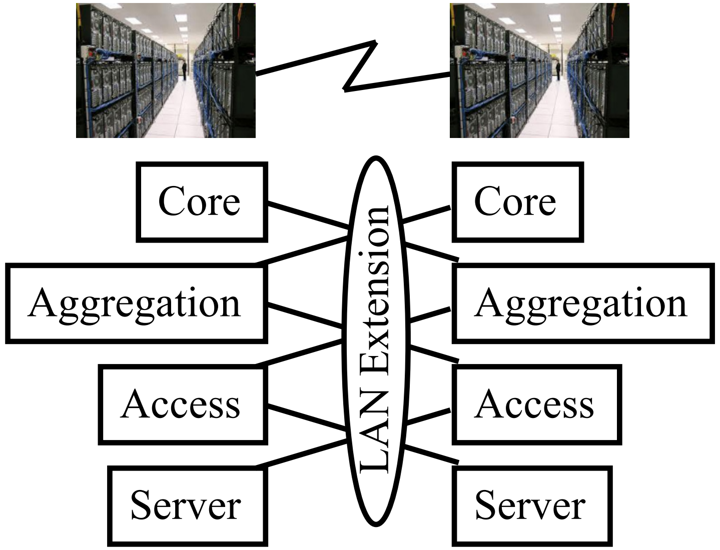
No. You can run/lease fibers between cities.

- How could DCI overcome the latency issue?

With CSMA/CD gone, there is no latency issue at L2. Active users are served mainly by the closest data center.

Challenges of LAN Extension

- **Broadcast storms**: Broadcast, Unknown, and Multicast (BUM) may create an excessive floo
- **Loops**: Easy to form loops in a large network.
- **STP Issues**:
  - High spanning tree diameter (leaf-to-leaf): More than 7.
  - The root can become a bottleneck and a single point of failure
  - Multiple paths remain unused
- **Tromboning**: Dual attached servers and switches generate excessive cross-traffic
- **Security**: Data on LAN extension must be encrypted



Student Questions

- Can you explain trombone again?

Tromboning: If two switches in different data centers try to be on standby for each other, they will have to exchange state, which could result in excessive long-distance traffic.

- Why is security an issue for LAN extension if, at the OS level, the encryption could be done?

On internal LANs, no encryption is not needed/used.

TRILL

- Transparent Interconnection of Lots of Links
- Allows a large campus to be a single extended LAN
- LANs allow free mobility inside the LAN but:
  - Inefficient paths using Spanning tree
  - Inefficient link utilization since many links are disabled
  - Inefficient link utilization since multipath is not allowed.
  - Unstable: small changes in network ⇒ large changes in spanning tree
- IP subnets are not good for mobility because IP addresses change as nodes move and break transport connections, but:
  - IP routing is efficient, optimal, and stable
- Solution: Take the best of both worlds
  ⇒ Use MAC addresses and IP routing

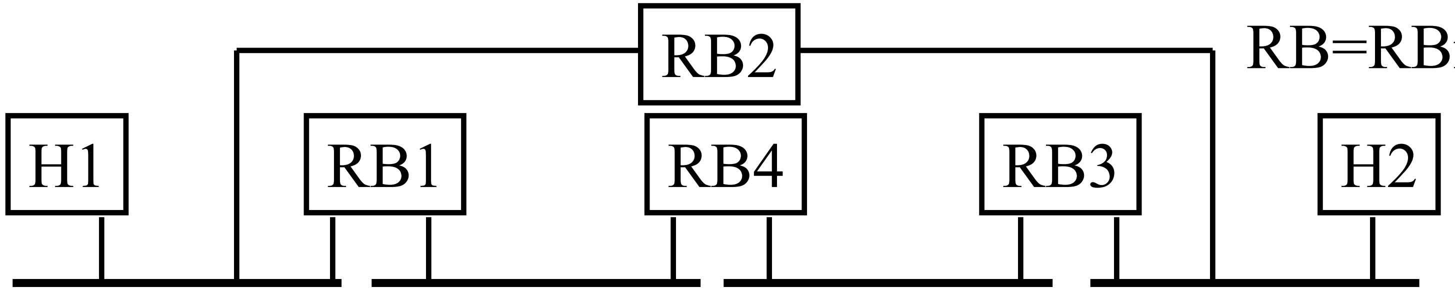
Ref: RFCs 5556, 6325, 6326, 6327, 6361, 6439

Student Questions

TRILL Architecture

- Routing Bridges (RBridges) encapsulate L2 frames and route them to destination Rbridges, which decapsulate and forward
- Header contains a hop limit to avoid looping
- RBridges run IS-IS to compute pair-wise optimal paths for unicast and distribution trees for multicast
- RBridge learn MAC addresses by source learning and by exchanging their MAC tables with other RBridges
- Each VLAN on the link has one (and only one) designated RBridge using IS-IS election protocol

H= Host
RB=RBridge

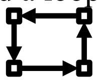


Ref: R. Perlman, "RBridges: Transparent Routing," Infocom 2004

Student Questions

- How does the hop-limit avoid a loop?

Same as in IP.



- What to do if the hop limit is exceeded?

Drop the packet

- What do you mean by VLAN on the link?

There are multiple VLANs on a network. Some VLANs may go through this link. Others may not.

- What is the job of the designated RBridge? Is it the root? *Designated=Default*
- Can TRILL compute or determine the shortest path before sending packets?

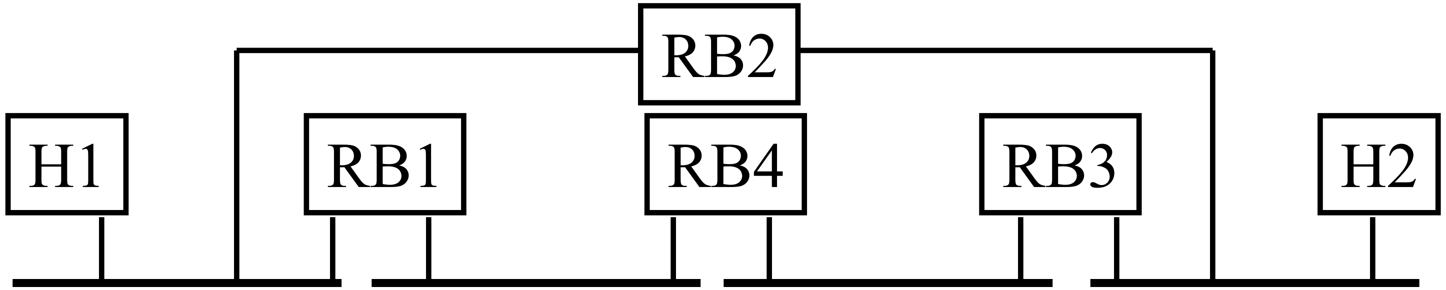
Yes. RBridges compute the path to all other bridges as in IP.

- What is H1, H2?

Now added to the figure.

TRILL Architecture

- Routing Bridges (RBridges) encapsulate L2 frames and route them to destination Rbridges, which decapsulate and forward
- Header contains a hop limit to avoid looping
- RBridges run IS-IS to compute pair-wise optimal paths for unicast and distribution trees for multicast
- RBridge learn MAC addresses by source learning and by exchanging their MAC tables with other RBridges
- Each VLAN on the link has one (and only one) designated RBridge using IS-IS election protocol



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

Student Questions

- Does the hop limit solve the problem of loops? It only seems to prevent infinite forwarding, but doesn't prevent the loop from occurring in the first place or if the packet is retransmitted and loops again.

It limits the damage until the problem is solved.

TRILL Encapsulation Format

Outer Header	TRILL header	Original 802.1Q packet

TRILL header:

Version	Res.	Multi-Destination	Options Length	Hops to Live	Egress RBridge	Ingress RBridge	Options
2b	2b	1b	5b	6b	16b	16b	

- For outer headers, both PPP and Ethernet headers are allowed. PPP for the long haul.
- Outer Ethernet header can have a VLAN ID corresponding to the VLAN used for TRILL.
- Priority bits in outer headers are copied from inner VLAN

Student Questions

- Each time the frame is sent from one RBridge to another RBridge, do we change the Egress and Ingress bridges in the packet's header?

No. Those are edge bridges. That is where the TRILL header is added or removed.

- What is PPP?
- Point-to-Point Protocol

This was the most common L2 protocol in use before Ethernet.

- Who decides what address is in the egress and ingress fields?

The first bridge puts its address as ingress RBridge. Computes the shortest path to the destination and puts the Egress RBridge.

- Is TRILL similar to MPLS in that at every hop, one is swapping labels, and the other is swapping addresses? *Yes.*

TRILL Encapsulation Format

Outer Header	TRILL header	Original 802.1Q packet

Version	Res.	Multi-Destination	Options Length	Hops to Live	Egress RBridge	Ingress RBridge	Options
2b	2b	1b	5b	6b	16b	16b	

- ❑ For outer headers, both PPP and Ethernet headers are allowed. PPP for the long haul.
- ❑ Outer Ethernet header can have a VLAN ID corresponding to the VLAN used for TRILL.
- ❑ Priority bits in outer headers are copied from inner VLAN

Student Questions

- ❑ You say the outer Ethernet header can have a VLAN ID; does this mean it doesn't need it? If it doesn't have one, will TRILL take care of it?

TRILL is in the payload. It can be sent with or without the VLAN header.

TRILL Features

- Transparent: No change to capabilities.
  Broadcast, Unknown, Multicast (**BUM**) support. Auto-learning.
- Zero Configuration: RBridges discover their connectivity and learn MAC addresses automatically
- Hosts can be multi-homed
- VLANs are supported
- Optimized route
- No loops
- Legacy bridges with spanning trees in the same extended LAN

Student Questions

- At present, is the TRILL structure the best application in the data center?

No. Didn't catch on.

- What do you mean by Auto-Learning?

They are learning by looking at the source addresses.

- Is TRILL still being used now? if not, why? What is the most common L2 routing protocol currently used?

TRILL is not currently used. The protocols were discussed along with shortest path routing.

TRILL: Summary

- TRILL allows a large campus to be a single Extended LAN
- Packets are encapsulated and routed using IS-IS routing

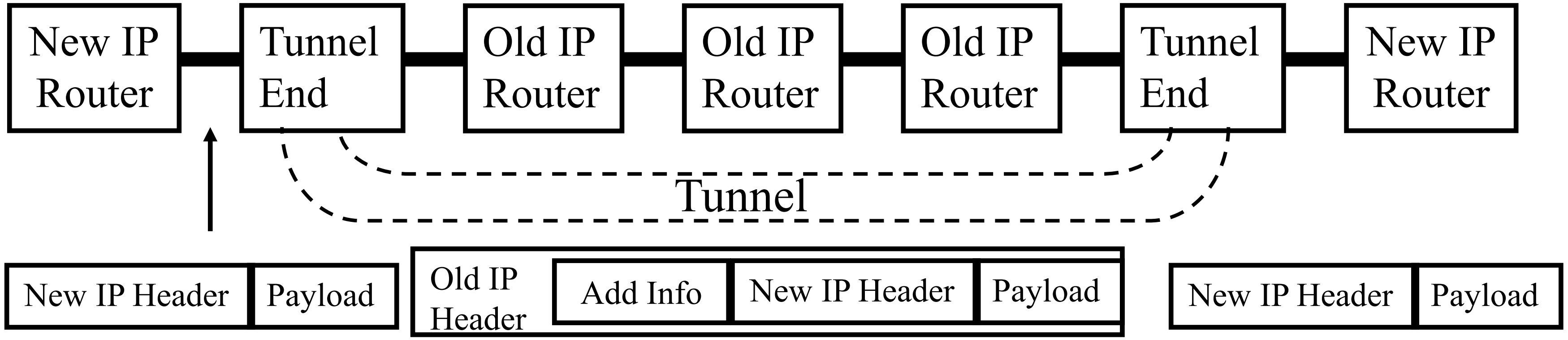
Student Questions

- How did 100Mbit Ethernet get away with copying FDDI protocols?

The same companies developed both.

GRE

- Any new feature in IP requires *encapsulation*, a.k.a. *tunneling*
- Generic Routing Encapsulation (RFC 1701/1702)
- Generic ⇒ X over Y for any X or Y protocols
- Given *n* protocols, we need O(n^2) encapsulation formats, GRE converts this to O(1) format.
- Encapsulations may require the following services:
  - Stream multiplexing: Which recipient at the other end?
  - Source Routing: what path to take?
  - Packet Sequencing



Student Questions

- Can you explain more about GRE tunneling? *Tunneling = Encapsulation = X over Y*
- I don't understand the 4th item. Could you please explain it in more detail? *Sure.*
- When is the GRE packet constructed? Is it at the router at the start of the tunnel?

Yes.

- What are some practical use cases or applications of the GRE?

Encapsulation

GRE (Cont)

- GRE provides all of the above encapsulation services
- Over IPv4, GRE packets use a protocol type of 47
- Optional: Checksum, Loose/strict Source Routing, Key
- Key is used either to authenticate the source or to distinguish different substreams
- Recursion Control: # of additional encapsulations allowed. 0 ⇒ Restricted to a single provider network ⇒ end-to-end
- Offset: Points to the next source route field to be used
- IP or IPSec are commonly used as delivery headers

Delivery Header	GRE Header	Payload

Check-sum Present	Routing Present	Key Present	Seq. # Present	Strict Source Route	Recursion Control	Flags	Ver. #	Prot. Type	Offset	Check sum (Opt)	Key (Opt)	Seq. # (Opt)	Source Routing List (Opt)
1b	1b	1b	1b	1b	3b	5b	3b	16b	16b	16b	32b	32b	Variable

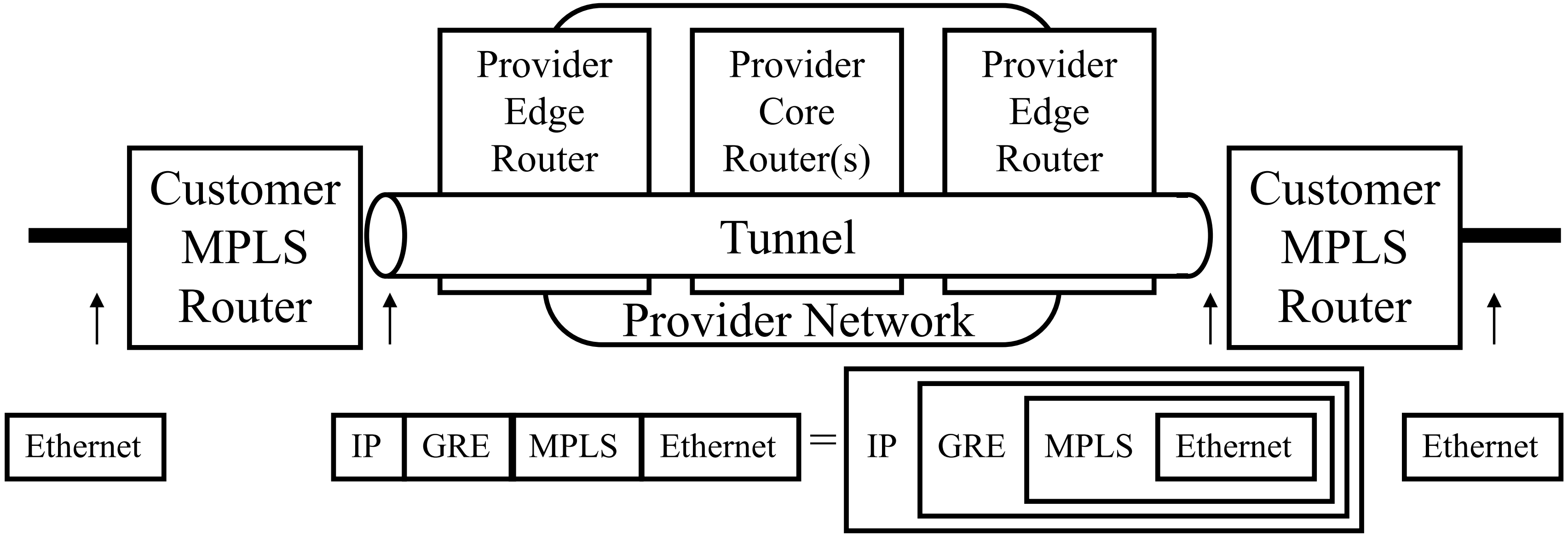
Student Questions

- Are there any restrictions for the loose routing since it does not need to go through other routers？ *No. Any router at the edge will do.*
- The key here is referring to the public key, correct? Why, if the recursion control=0, don't encrypt?

No. I may not want to go over other domains.

EoMPLSoGRE

- Ethernet over MPLS over GRE (point-to-point)
  VPLS over MPLS over GRE (Multipoint-to-multipoint)
- Used when a provider offers only L3 connectivity
  Subscribers use their own MPLS over GRE tunnels
- VPLSoGRE or Advanced-VPLSoGRE can also be used
- GRE offers an IPSec encryption option



Student Questions

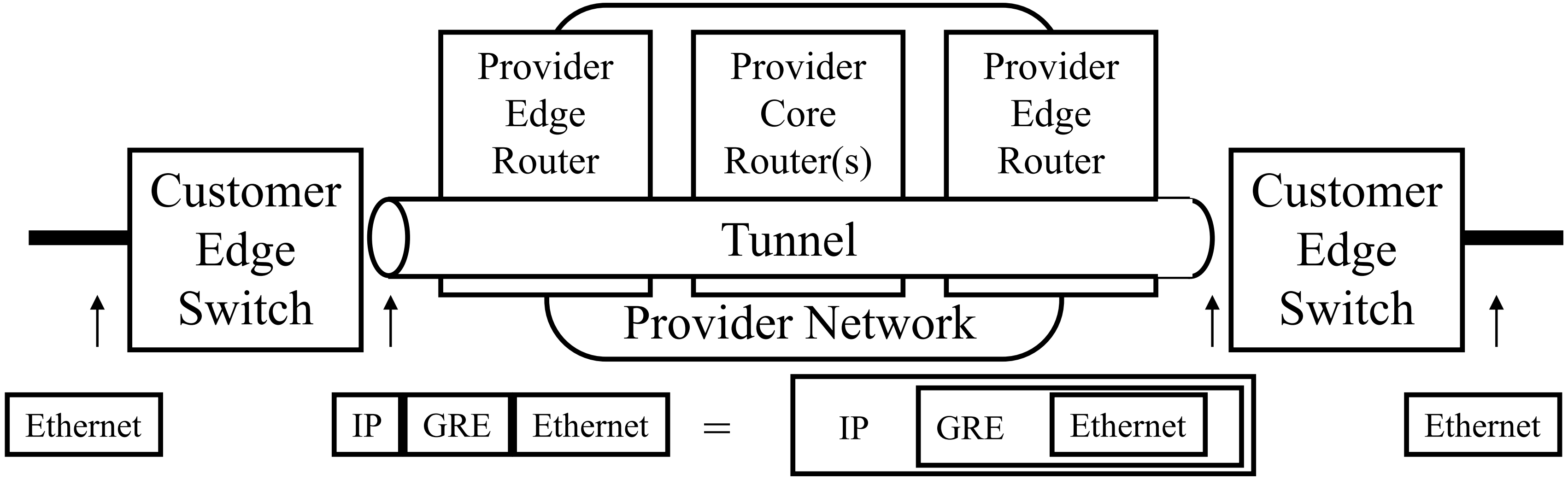
- what do you mean by "Advanced" VPLSoGRE? *VPLSoGRE v2*
- Why would the customer want to carry Ethernet over MPLS to a remote location?

A US company has a branch in Korea. The traffic may go through Russia.

End of Part 1

NVGRE

- Ethernet over GRE over IP (point-to-point)
- A unique 24-bit Virtual Subnet Identifier (VSID) is used as the lower 24 bits of the GRE key field ⇒ 2^{24} tenants can share
- A unique IP multicast address is used for BUM (Broadcast, Unknown, Multicast) traffic on each VSID
- Equal Cost Multipath (ECMP) allowed on point-to-point tunnels



Ref: P. Garg, Y. Wang, et al. "NVGRE: Network Virtualization Using Generic Routing Encapsulation Encapsulation", RFC 7637, IETF, September 2015.

Student Questions

- What's the main difference between NVGRE vs. VXLAN? Do they both provide the same functionality?

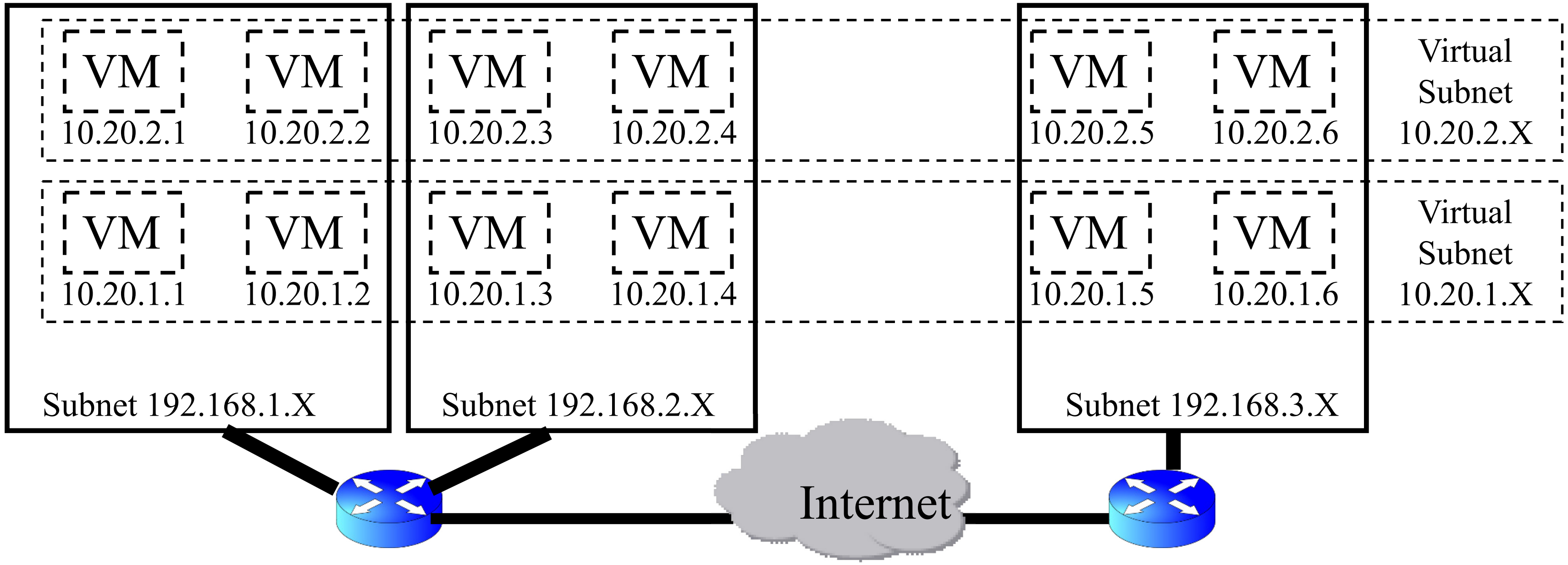
NVGRE is an older technology based on GRE. It has less functionality than VXLAN. NVGRE was designed for long-haul traffic. VXLAN is explicitly designed for VMs in data centers.

- Can you explain more about how this uses Ethernet to be L2?

Ethernet is always L2, regardless of how the packets are sent. This uses IP to send the packets.

NVGRE (Cont)

- In a cloud, a pSwitch or a vSwitch can serve as a tunnel endpoint
- VMs need to be in the same VSID to communicate
- VMs in different VSIDs can have the same MAC address
- Inner IEEE 802.1Q tag, if present, is removed.



Ref: Emulex, "NVGRE Overlay Networks: Enabling Network Scalability," Aug 2012, 11pp., http://www.emulex.com/artifacts/074d492d-9dfa-42bd-9583-69ca9e264bd3/elx_wp_all_nvgre.pdf

Student Questions

- Could you please explain this picture again? We can't see where the mouse is pointing.

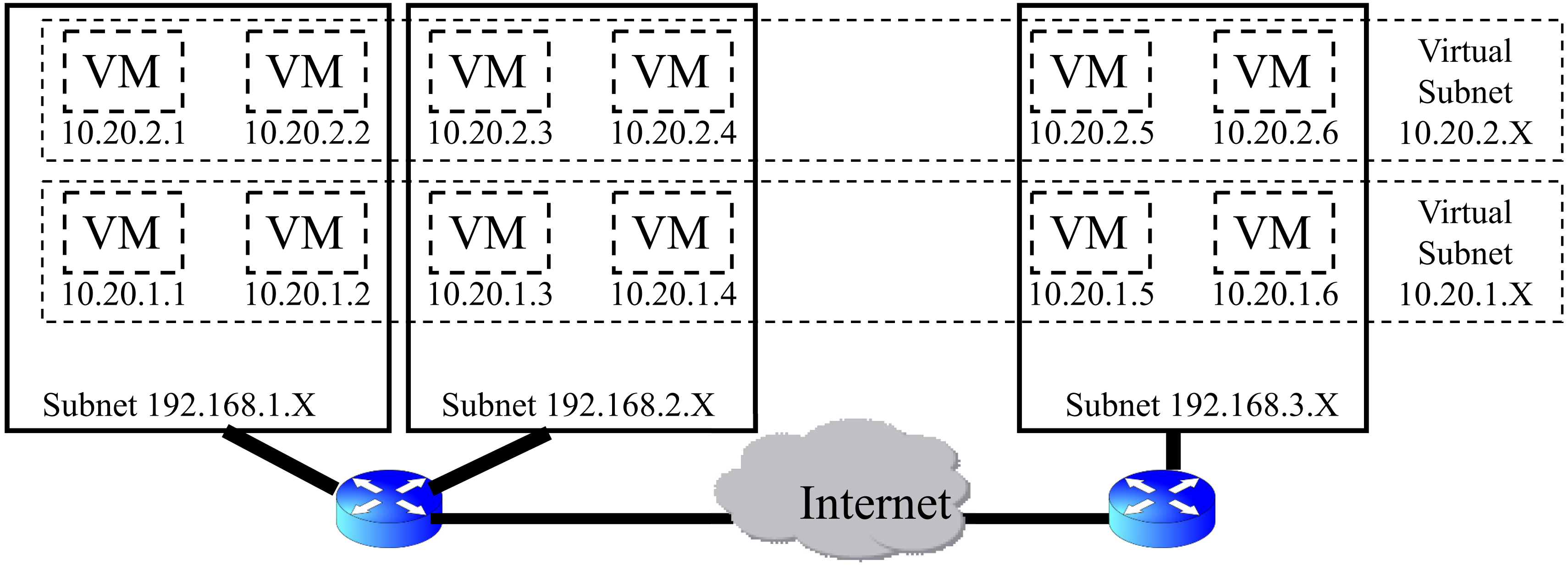
Two customers: A and B.
A = Virtual subnet 10.20.2.X
B = Virtual sunet 10.20.1.X
The provider has three physical subnets, 192.168.1.X, 192.168.2.X, and 192.168.3.X, at two locations connected via the Internet. Each can select their IP address spaces independent of the other.

- So, VMs could have the same MAC address in different VSIDs, but could they have the same IP address (in different VSIDs)?

Yes. IP addresses are private addresses and are assigned by the subnet.

NVGRE (Cont)

- In a cloud, a pSwitch or a vSwitch can serve as a tunnel endpoint
- VMs need to be in the same VSID to communicate
- VMs in different VSIDs can have the same MAC address
- Inner IEEE 802.1Q tag, if present, is removed.



Ref: Emulex, "NVGRE Overlay Networks: Enabling Network Scalability," Aug 2012, 11pp., http://www.emulex.com/artifacts/074d492d-9dfa-42bd-9583-69ca9e264bd3/elx_wp_all_nvgre.pdf

Student Questions

- If an application is on a virtual server (Like an EC2 instance on AWS), What will be its IP address resolution using its public domain name? Will the result be different from within the AWS private network or outside the AWS private network?

Outside public addresses remain the same but are translated inside the AWS to private addresses. Please read about network address translation (NAT).

NVO3

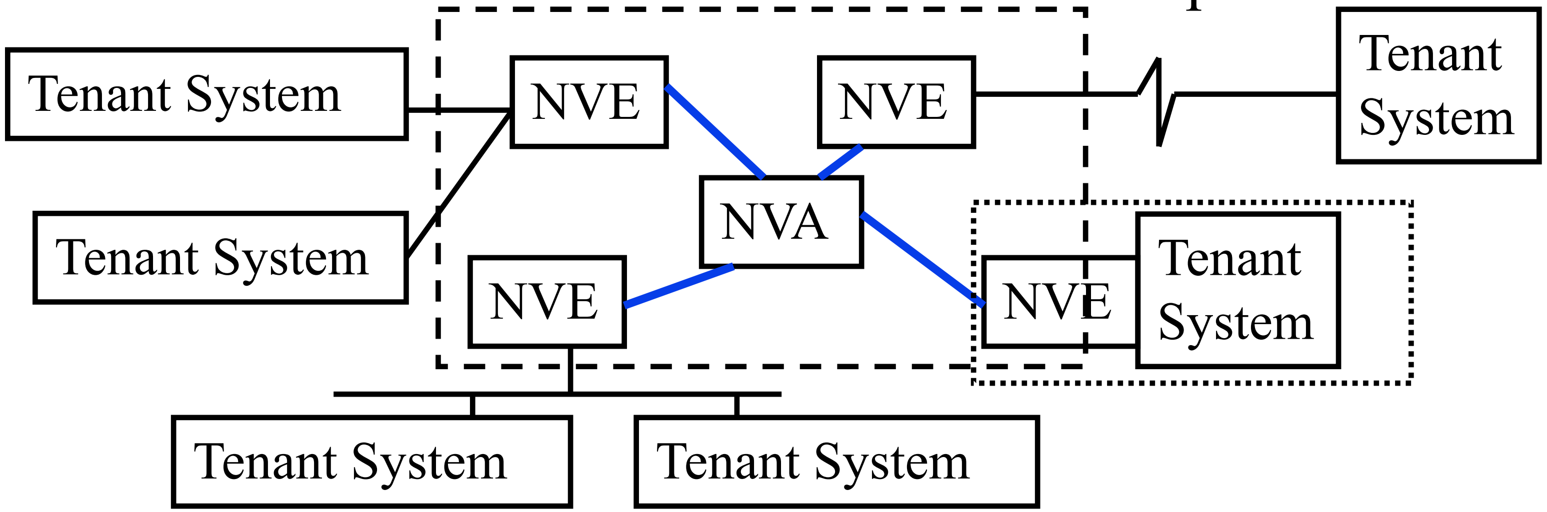
- Network Virtualization Overlays using L3 techniques
- **Problem**: Data Center Virtual Private Network (DCVPN) in a multi-tenant datacenter
- Issues:
  - Scale in Number of Networks: Hundreds of thousands of DCVPNs in a single administrative domain
  - Scale in Number of Nodes: Millions of VMs
  - VM (or pM) Migration
  - Support both L2 and L3 VPNs
  - Dynamic provisioning
  - Addressing independence: Each tenant should be able to select its address space
  - Virtual Private ⇒ Other tenants do not see your frames
  - Optimal Forwarding: VMs should not be tied to a single designated router that may be far away.

Ref: T. Narten, Ed., "Problem Statement: Overlays for Network Virtualization," IETF RFC 7364, Oct 14, 23 pp.

Student Questions

NVO3 Terminology

- **Tenant System (TS)**: VM or pM
- **Virtual Network (VN)**: L2 or L3 Tenant networks
- **Network Virtualization Edges (NVEs)**: Entities connecting TSs (virtual/physical switches/routers)
- NVEs could be in vSwitches, external pSwitches, or span both.
- **Network Virtualization Authority (NVA)**: Manages forwarding info for a set of NVEs
- NVA could be distributed or centralized and replicated.



Student Questions

- What would a distributed NVA look like? Since it is not centralized, no single entity manages the virtual network. Then, in a sense, NVA does not exist and is distributed in the edges who exchange information to act as authority.

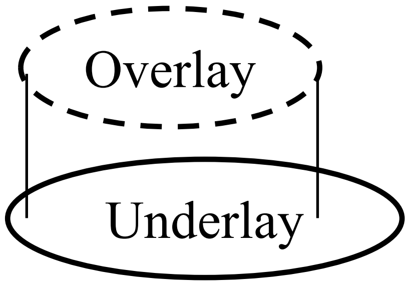
A distributed NVA can have a central controller.

- If separate tenants are connected to the same NVE, do they have to be a part of the same VN?

No. NVE tags them to keep them separate.

NVO3 Terminology (Cont)

- **Virtual Network (VN):** Provides L2/L3 services to a set of tenants
- **VN Context ID**: A field in the header that identifies a VN instance (VNI).
- **Overlay header** = inner header = Virtual Network Header
- **Underlay header** = outer header = Physical Network Header
- **Tenant Separation**: A tenant's traffic cannot be seen by another tenant



Student Questions

- NVGRE uses virtual subnets, and NVO3 uses virtual networks. What is the difference? If they are both virtual the customer/tenant would not notice a difference between subnet or network, correct?

Subnet = One LAN = One VLAN
There is no difference.

Current NVO Technologies

- BGP/MPLS IP VPNs: Widely deployed in enterprise networks. Difficult in data centers because hosts/hypervisors do not implement BGP.
- BGP/MPLS Ethernet VPNs: Deployed in carrier networks. Difficult in data centers.
- **802.1Q**, PB, PBB VLANs
- **Shortest Path Bridging**: IEEE 802.1aq
- Virtual Station Interface (VSI) Discovery and Configuration Protocol (VDP): IEEE 802.1Qbg
- Address Resolution for Massive numbers of hosts in the Data Center (ARMD): RFC6820
- **TRILL**
- **L2VPN**: Provider provisioned L2 VPN
- Proxy Mobile IP: Does not support multi-tenancy
- LISP: RFC 6830

Student Questions

VXLAN

- Virtual eXtensible Local Area Networks (VXLAN)
- L3 solution to isolate multiple tenants in a data center (L2 solution is Q-in-Q and MAC-in-MAC)
- Developed by VMware. Supported by many companies in the IETF NVO3 working group
- Problem:
  - 4096 VLANs are not sufficient in a multi-tenant data center
  - Tenants need to control their MAC, VLAN, and IP address assignments ⇒ Overlapping MAC, VLAN, and IP addresses
  - Spanning tree is inefficient with a large number of switches ⇒ Too many links are disabled
  - Better throughput with IP equal cost multipath (ECMP)

Ref: M. Mahalingam, D. G. Dutt, et al. "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks," IETF RFC 7348, August 2014.

Student Questions

- What is the difference between the transmission methods of VXLAN and NVGRE?

See the answer on Slide 8-14.

- What are the limitations of VXLAN? Since it covers the scaling limitation of VLAN.

Overhead.

- Why is a spanning tree inefficient with a large number?

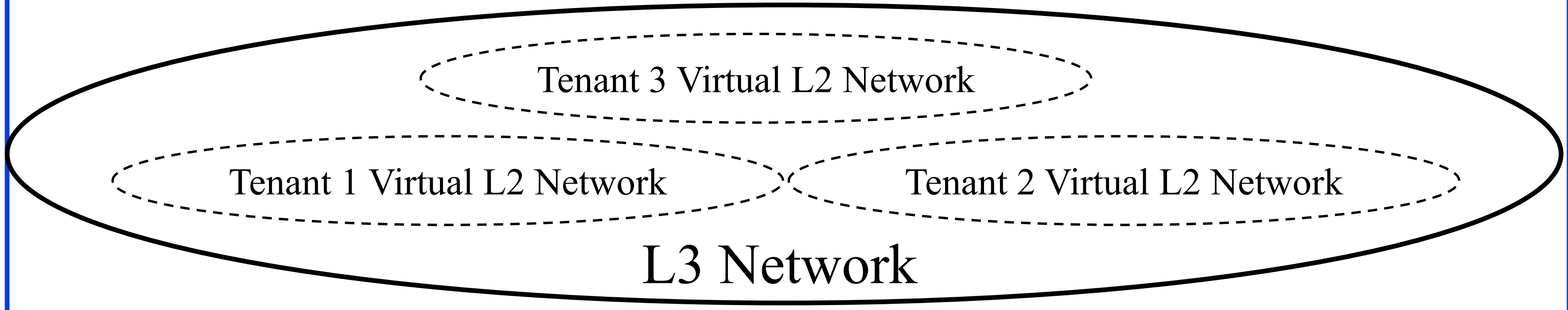
A large number of nodes have to agree to a single root. It will take a long time and too many messages.

- If the Spanning tree is not used, then what is used? IS-IS?

IP equal cost multipath (ECMP) described in RFC 2992 is from IETF. IS-IS is an OSI standard. They are different.

VXLAN Architecture

- Create a virtual L2 overlay (called VXLAN) over L3 networks
- 2^{24} Virtual Network Instances (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).
- Segments may have overlapping MAC addresses and VLANs, but L2 traffic never crosses a VNI



Student Questions

- Can you please explain the function of VTEP in point 4? Is it just the name of the end points of the VXLAN tunnel?

It is both name and function. See the next slide.

- VXLAN is L2 over L4 or L2 over L3?

You could call it L2 over L4 since there is a thin L4 (UDP).

- So, one LAN could have multiple VNIs?

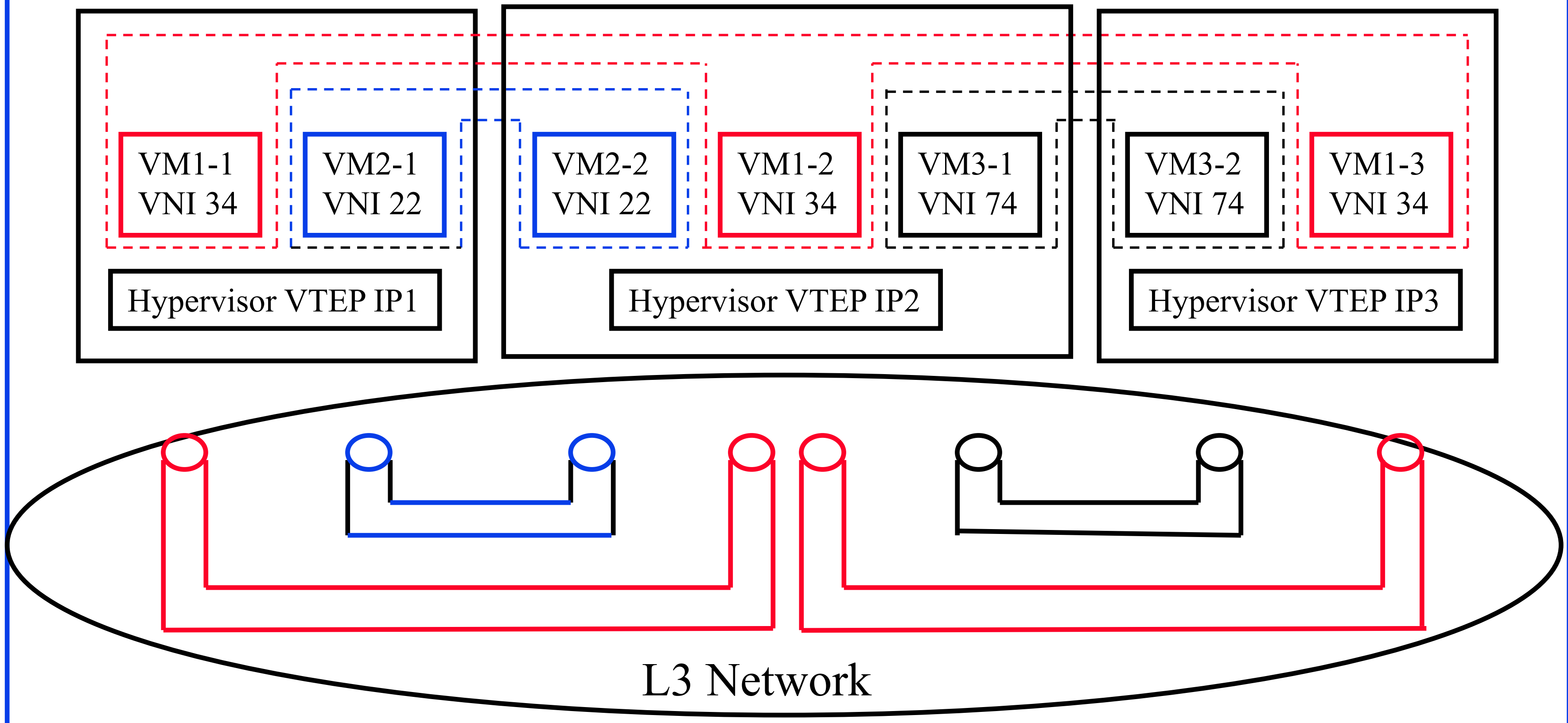
Yes, one LAN can have multiple VLANs. But here, multiple LANs combined can have multiple shared VLANs.

- To confirm VXLAN is in the L2 layer?

VXLAN creates a virtual L2 network using L3 layer protocols.

VXLAN Deployment Example

Example: Three tenants. 3 VNIs. 4 Tunnels for unicast.
+ 3 tunnels for multicast (not shown)



Student Questions

- Can you illustrate this with the laser pointer, please? Would the red multicast be the combination of the two red unicast tunnels?

Yes. RED VLAN uses two red tunnels.

- Can you explain this slide again? *Sure.*
- Could you explain '4 tunnels for unicast + 3 tunnels for multicast' in more detail?

The Red Multicast tunnel consists of two unicast tunnels.

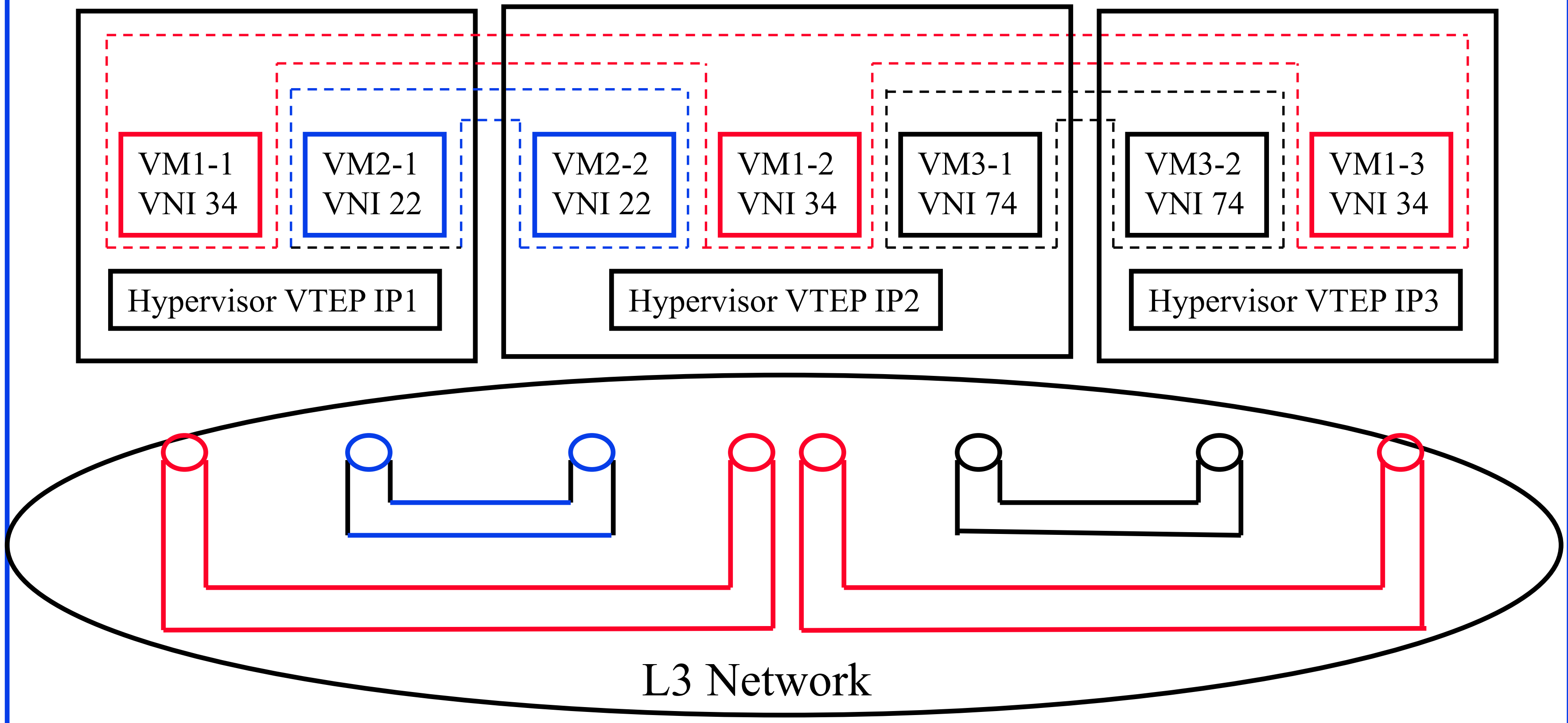
- Are there any connections or differences between Virtual Subnet and VNI?

VNI is the ID of the Virtual subnet. A numeric ID identifies each subnet.

- Could you explain the structure of VXLAN in this example in detail, please? *Sure.*

VXLAN Deployment Example

Example: Three tenants. 3 VNIs. 4 Tunnels for unicast.
+ 3 tunnels for multicast (not shown)



Student Questions

- How would the network change if VMs were aware of if they were operating on VLAN or XVLAN?

VMs can only do something different with that knowledge. Not knowing makes it easier to package VMs.

- If VXLAN is implemented within a single data center, why aren't tenant VMs co-located on the same hypervisor?

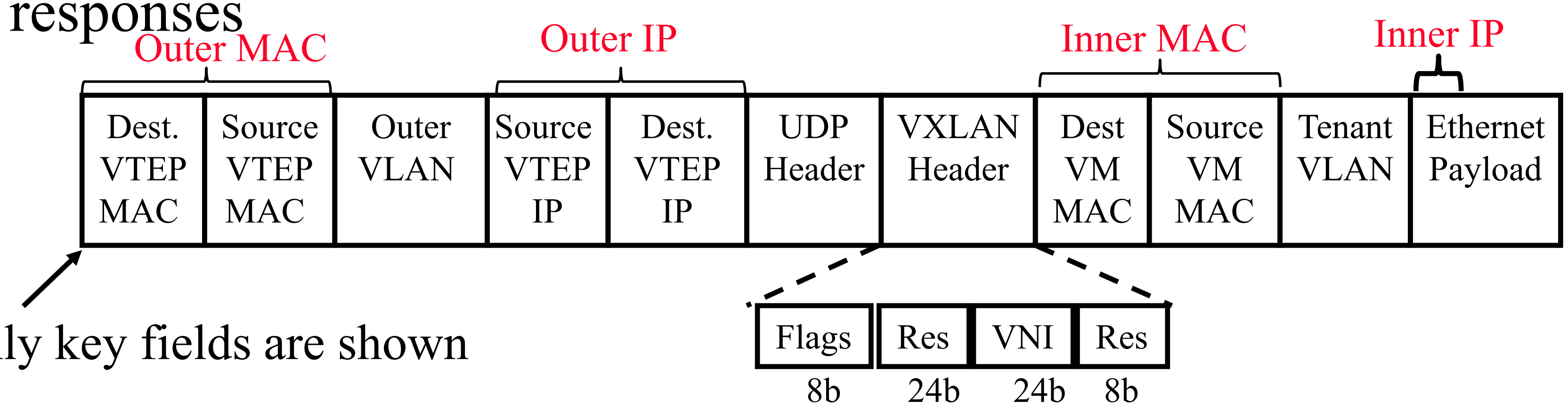
VMs are allocated on demand. The hypervisor may have already used up its capacity.

- Which part of the graph is vSwitch?

It is built-in inside the Hypervisor.

VXLAN Encapsulation Format

- Outer VLAN tag is optional.
  Used to isolate VXLAN traffic on the LAN
- Source VM ARPs to find the Destination VM's MAC address.
  All L2 multicasts/unknowns are sent via IP multicast.
  Destination VM sends a standard Ethernet ARP response.
- Destination VTEP learns Inner-Src-MAC-to-Outer-Src-IP mapping ⇒ Avoids unknown destination flooding for returning responses



Only key fields are shown

Student Questions

- Can you point out which is inner-src-MAC and outer-src-IP?

See updates in red on the left.

- "Destination VM sends a standard IP unicast ARP response." We know the destination VM does not know it is in a virtualized domain. So, it should send a normal ARP packet, which is not carried normally over IP. I think the VTEP will add the IP and the rest of the VXLAN header. Is my understanding correct?

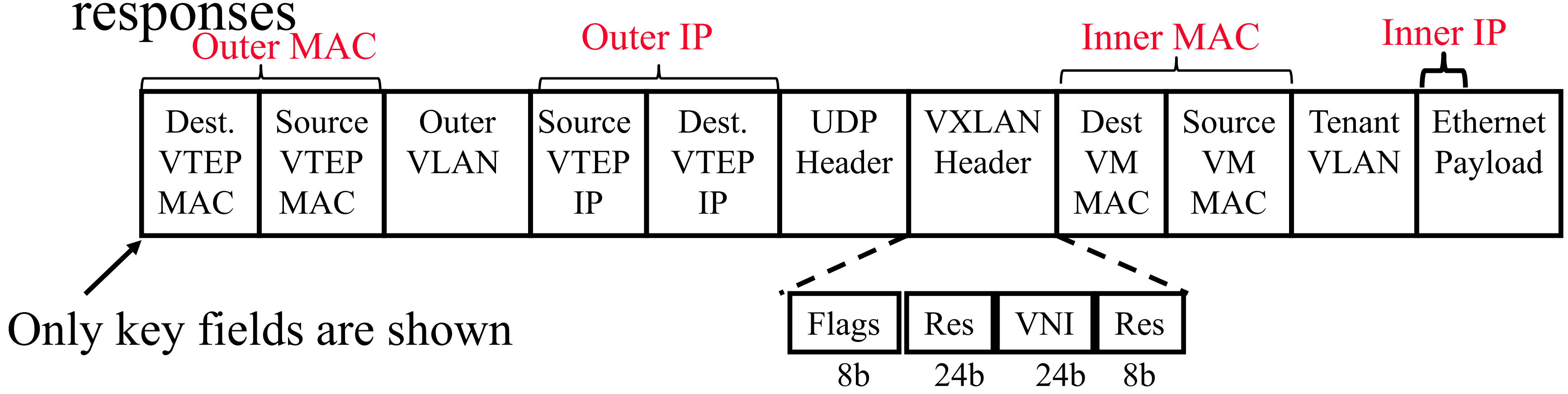
Your understanding is correct. The 2nd point on the slide has been corrected to reflect this.

- Does this require much memory to store as the endpoint?

All endpoints are required to store Ethernet-to-IP translation even before all this virtualization.

VXLAN Encapsulation Format

- Outer VLAN tag is optional.
  Used to isolate VXLAN traffic on the LAN
- Source VM ARPs to find the Destination VM's MAC address.
  All L2 multicasts/unknowns are sent via IP multicast.
  Destination VM sends a standard Ethernet ARP response.
- Destination VTEP learns Inner-Src-MAC-to-Outer-Src-IP mapping ⇒ Avoids unknown destination flooding for returning responses



Student Questions

- So, is it right to think that VTEPs cover L2 and L3?

Yes. Anything that covers both L2 and L3 is called L3. So, VTEPs are L3.

VXLAN Encapsulation Format (Cont)

- Internet Group Multicast Protocol (IGMP) is used to prune multicast trees
- 7 of 8 bits in the flag field are reserved.
  One flag bit is set if the VNI field is valid
- 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
- VTEPs need to learn the MAC address of other VTEPs and client VMs of VNIs they are handling.
- A VXLAN gateway switch can forward traffic to/from non-VXLAN networks. Encapsulates or decapsulates the packets.

Student Questions

- You stated that ECMP is allowed on point-to-point tunnels. What is the benefit of using ECMP in this case? Because we have no control/information over the carrier provider network. So, why do we use multiple paths?

Multiple paths inside the data center.

- How large is the hash, and how many unique source ports could VXLAN support?

All 2^{16} ports can be supported. Hashing does not require storage. It is only a computation.

- Could you express why hashing the UDP source port benefits the load balance again, please?

UDP source port identifies the flows. Hashing helps keep the flows from taking different paths.

VXLAN: Summary

- VXLAN solves the problem of multiple tenants with overlapping MAC addresses, VLANs, and IP addresses in a cloud environment.
- A server may have VMs belonging to different tenants
- No changes to VMs. Hypervisors are responsible for all details.
- Uses UDP over IP encapsulation to isolate tenants

Student Questions

End of Part 2

Stateless Transport Tunneling Protocol (STT)

- Ethernet over **TCP-Like** over IP tunnels.
  GRE and IPSec tunnels can also be used if required.
- Designed for large storage blocks **64kB**. Fragmentation allowed.
- Most other overlay protocols use UDP and disallow fragmentation ⇒ Maximum Transmission Unit (MTU) issues.
- 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.
  New work on Geneve.

Ref: B. Davie and J. Gross, "A Stateless Transport Tunneling Protocol for Network Virtualization (STT)," Sep 2013, http://tools.ietf.org/html/draft-davie-stt-04

Student Questions

- Can you please provide us with a reference to learn more about the "TCP-Like" protocol?

TCP-Like = Transport protocols with reliable/in-order packet delivery, e.g., Stream Control Transmission Protocol (SCTP), Datagram Congestion Control Protocol (DCCP), Reliable User Datagram Protocol (RUDP), Scalable TCP (STCP), Transactional TCP (T/TCP), Reliable Datagram Sockets (RDS), RDMA over Converged Ethernet (RoCE)
UDP-Like = Transport protocols with unreliable/out-of-order packet delivery.
Ref: https://en.wikipedia.org/wiki/Transport_layer

- Why was STT modeled after the TCP header if it doesn't use most of TCP's features?

Often, it helps to copy only the good parts of an existing scheme.

Geneve

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

Ref: J. Gross, et al, "Geneve: Generic Network Virtualization Encapsulation" IETF Internet Draft, draft-ietf-nvo3-geneve-14, Sep. 12, 2019, https://tools.ietf.org/pdf/draft-ietf-nvo3-geneve-14.txt

Student Questions

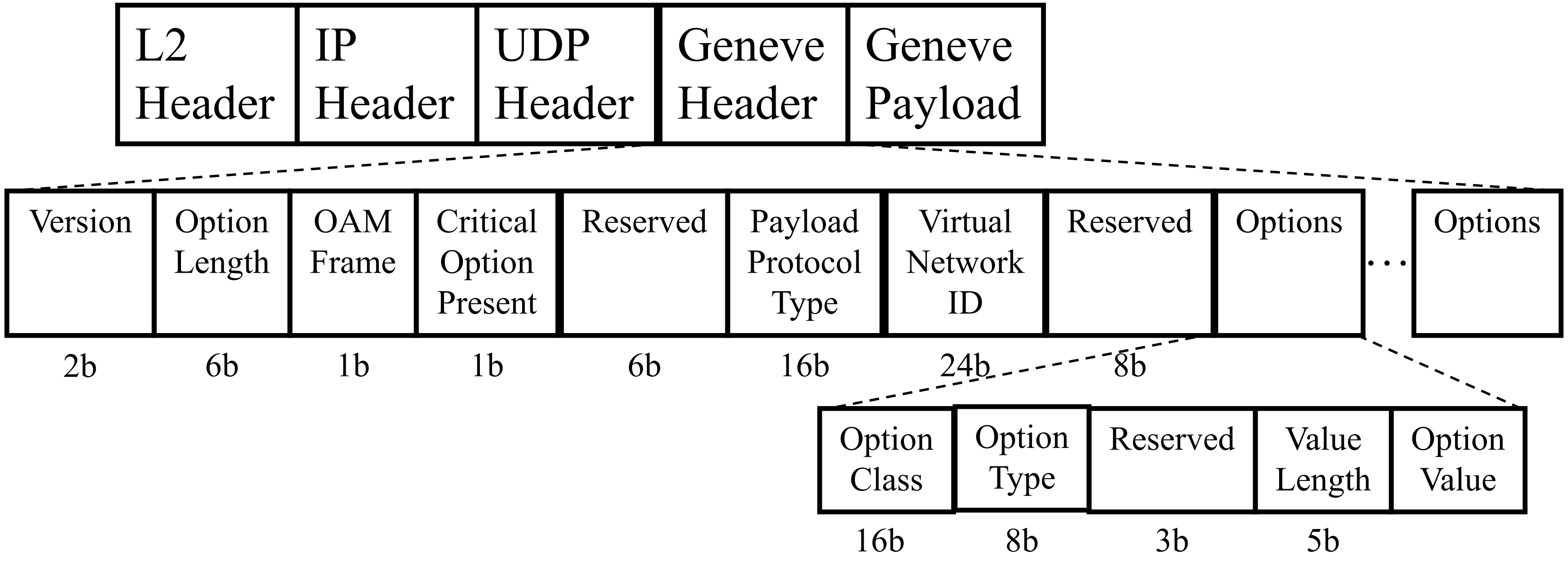
- How does Geneve differ from other tunneling protocols, such as VXLAN and GRE?

See the list of features on the slide. 1. Generic, 2. Affects only endpoints, 3. No modifications at intermediate points.

Next Q on 28

Geneve Frame Format

- **Highly Extensible**: Variable number of variable size options
- Any vendor can extend it in its way by getting an "Option Class" from IANA (Internet Assigned Number Authority)
- Options are encoded in a **TLV** (Type-Length-Value) format



Student Questions

- What are all the reserved bits for?

Future extensions.

- Will Geneve be popular in the future? *Maybe.*
- What is the IANA, and does it have denial powers?

Internet Assigned Numbers Authority assigns and registers all Internet Parameter numbers. Used to be "John Postel," but now by an affiliate of the Internet Corporation for Assigned Names and Numbers (ICANN).

Ref:https://en.wikipedia.org/wiki/Internet_Assigned_Numbers_Authority

- In the Geneve Frame Format, we can find two reserved areas; I wonder if there are any rules for the reserved areas (where to place them and how many bits of them)?

1. For Byte alignment

2. Expect to extend a particular field

Geneve Frame Format (Cont)

- **Option Length** (6 bits): Length of options field in 4B (does not include the rest of the Geneve header)
- **OAM Frame** (1 bit): Control packet. Does not contain user data. Must be passed on to the control plane CPU
- **Critical Options Present** (1 bit): One or more options are critical.
  Drop the packet if you don't understand a critical option
- **Payload Protocol Type** (16 bits): 0x6558 for Ethernet
- **Virtual Network ID** (24 bits): Tenant ID
- **Option Class** (16 bits): Who designed this option? Vendor, technologies, organizations, …
- **Option Type** (8 bits) : msb (most significant bit) =1 => Critical
- **Option Value Length** (5 bits): in units of 4 bytes

Student Questions

- The option length field of 6 bits is indicated to be wrong in the video, so what's the correct value for it? *Six is correct.*
- What's the difference between option length and the length of options?

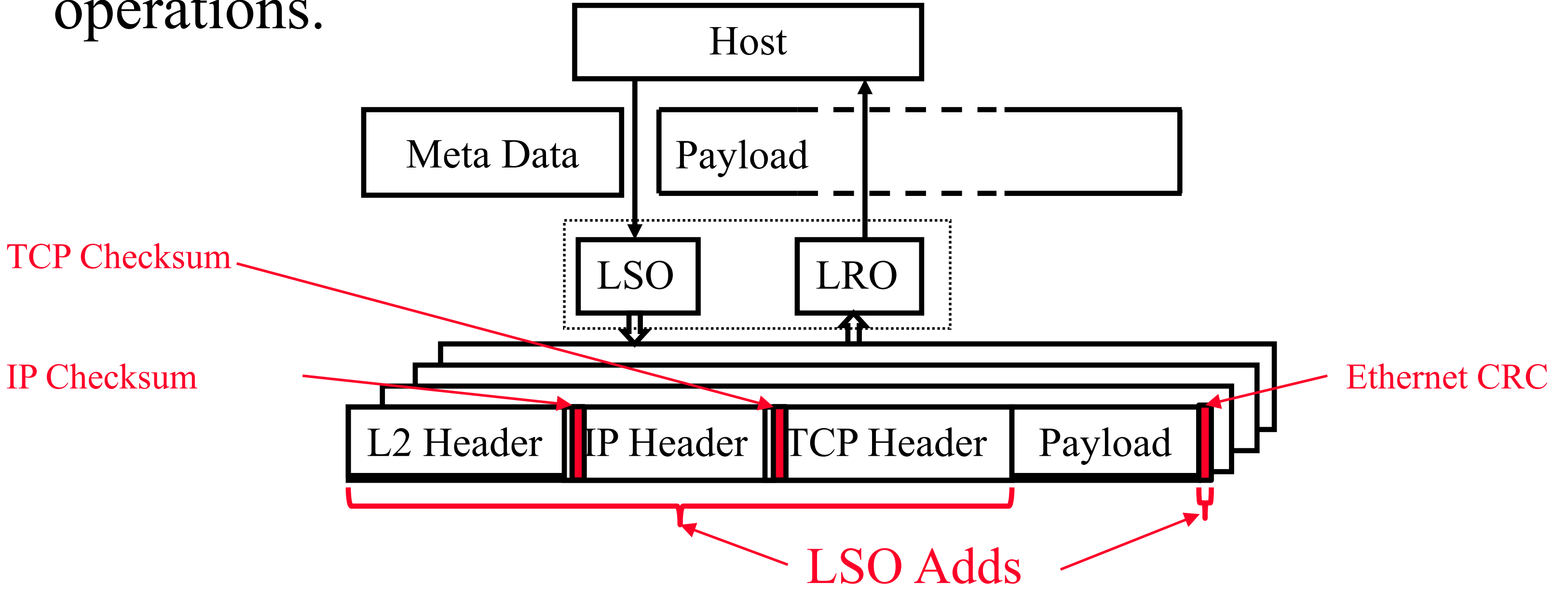
If there are many options, each option may have a different length. In the TLV-encoded option, option length is encoded in the L field.

- Knowing that a packet will be dropped if the critical options are not understood, is the critical option sometimes used as a filter?

No. Some information is necessary for processing a packet. If that information is not there, we cannot do much. We drop it.

LSO and LRO

- **Large Send Offload** (LSO): The host hands a large chunk of data to NIC and metadata. NIC makes MSS size segments and adds checksum, TCP, IP, and MAC headers to each segment.
- **Large Receive Offload** (LRO): NICs attempt to reassemble multiple TCP segments and pass larger chunks to the host. The host does the final reassembly with fewer per-packet operations.



Student Questions

- How much is the MSS size？

Host Administrator can set MSS size. Minimum, 536 bytes, but it can be more significant. MTU=576B ⇒ MSS=576-20-20

*Ref:*https://en.wikipedia.org/wiki/Maximum_segment_size

- Is the checksum for LSO or LSO and more?

LSO takes the TCP payload and adds all headers and trailers, as shown in the updated slide.

- "I thought of NICs as a type of I/O device. When you mentioned that 'every I/O device reads data at 64 kilobytes/s,' were you referring to block devices like disks?"

64 kB is the size of a disk read. NICs are also I/O, but they used only 1500-byte packets before.

Geneve Implementation Issues

- **Fragmentation**: Use Path MTU (Maximum Transmission Unit) discovery to avoid fragmentation on the path
- **DSCP** (Differentiated Services Control Point): DSCP bits in the outer header may or may not be the same as in the inner header. Decided by the policy of the network service provider
- **ECN** (Explicit Congestion Notification): ECN bits should be copied from the inner header on entry to the tunnel and copied back to the inner header on exit from the tunnel
- **Broadcast and Multicast**: Use underlying networks' multicast capabilities if available. Use multiple point-to-point tunnels if multicast is not available.

Student Questions

- DSCP differs (only) when customers request services they were not entitled to have (paid for). Is this correct?

Some services may be too much work to bill and may be included if requested, such as audio transmission, video transmission, and data transmission.

- How can a P2P tunnel replace the multicast capability?

Ethernet is both P2P and multicast. IP can also provide multicast even though the packets mostly go over point-to-point links.

Geneve Implementation Issues (Cont)

- **LSO**: Replicate all Geneve headers and options on all outgoing packets.
- **LRO**: Merge all packets with the identical Geneve headers
- **Option Order**: Not significant. Options can be in any order.
- **Inner VLAN**: Tunnel endpoints decide whether to differentiate packets with different inner VLAN values.

Student Questions

- Can you be more specific about why the software would have issues performing jobs done by LSO and LRO? e.g., adding checksums, TCP, IP, and MAC headers to each segment. Why can't software do this?

The software would be slow. Specially designed hardware would be much faster.

- Does the "inner" in "Inner VLAN" means that it's a VLAN in a bigger VLAN or is in a bigger Physical LAN?

Here inner VLAN is for inner VLAN header. The packet may have several headers.

Outer	Inner

Geneve Implementation Issues (Cont)

- **LSO**: Replicate all Geneve headers and options on all outgoing packets.
- **LRO**: Merge all packets with the identical Geneve headers
- **Option Order**: Not significant. Options can be in any order.
- **Inner VLAN**: Tunnel endpoints decide whether to differentiate packets with different inner VLAN values.

Student Questions

- Since LSO and LRO apply headers to each packet "chunk," can packets be sent out of order?

No. Packets are still sent in order. Sometimes, they may get out-of-order on the path as usual.

Geneve Summary

1. UDP over IP encapsulation
2. Geneve header is extensible by vendors
3. Generally, variable length headers are considered hard for hardware implementation
4. Vendor extensibility requires a system to register options and may result in interoperability issues
5. This is subject to change since it is in the draft stage.

Student Questions

- Can you provide some examples of interoperability issues that would happen?

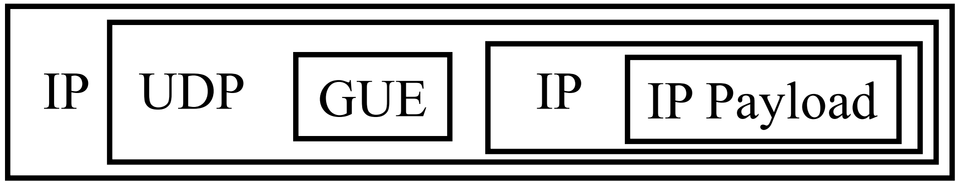
One vendor extends one way, and the other extends the other way. They may not inter-operate. For, one may have 64-bit CRC; another may be unable to process that.

- Why would vendors registering options result in interoperability issues?

Vendor registering reduces interoperability since others know what this option means.

Generic UDP Encapsulation (GUE)

- Using UDP to encapsulate IP protocols
- Allows the use of efficient hardware implementations of UDP
- Generic $\Rightarrow$ Any IP payload
- Optional data in the header: VNI, Authentication, Security, congestion control, vendor extensions, etc.
- Allows Non-TCP/Non-UDP IP payloads over networks that filter all non-TCP/non-UDP packets.



Ref: T. Herbert, L. Yong, and O. Zia, “Generic UDP Encapsulation,” IETF Draft, March 7, 2019, draft-ietf-intarea-gue-07

Student Questions

- Why is the consensus UDP for these tunneling protocols? Hardware is more efficient for UDP, but then why not use UDP always?

UDP does not do many functions, so it is simpler/faster, but if those functions are needed, TCP has to be used, such as lost packet retransmission.

- What's the difference between GUE and GRE?

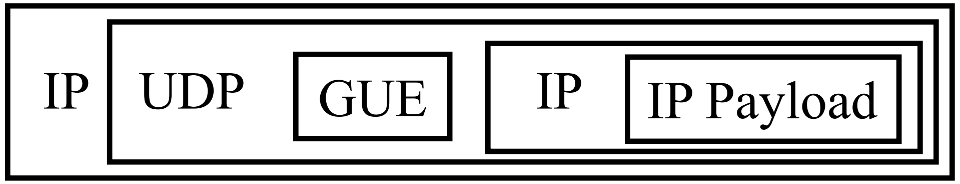
GRE runs on IP. GUE uses UDP\ to provide protocol multiplexing and optional checksums.

- Wouldn't it be a security risk if GUE can pass the non-TCP/non-UDP filter?

Sometimes, it is necessary to use non-TCP/non-UDP payloads. The recipients can filter those out if needed.

Generic UDP Encapsulation (GUE)

- Using UDP to encapsulate IP protocols
- Allows the use of efficient hardware implementations of UDP
- Generic ⇒ Any IP payload
- Optional data in the header: VNI, Authentication, Security, congestion control, vendor extensions, etc.
- Allows Non-TCP/Non-UDP IP payloads over networks that filter all non-TCP/non-UDP packets.



Ref: T. Herbert, L. Yong, and O. Zia, "Generic UDP Encapsulation," IETF Draft, March 7, 2019, draft-ietf-intarea-gue-07

Student Questions

- Is it correct to say that GUE can pass the filter because it "Fakes" the packet as a UDP packet?

Yes. It only works at the firewall. The receiver can throw the packet if they were not expecting it.

GUE Packet Format

- Source Port: 6080 or any other port
- Dest Port: 6080 or any other port
- Len: Length of UDP header and payload in 4B unit
- Checksum: Standard UDP checksum
- Versions 0 and 1 are standardized. Version 0 is shown below.
- Hlen: Extension Field + Private Data length in 32-bit words
  Total GUE header length = HLEN*4+4 bytes
- Protocol/Control Type: C=1 ⇒ Control, C=0 ⇒ Protocol type of the payload
- Flags indicate the presence of various extension fields and private data
- Version 1: No Extension or private data. The first 4 bits of the IP header are version 0100 or 0110. The first 2 bits indicate the GUE version.

Source Port	Dest Port	Len	Check sum	Ver	C	Hlen	Proto/ Ctype	Flags	Extension Fields (Opt)	Private Data (Opt)
16b	16b	16b	16b	2b	1b	5b	8b	16b		

Student Questions

- Can you please explain Hlen? and why we multiply HLEN by 4?

Hlen= Header Length in words.

Each word is 4 bytes.

- Since the version only has two kinds, 00 and 01, why not just use one bit?

In the future, we may have more versions.

- How can GUE help virtualization? How can we encapsulate VNI and other information related to a virtualized domain into the GUE header? It does not have any field for them, unlike Geneve or VXLAN.

It is much older than VXLAN, which is older than Geneve.

- Since there are only version 0 and 1, why it needs two bits?

For future versions 2 and 3.

GUE Packet Format

- Source Port: 6080 or any other port
- Dest Port: 6080 or any other port
- Len: Length of UDP header and payload in 4B unit
- Checksum: Standard UDP checksum
- Versions 0 and 1 are standardized. Version 0 is shown below.
- Hlen: Extension Field + Private Data length in 32-bit words
  Total GUE header length = HLEN*4+4 bytes
- Protocol/Control Type: C=1 ⇒ Control, C=0 ⇒ Protocol type of the payload
- Flags indicate the presence of various extension fields and private data
- Version 1: No Extension or private data. The first 4 bits of the IP header are version 0100 or 0110. The first 2 bits indicate the GUE version.

Source Port	Dest Port	Len	Check sum	Ver	C	Hlen	Proto/ Ctype	Flags	Extension Fields (Opt)	Private Data (Opt)
16b	16b	16b	16b	2b	1b	5b	8b	16b		

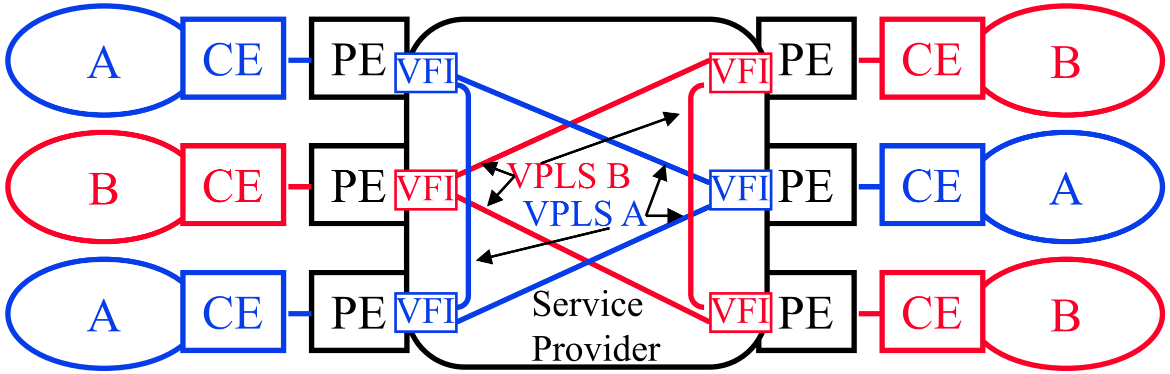
Student Questions

- Which point is the IP address starting when the version is 1? From the version part?

The dotted part on the right.

Virtual Private LAN Service (VPLS)

- Allows *multi-point* Ethernet services over MPLS networks using a *full-mesh point-to-point pseudo-wires*
- **Virtual Forwarding Instance (VFI)**: A virtual LER instance in the provider edge router specific to each customer LAN
- **VPLS Instance**: Set of VFIs and PWs connecting them. Creates a single "VLAN" broadcast domain connecting VFIs.
- Widely deployed but does not meet data center requirements ⇒ BGP MPLS-Based Ethernet VPN (EVPN)



CE: Customer Edge
PE: Provider Edge

Ref: G. Santana, "Datacenter Virtualization Fundamentals," Cisco Press, 2014, ISBN: 1587143240
K. Kompella and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling," IETF RFC 4761, Jan 2007, https://tools.ietf.org/pdf/rfc4761

Student Questions

- If VPLS does not meet data center requirements, why would it be widely deployed?

It is used outside the data center by carriers.

- If VPLS does not meet DCN requirements, where is it typically used?

See above.

- Are the VPLS instances tunnels made of pseudowires?

Yes.

EVPN Features

- Multicast Optimization: MP2MP LSPs
- Ease of Provisioning: Auto-discovery of PEs, Customer Site ID, Automated designated Forwarder election among PEs, MPLS parameters based on VLAN parameters
- New Service Interface: Port = VLAN, Multiple VLANs per port, VLAN bundles are treated as one VLAN
- Fast Convergence
- Flood Suppression
- Flexible VPN topologies and Policies

Student Questions

- You mention VPLS is a natural transition to EVPN. Can you explain why?

EVPN is on Ethernet and offers multicast. VPLS is a similar service to MPLS.

- I don't see any references to VPLS on the EVPN slides.

Ref: https://en.wikipedia.org/wiki/Virtual_Private_LAN_Service

- Is EVPN VPLS with added automation?

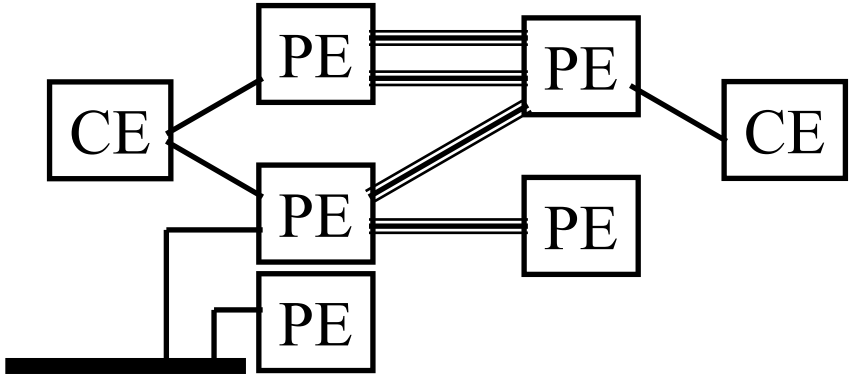
No. EVPN does not require MPLS. VPLS requires MPLS.

- How does EVPN handle multiple VLANs per port?

It reads the packet headers and allows multiple VLANs per port.

EVPN Redundancy Features

- Flow-based load balancing: CE connected to multiple PEs. Select path by hashing 7-tuple (L2, L3, L4 addresses, VLAN)
- Flow-based multi-pathing: Multiple LSPs between PEs
- Geo-Redundant PE Nodes: CE connected to multiple PEs in different PoPs
- Optimal Traffic Forwarding: Single/multi-homed CE to Single/multi-homed CE. Packet not forwarded to PEs not connected to the destination CE
- Flexible Redundancy Grouping: PEs are grouped for redundancy.
- Multi-homed Network: Entire Ethernet is connected to multiple PEs.



Ref: J. Rabadan, et al., "Applicability of EVPN to NVO3 Networks," IETF Draft, July 8, 2019, https://tools.ietf.org/html/draft-ietf-nvo3-evpn-applicability
A Sajassi, et al., "BGP MPLS-Based Ethernet VPN," IETF RFC 7432, Feb 2015, https://tools.ietf.org/html/rfc7432
A. Sajassi, et al., "Requirements for Ethernet VPN (EVPN)," May 2014, https://tools.ietf.org/html/rfc7209

Student Questions

- Can you explain the Geo-Redundant PE Nodes again?

Assume that the top left PE in the figure is in Kansas City and the middle left PE is in St. Louis. They would be Geo-redundant.

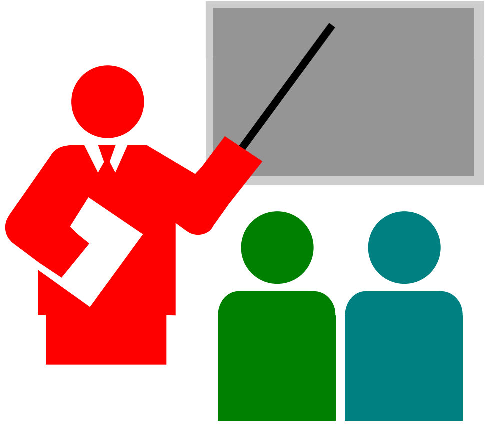
- What is Pops?

PoP = Point of Presence = Carrier Building

- Since EVPN is better, have carriers switched to it from VPLS?

Yes.

Summary



1. TRILL uses "Routing Bridges" to transport Ethernet packets on a campus network. RBs use IS-IS to find the shortest path.
2. NVO3 is a generalized framework for network virtualization and partitioning for multiple tenants over L3. It covers both L2 and L3 connectivity.
3. NVGRE uses Ethernet over GRE for L2 connectivity.
4. VXLAN uses Ethernet over UDP over IP
5. Geneve uses Any protocol over UDP over IP encapsulation.

Student Questions

References

- Emulex, "NVGRE Overlay Networks: Enabling Network Scalability," Aug 2012, 11pp., http://www.emulex.com/artifacts/074d492d-9dfa-42bd-958369ca9e264bd3/elx_wp_all_nvgre.pdf
- G. Santana, "Datacenter Virtualization Fundamentals," Cisco Press, 2014, ISBN: 1587143240 (Safari Book)
- J. Gross, et al, "Geneve: Generic Network Virtualization Encapsulation" IETF Internet Draft, Sep. 12, 2019, https://tools.ietf.org/pdf/draft-ietf-nvo3-geneve-14.txt
- M. Lasserre, et al., "Framework for Data Center (DC) Network Virtualization," IETF RFC 7365, Oct 2014, 26 pp., https://tools.ietf.org/pdf/rfc7365
- M. Mahalingam, et al, "*VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks*," IETF RFC 7348, https://tools.ietf.org/pdf/rfc7348
- P. Garg, Y. Wang, "NVGRE: Network Virtualization using GRE," Sep 2015, IETF RFC 7637, https://tools.ietf.org/pdf/rfc7637

Student Questions

References (Cont)

- R. Perlman, "RBridges: Transparent Routing," Infocom 2004
- TRILL RFCs 5556, 6325, 6326, 6327, 6361, 6439
- T. Narten, Ed., "Problem Statement: Overlays for Network Virtualization," IETF RFC 7364, Oct 14, 23 pp., https://tools.ietf.org/html/rfc7364

Student Questions

Wikipedia Links

- http://en.wikipedia.org/wiki/Generic_Routing_Encapsulation
- http://en.wikipedia.org/wiki/Locator/Identifier_Separation_Protocol
- http://en.wikipedia.org/wiki/Large_segment_offload
- http://en.wikipedia.org/wiki/Large_receive_offload
- https://en.wikipedia.org/wiki/Virtual_Extensible_LAN

Student Questions

Acronyms

- ARMD Address Resolution for Massive numbers of hosts in the Data center
- ARP Address Resolution Protocol
- BGP Border Gateway Protocol
- BUM Broadcast, Unknown, Multicast
- CPU Central Processing Unit
- DC Data Center
- DCI Data Center Interconnection
- DCN Data Center Networks
- DCVPN Data Center Virtual Private Network
- DSCP Differentiated Services Control Point
- ECMP Equal Cost Multi Path
- EoMPLSoGRE Ethernet over MPLS over GRE
- ECN Explicit Congestion Notification
- EVPN Ethernet Virtual Private Network
- GRE Generic Routing Encapsulation

Student Questions

Acronyms (Cont)

- IANA Internet Address and Naming Authority
- ID Identifier
- IEEE Institution of Electrical and Electronic Engineers
- IETF Internet Engineering Task Force
- IGMP Internet Group Multicast Protocol
- IP Internet Protocol
- IPSec IP Security
- IPv4 Internet Protocol V4
- IS-IS Intermediate System to Intermediate System
- LAN Local Area Network
- LISP Locator ID Separation Protocol
- LRO Large Receive Offload
- LSO Large Send Offload
- MAC Media Access Control
- MPLS Multi Protocol Label Switching
- MSS Maximum Segment Size

Student Questions

Acronyms (Cont)

- MTU — Maximum Transmission Unit
- NIC — Network Interface Card
- NV — Network Virtualization
- NVA — Network Virtualization Authority
- NVEs — Network Virtualization Edge
- NVGRE — Network Virtualization Using GRE
- NVO3 — Network Virtualization over L3
- OAM — Operation, Administration and Management
- OTV — Overlay Transport Virtualization
- PB — Provider Bridges
- PBB — Provider Backbone Bridge
- pM — Physical Machine
- pSwitch — Physical Switch
- QoS — Quality of Service
- RB — Routing Bridge
- RFC — Request for Comment

Student Questions

Acronyms (Cont)

- RS Routing System
- STT Stateless Transport Tunneling Protocol
- TCP Transmission Control Protocol
- TLV Type-Length-Value
- TRILL Transparent Routing over Lots of Links
- TS Tenant System
- UDP User Datagram Protocol
- VDP VSI Discovery and Configuration Protocol
- VLAN Virtual Local Area Network
- VM Virtual Machine
- VN Virtual Network
- VNI Virtual Network Instance/Virtual Network Context ID
- VPLS Virtual Private LAN Service
- VPLSoGRE Virtual Private LAN Service over GRE
- VPN Virtual Private Network

Student Questions

Acronyms (Cont)

VRRP	Virtual Router Redundancy Protocol
VSI	Virtual Station Interface
VSID	Virtual Subnet Identifier
vSwitch	Virtual Switch
VTEP	VXLAN Tunnel End Point
VXLAN	Virtual Extensible Local Area Network

Student Questions

Scan This to Download These Slides

Raj Jain
http://rajjain.com

http://www.cse.wustl.edu/~jain/cse570-23/m_08dmt.htm

Student Questions

- We have gone through lots of similar technologies in terms of their purpose. (NVGRE, TRILL, etc.) Can we have a summary and comparison to figure out who in what condition uses which technology to do what? *Good Idea. Will try.*

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=PLjGG94etKypJWOSPMh8Azcgy5e_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

Student Questions