

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/cse570-18/>



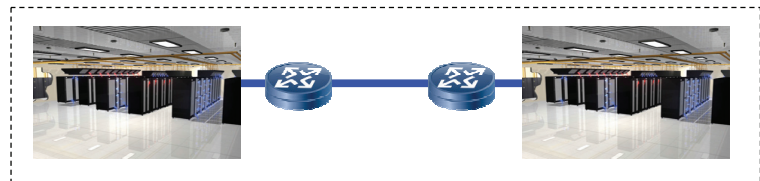
1. TRILL
2. NVGRE
3. VXLAN
4. NVO3
5. Geneve

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, application and data are loaded in standby. Manual switchover ⇒ Significant downtime. (1970-1990)
- ❑ Hot-Standby: Two servers in different 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.

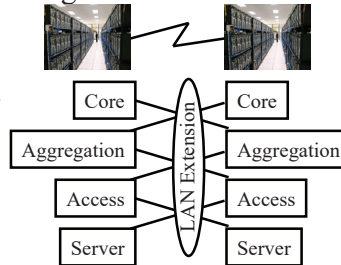
Data Center Interconnection (DCI)

- ❑ 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 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.
 - Root can become 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



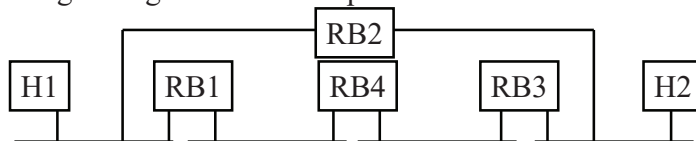
TRILL

- ❑ 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 \Rightarrow 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
 - \Rightarrow Use MAC addresses and IP routing

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

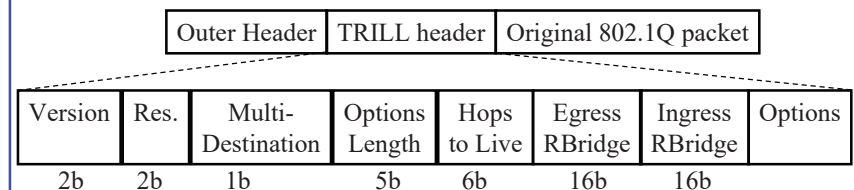
TRILL Architecture

- ❑ 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

TRILL Encapsulation Format



- ❑ For outer headers both PPP and Ethernet headers are allowed. PPP for 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

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 tree in the same extended LAN

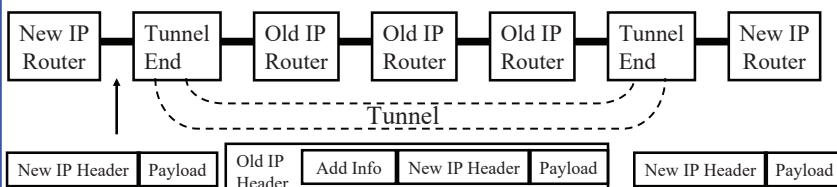


TRILL: Summary

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

GRE

- ❑ Any new feature in IP requires *encapsulation*, a.k.a. *tunneling*
- ❑ Generic Routing Encapsulation (RFC 1701/1702)
- ❑ Generic \Rightarrow 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



GRE (Cont)

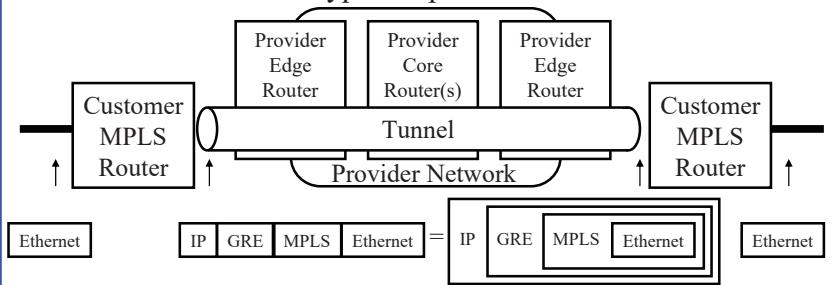
- ❑ GRE provides all of the above encapsulation services
- ❑ Over IPv4, GRE packets use a protocol type of 47
- ❑ Optional: Checksum, Loose/strict Source Routing, Key
- ❑ Key is used either to authenticate the source or to distinguish different substreams
- ❑ Recursion Control: # of additional encapsulations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



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

EoMPLSoGRE

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



Washington University in St. Louis

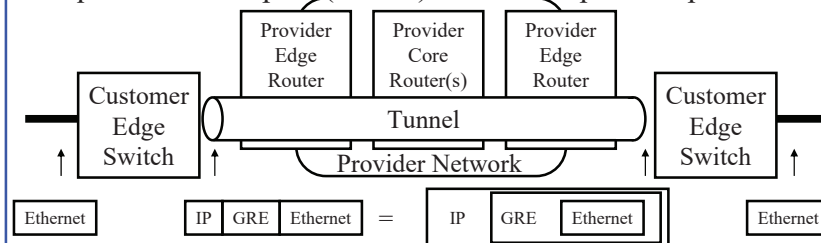
<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 Raj Jain

9-13

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 GRE key field $\Rightarrow 2^{24}$ tenants can share
- ❑ 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: M. Sridharan, "NVGRE: Network Virtualization using GRE," Aug 2013,

<http://tools.ietf.org/html/draft-sridharan-virtualization-nvgre-03>

Washington University in St. Louis

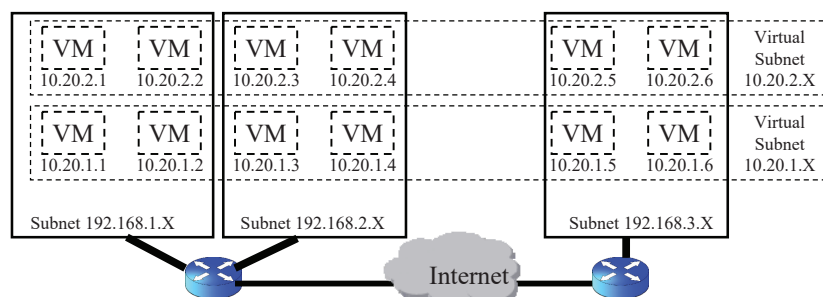
<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 Raj Jain

9-14

NVGRE (Cont)

- ❑ In a cloud, a pSwitch or a vSwitch can serve as 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/els_wp_all_nvgre.pdf

Washington University in St. Louis

<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 Raj Jain

9-15

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 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 \Rightarrow Overlapping MAC, VLAN, and IP addresses
 - Spanning tree is inefficient with large number of switches \Rightarrow Too many links are disabled
 - Better throughput with IP equal cost multipath (ECMP)

Ref: M. Mahalingam, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks," draft-mahalingam-dutt-dcops-vxlan-04, May, 8, 2013, <http://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-04>

Washington University in St. Louis

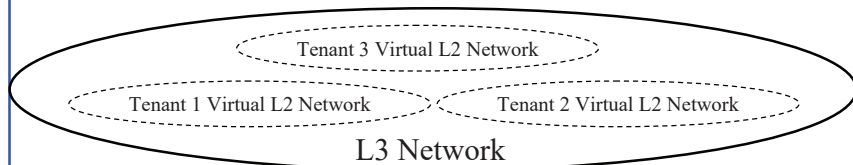
<http://www.cse.wustl.edu/~jain/cse570-18/>

©2018 Raj Jain

9-16

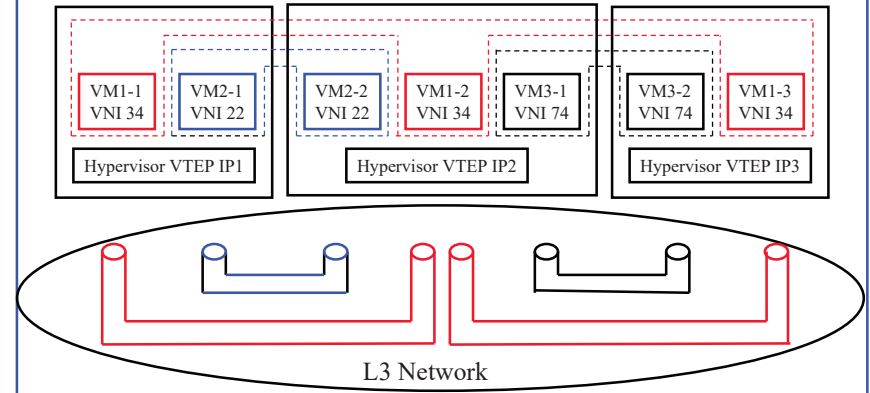
VXLAN Architecture

- ❑ Create a 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).
- ❑ Segments may have overlapping MAC addresses and VLANs but L2 traffic never crosses a VNI



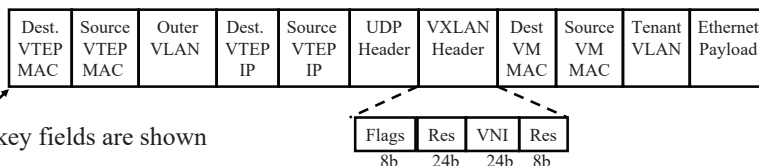
VXLAN Deployment Example

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



VXLAN Encapsulation Format

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



VXLAN Encapsulation Format (Cont)

- ❑ IGMP is used to prune multicast trees
- ❑ 7 of 8 bits in the flag field are reserved.
One flag bit is set if 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 MAC address of other VTEPs and of client VMs of VNIs they are handling.
- ❑ 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 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 responsible for all details.
- ❑ Uses UDP over IP encapsulation to isolate tenants

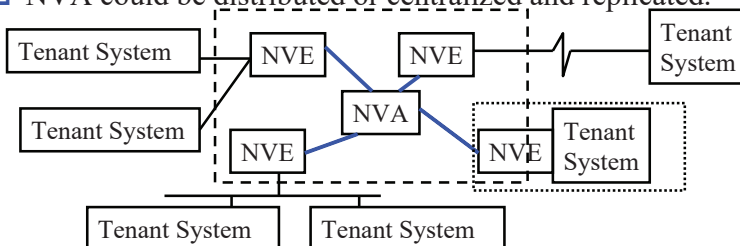
NVO3

- ❑ 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 select its address space
 - Virtual Private \Rightarrow 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.

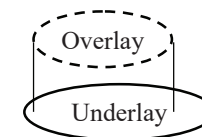
NVO3 Terminology

- ❑ **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.



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



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

Stateless Transport Tunneling Protocol (STT)

- ❑ Ethernet over **TCP-Like** over IP tunnels.
GRE, 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>

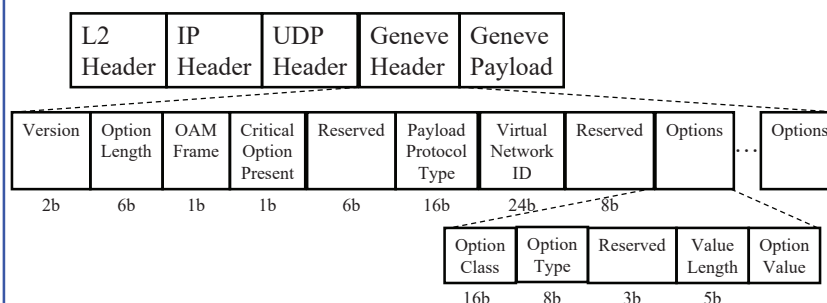
Geneve

- ❑ 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-00, May 8, 2015

Geneve Frame Format

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

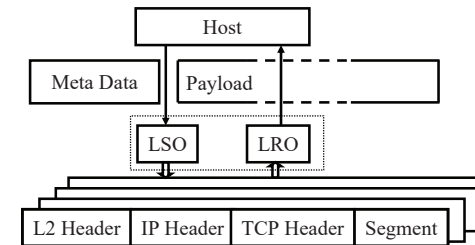


Geneve Frame Format (Cont)

- ❑ **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 bits): 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

LSO and LRO

- ❑ **Large Send Offload (LSO)**: Host hands a large chunk of data to NIC and meta data. NIC makes MSS size segments, 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. Host does the final reassembly with fewer per packet operations.



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 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.

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.

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. All of this is subject to change since it is in the draft stage.

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.

Reading List

- B. Davie and J. Gross, "A Stateless Transport Tunneling Protocol for Network Virtualization (STT)," IETF expired draft, Nov 29, 2016, <https://tools.ietf.org/pdf/draft-davie-stt-08.pdf>
- 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, <draft-ietf-nvo3-geneve-05>, Sep 13, 2017
- 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>

Reading List (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>

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

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

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

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

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 Identifier
- ❑ VPLS Virtual Private LAN Service
- ❑ VPLSoGRE Virtual Private LAN Service over GRE
- ❑ VPN Virtual Private Network

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






Scan This to Download These Slides



Raj Jain

<http://rajjain.com>

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>