Multi-Tenant Isolation 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/cse570-13/

Overview

1. NVO3
2. VXLAN
3. NVGRE
4. STT

Note: Data center interconnection and LAN extension techniques are covered in another module which includes OTV, TRILL, and LISP.

Network Virtualization

1. Network virtualization allows tenants to form an overlay network in a multi-tenant network such that tenant can control:
   1. Connectivity layer: Tenant network can be L2 while the provider is L3 and vice versa
   2. Addresses: MAC addresses and IP addresses
   3. Network Partitions: VLANs and Subnets
   4. Node Location: Move nodes freely
2. Network virtualization allows providers to serve a large number of tenants without worrying about:
   1. Internal addresses used in client networks
   2. Number of client nodes
   3. Location of individual client nodes
   4. Number and values of client partitions (VLANs and Subnets)
3. Network could be a single physical interface, a single physical machine, a data center, a metro, … or the global Internet.
4. Provider could be a system owner, an enterprise, a cloud provider, or a carrier.

Network Virtualization Techniques

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

*All L2/L3 technologies for L2 Network partitioning and aggregation can also be used for L3 network partitioning and aggregation, respectively, by simply putting L3 packets in L2 payloads.

**The aggregation technologies can also be seen as partitioning technologies from the provider point of view.

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 on hundred thousands of physical servers
  - VM (or pM) Migration
  - Support both L2 and L3 VPNs
  - Dynamic provisioning
  - Addressing independence
  - Virtual Private ⇒ Other tenants do not see your frames
  - Optimal Forwarding (VRRP inefficient in a large network)

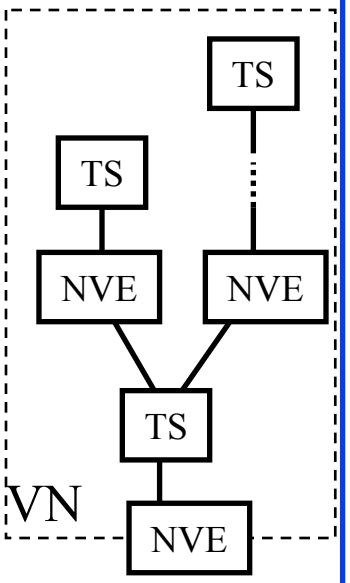
Ref: Network Virtualization Overlays (nvo3) charter, http://datatracker.ietf.org/wg/nvo3/charter/

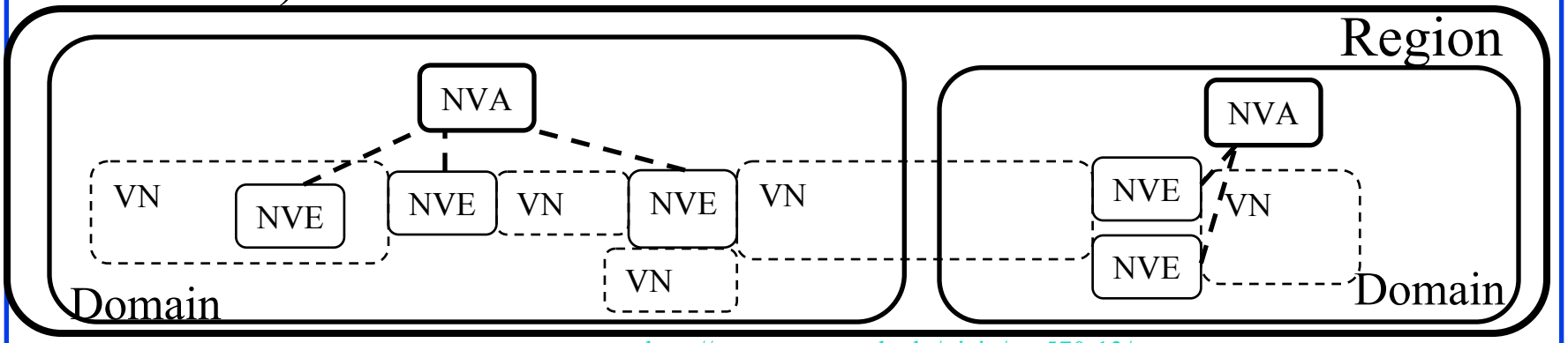
NVO3 Goals

- Develop a general architectural framework
  - Identify key functional blocks.
  - Indentify alternatives for each functional block
  - Deployments can mix and match these alternatives
  - Analyze which requirements are satisfied by different alternatives
  - Operation, Administration and Management (OAM)

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)
- **Network Virtualization Authority (NVA)**: Manages forwarding info for a set of NVEs
- **NV Domain**: Set of NVEs under one authority
- **NV Region**: Set of domains that share some information (to support VNs that span multiple domains)





NVO3 Components

- Underlay Network: Provides overlay network service
- Orchestration Systems: Create new VMs and associated vSwitches and other networking entities and properties. May share this information with NVAs.
- NVEs could be in vSwitches, external pSwitches or span both.
- NVA could be distributed or centralized and replicated.
- NVEs get information from hypervisors and/or NVA.
  - Hypervisor-to-NVE Protocol (data plane learning)
  - NVE-NVA Protocol: Push or Pull (on-demand) model. Control plane learning.
- Map and Encap: Find destination NVE (map) and send (encap)

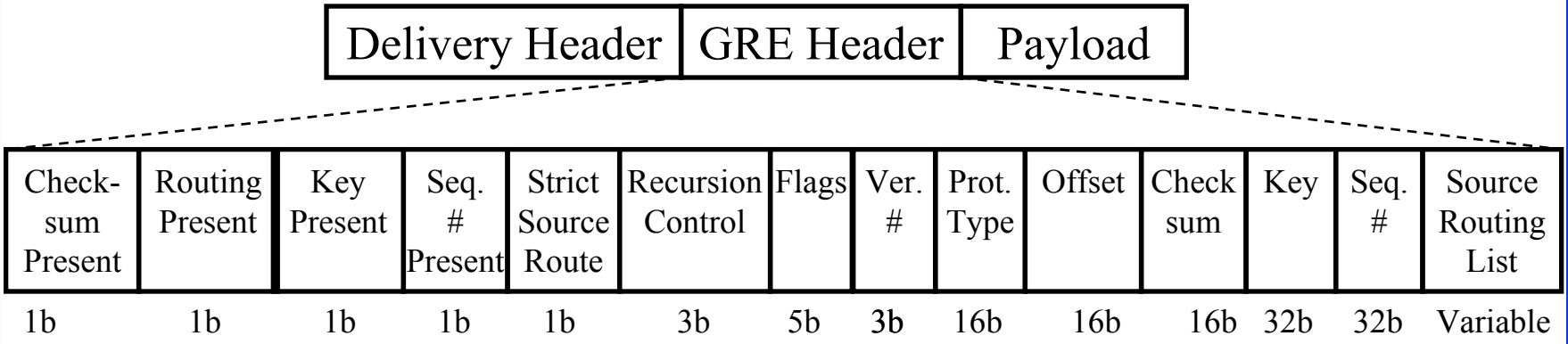
Ref: T. Narten, et al., "An Architecture for Overlay Networks (NVO3)," http://datatracker.ietf.org/doc/draft-narten-nvo3-arch/

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

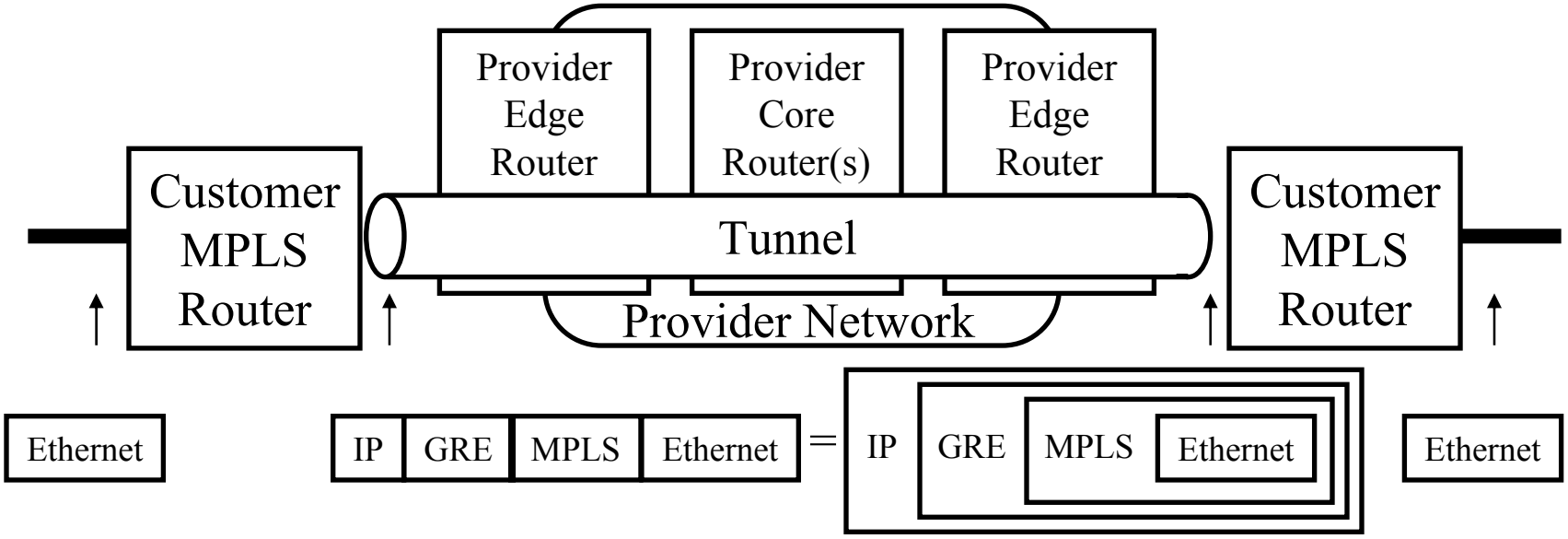
GRE

- Generic Routing Encaptulation (RFC 1701/1702)
- Generic ⇒ X over Y for any X or Y
- Over IPv4, GRE packets use a protocol type of 47
- Optional Checksum, Loose/strict Source Routing, Key
- Key is used to authenticate the source
- 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



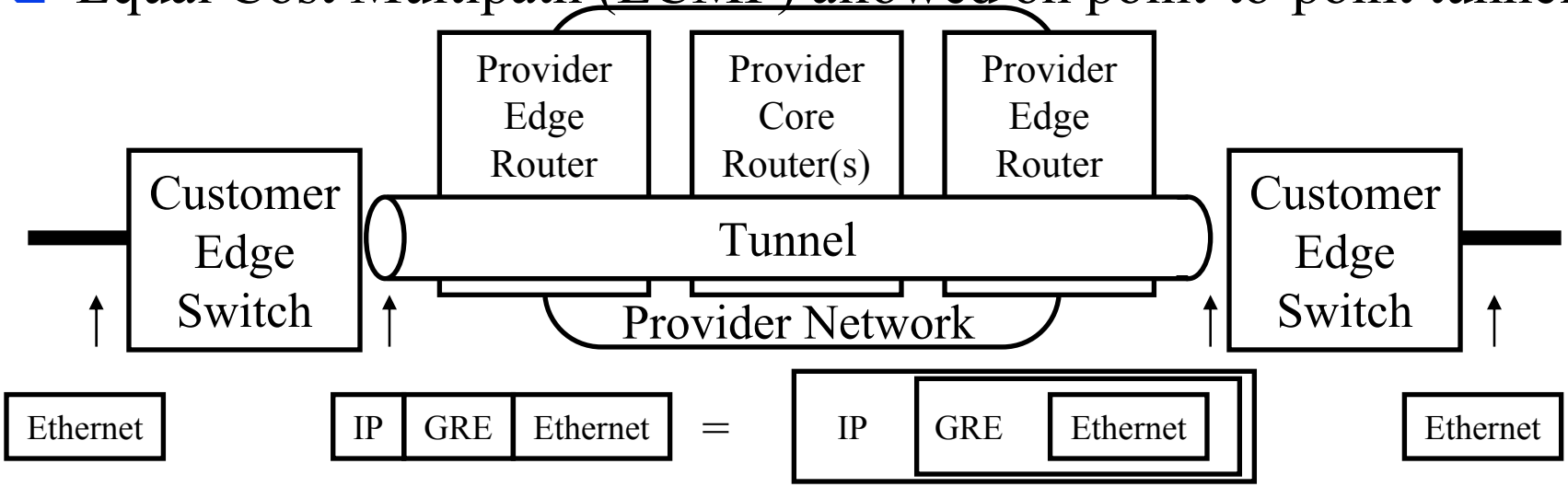
EoMPLSoGRE

- 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



NVGRE

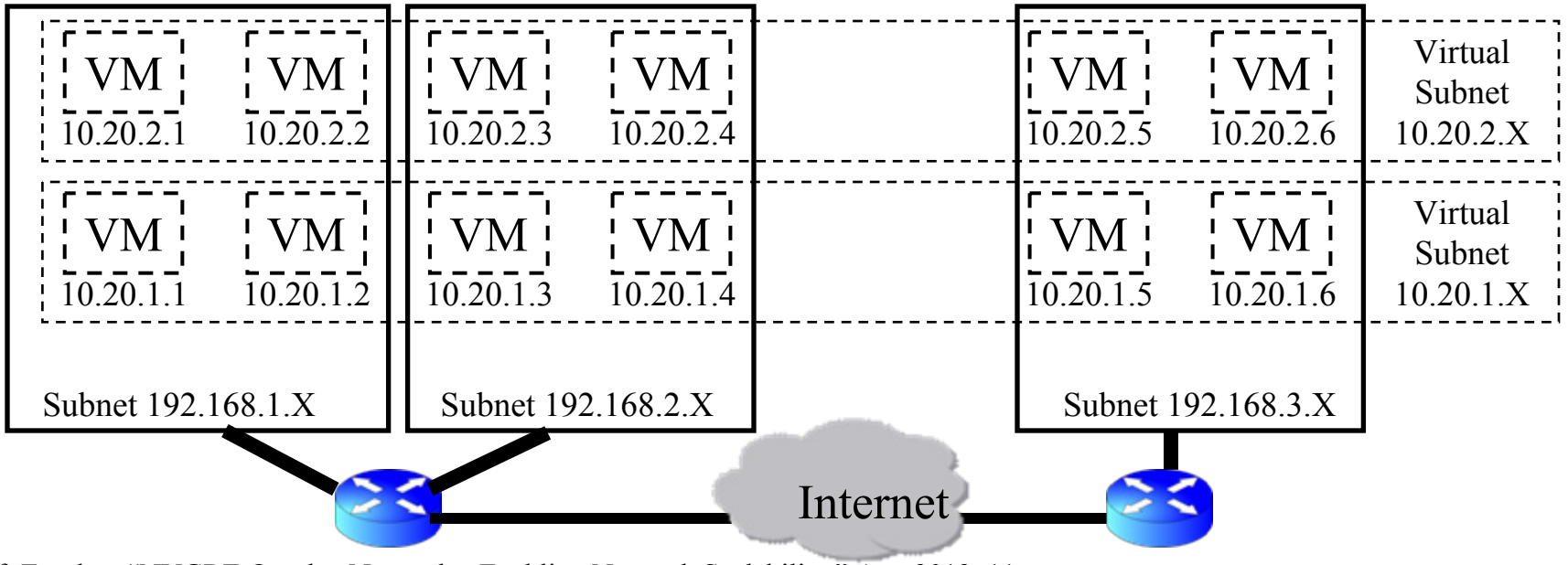
- Ethernet over GRE over IP (point-to-point)
- A unique 24-bit Virtual Subnet Identifier (VSID) is used as the lower 24-bits of GRE key field ⇒ 2^{24} tenants can share
- Unique IP multicast address is used for BUM (Broadcast, Unknown, Multicast) traffic on each VSID
- Equal Cost Multipath (ECMP) allowed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



Ref: M. Sridharan, "MVGRE: Network Virtualization using GRE," Aug 2013, http://tools.ietf.org/html/draft-sridharan-virtualization-nvgre-03

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/elx_wp_all_nvgre.pdf

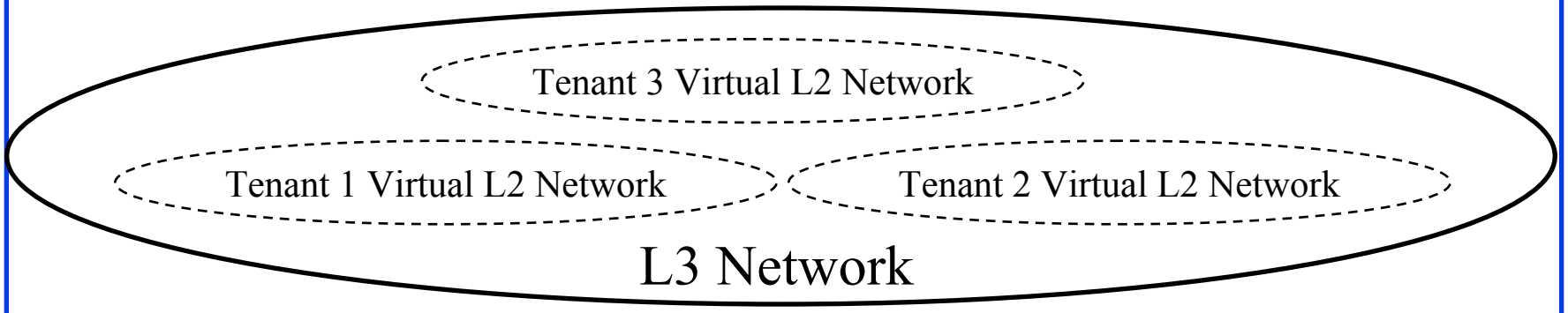
VXLAN

- 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 ⇒ Overlapping MAC, VLAN, and IP addresses
  - Spanning tree is inefficient with large number of switches ⇒ 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

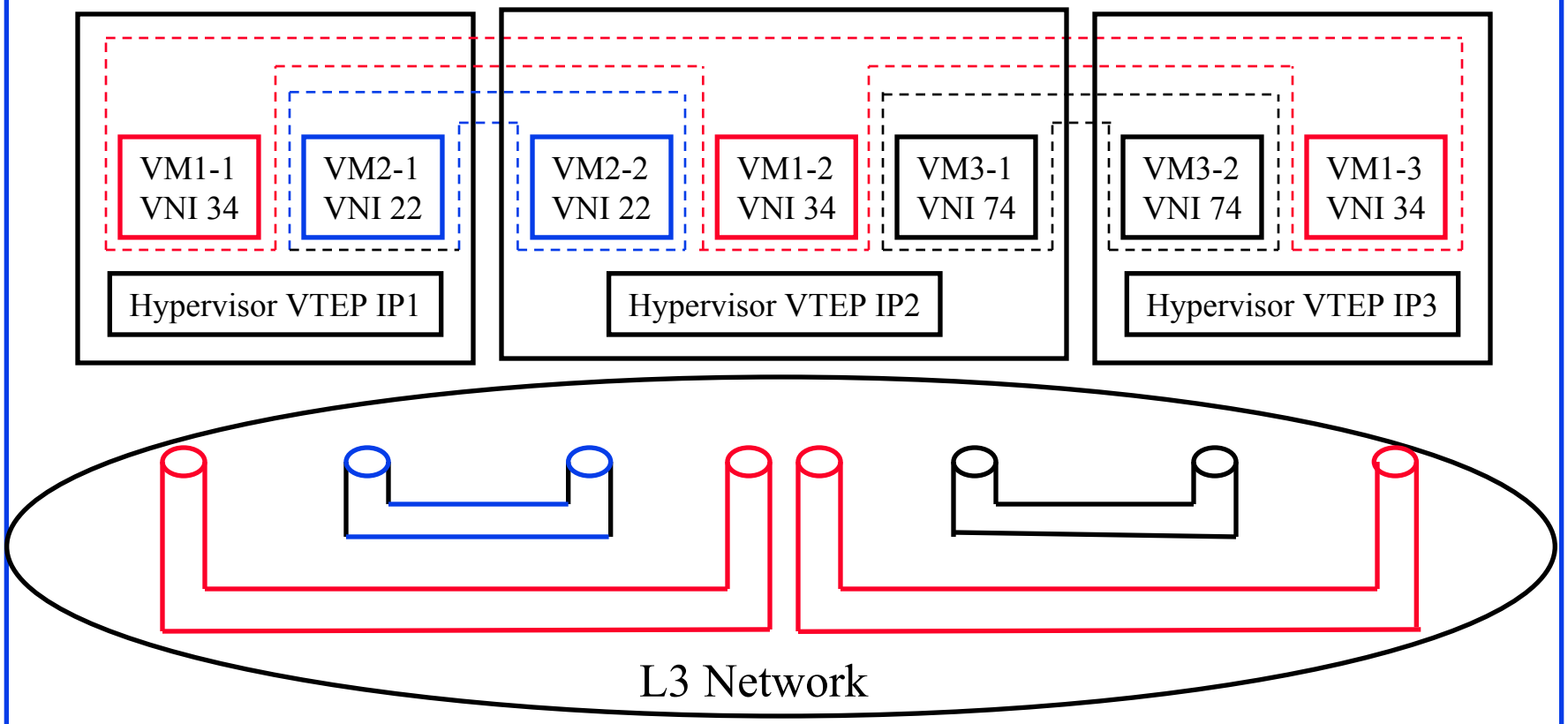
VXLAN Architecture

- 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).
  $\Rightarrow$ 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



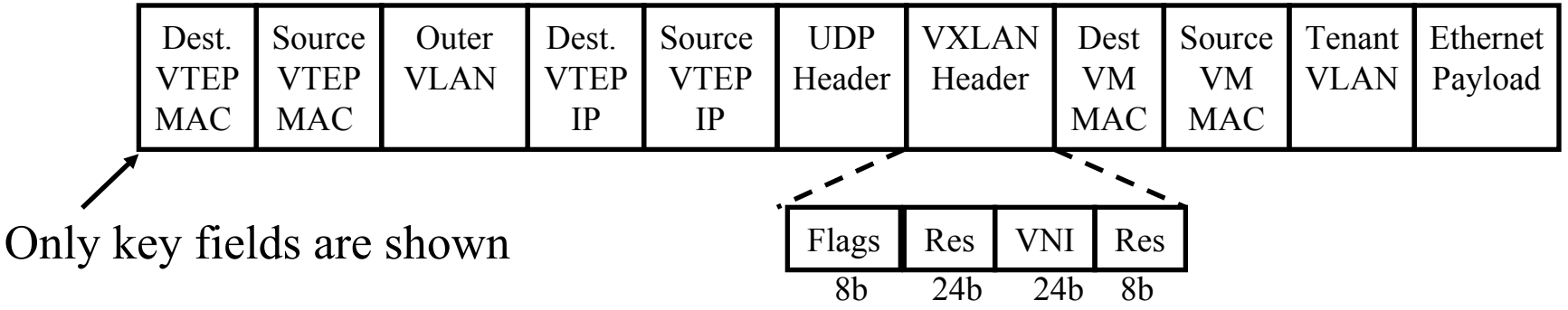
VXLAN Deployment Example

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



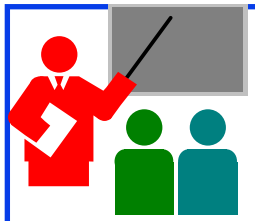
VXLAN Encapsulation Format

- 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.
  I flag bit is set if VNI field is valid
- UDP source port is a hash of the inner MAC header
  $\Rightarrow$ 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

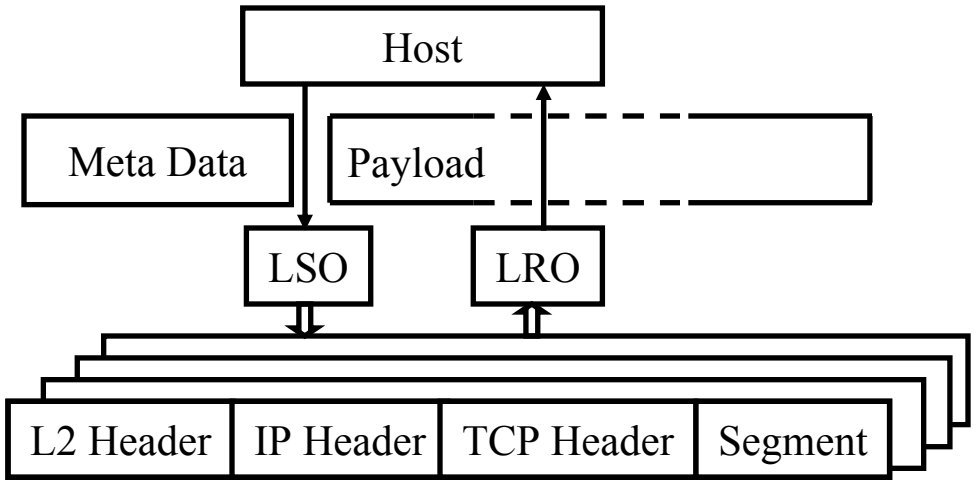
Stateless Transport Tunneling Protocol (STT)

- Ethernet over TCP-Like over IP tunnels.
  GRE, IPSec tunnels can also be used if required.
- Tunnel endpoints may be inside the end-systems (vSwitches)
- 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).
- Broadcast, Unknown, Multicast (BUM) handled by IP multicast tunnels

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

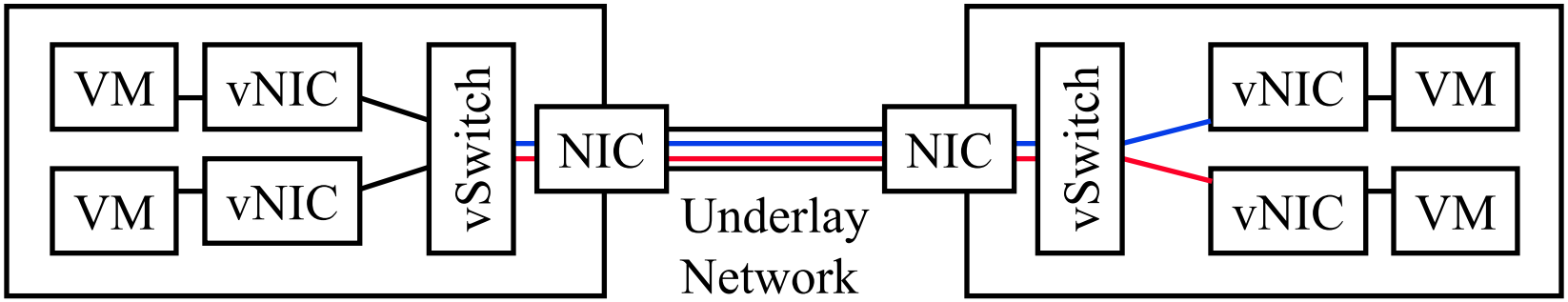
LSO and LRO

- 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.
- STT takes advantage of LSO and LRO features, if available.
- Using a protocol number other than 6 will not allow LSO/LRO to handle STT



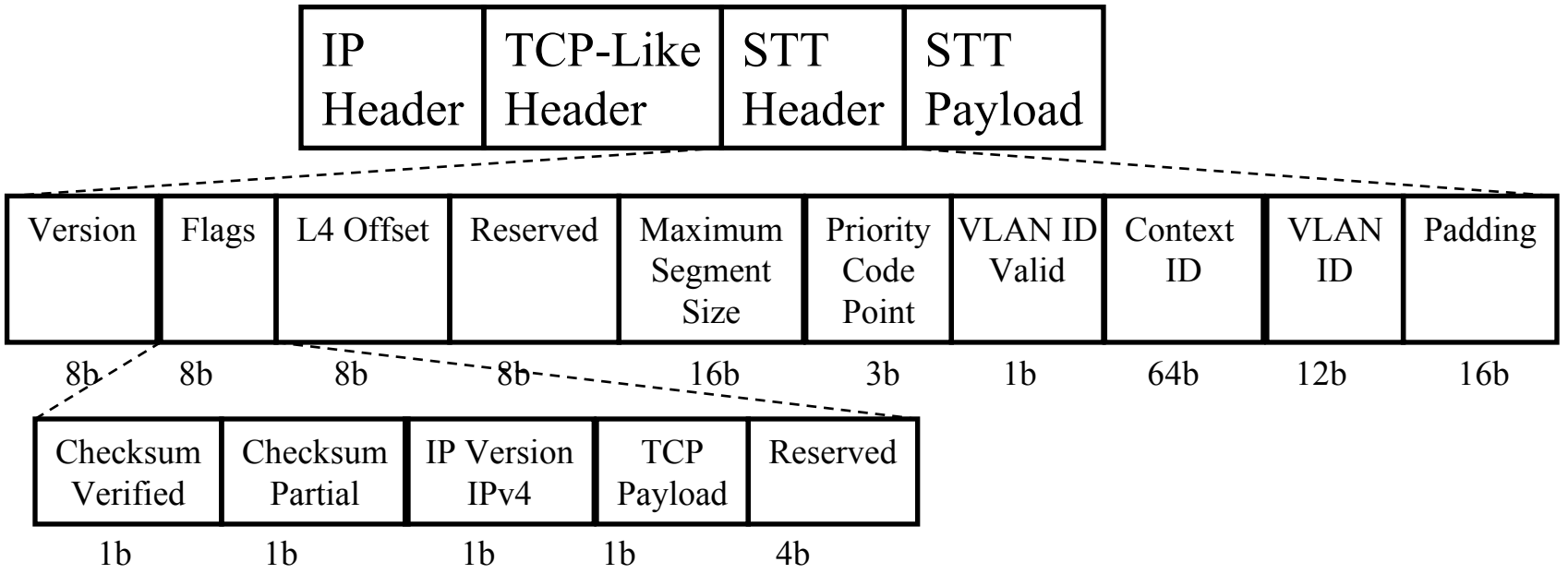
STT Optimizations

- Large data size: Less overhead per payload byte
- Context ID: 64-bit tunnel end-point identifier
- Optimizations:
  - 2-byte padding is added to Ethernet frames to make its size a multiple of 32-bits.
  - Source port is a hash of the inner header ⇒ ECMP with each flow taking different path and all packets of a flow taking one path
- No protocol type field ⇒ Payload assumed to be Ethernet, which can carry any payload identified by protocol type.



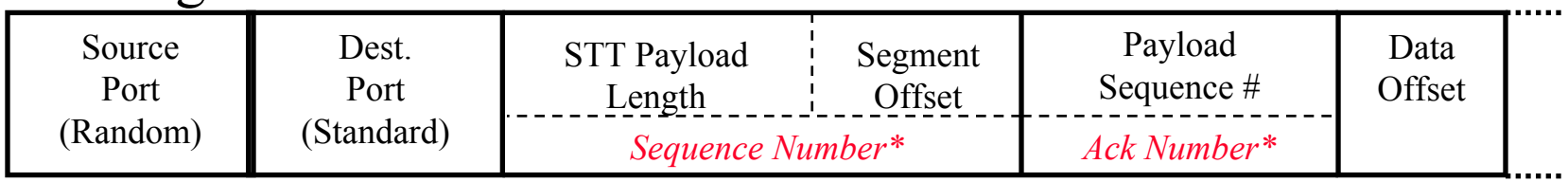
STT Frame Format

- 16-Bit MSS $\Rightarrow 2^{16}$ B = 64K Byte maximum
- L4 Offset: From the of STT header to the start of encapsulated L4 (TCP/UDP) header $\Rightarrow$ Helps locate payload quickly
- Checksum Verified: Checksum covers entire payload and valid
- Checksum Partial: Checksum only includes TCP/IP headers



TCP-Like Header in STT

- Destination Port: Standard to be requested from IANA
- Source Port: Selected for efficient ECMP
- Ack Number: STT payload sequence identifier. Same in all segments of a payload
- Sequence Number (32b): Length of STT Payload (16b) + offset of the current segment (16b) ⇒ Correctly handled by NICs with Large Receive Offload (LRO) feature
- No acks. STT delivers partial payload to higher layers.
- Higher layer TCP can handle retransmissions if required.
- Middle boxes will need to be programmed to allow STT pass through

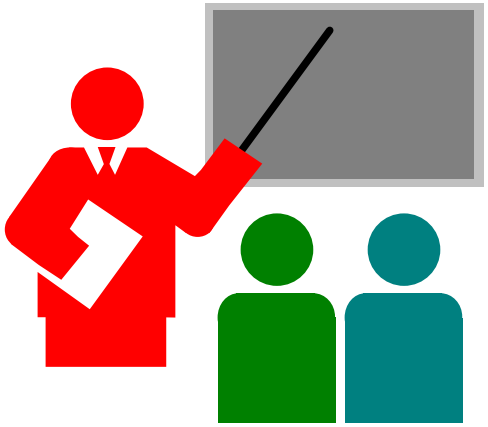


*Different meaning than TCP

STT Summary

- STT solves the problem of *efficient* transport of large 64 KB storage blocks
- Uses Ethernet over TCP-Like over IP tunnels
- Designed for software implementation in hypervisors

Summary



1. NVO3 is a generalized framework for network virtualization and partitioning for multiple tenants over L3. It covers both L2 and L3 connectivity.
2. NVGRE uses Ethernet over GRE for L2 connectivity.
3. VXLAN uses Ethernet over UDP over IP
4. STT uses Ethernet over TCP-like stateless protocol over IP.

Reading List

- 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
- 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
- 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
- M. Sridharan, "MVGRE: Network Virtualization using GRE," Aug 2013, http://tools.ietf.org/html/draft-sridharan-virtualization-nvgre-03
- Network Virtualization Overlays (nvo3) charter, http://datatracker.ietf.org/wg/nvo3/charter/
- T. Narten, et al., "An Architecture for Overlay Networks (NVO3)," http://datatracker.ietf.org/doc/draft-narten-nvo3-arch/
- V. Josyula, M. Orr, and G. Page, "Cloud Computing: Automating the Virtualized Data Center," Cisco Press, 2012, 392 pp., ISBN: 1587204347.

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
- DCN — Data Center Networks
- DCVPN — Data Center Virtual Private Network
- ECMP — Equal Cost Multi Path
- EoMPLSoGRE — Ethernet over MPLS over GRE
- EVPN — Ethernet Virtual Private Network
- GRE — Generic Routing Encapsulation
- IANA — Internet Address and Naming Authority
- ID — Identifier
- IEEE — Institution of Electrical and Electronic Engineers
- IETF — Internet Engineering Task Force

Acronyms (Cont)

- IGMP Internet Group Multicast Protocol
- IP Internet Protocol
- IPSec IP Security
- IPv4 Internet Protocol V4
- 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
- MTU Maximum Transmission Unit
- NIC Network Interface Card
- NV Network Virtualization
- NVA Network Virtualization Authority
- NVEs Network Virtualization Edge

Acronyms (Cont)

- 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
- RFC Request for Comment
- RS Routing System
- STT Stateless Transport Tunneling Protocol
- TCP Transmission Control Protocol
- TRILL Transparent Routing over Lots of Links
- TS Tenant System
- UDP User Datagram Protocol

Acronyms (Cont)

- 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
- VPLSoGRE VPLS over GRE
- VPN Virtual Private Network
- 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