Cloud e Datacenter Networking

Università degli Studi di Napoli Federico II
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione DIETI
Laurea Magistrale in Ingegneria Informatica

Prof. Roberto Canonico

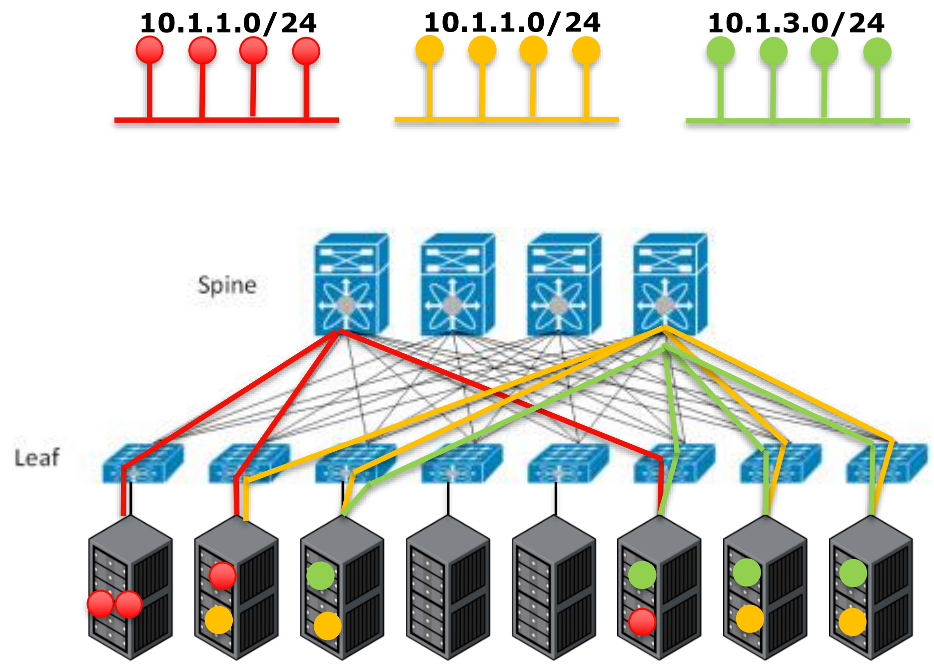
Datacenter networking and multitenancy

Lesson outline

- Multitenancy
- Virtual networking techniques in a datacenter
- Tunneling protocols

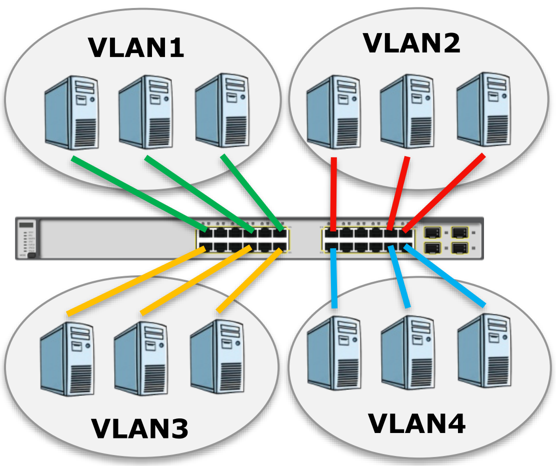
Virtual networking in a Cloud datacenter

- In a multi-tenant virtualized datacenter proper solutions are needed to map multiple independent virtual infrastructures (provided as a IaaS service) on top of a shared physical infrastructure
  - Requirements: isolation, fully flexible VM placement and migration, address independance
  - Challenges: address collisions, partitioning, mapping, …



VLANs

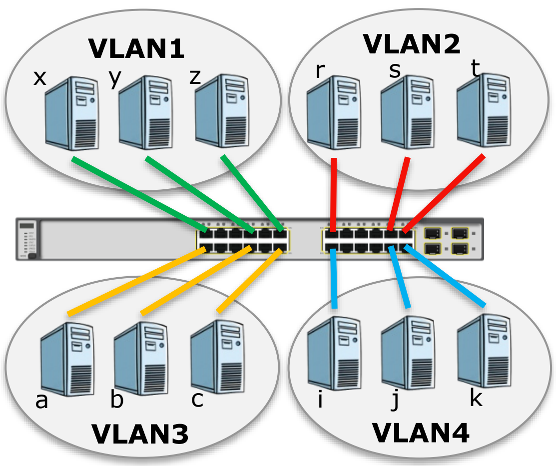
- VLANS are a first approach to network virtualization
- VLANs create separate broadcast domains within the same switch
  - Needed if multiple IP subnets need to coexist in the same switch
  - A router is needed to route traffic between VLANs
- In a single switch network, VLANs are typically assigned to ports by the admin



- Each switch port could be assigned to a different VLAN
- Ports assigned to the same VLAN share broadcasts
- Ports that do not belong to the same VLAN do not share broadcasts
- The default VLAN for every port in the switch is the “native VLAN”
  - The native VLAN is always VLAN 1 and may not be deleted
- All other ports on the switch may be reassigned to alternate VLANs

VLAN bridging tables

- Implementing VLANs on a switch causes the following to occur
  - The switch maintains a separate *bridging table* for each VLAN
  - If a frame comes in on a port in VLAN x, the switch searches the bridging table for VLAN x
  - When a frame is received, the switch adds the source address to the bridging table if it is currently unknown
  - The destination is checked so a forwarding decision can be made
  - For learning and forwarding the search is made against the address table for that VLAN only



VLAN1 bridging table

MAC address	port
x	1
y	7
z	11

VLAN2 bridging table

MAC address	port
r	13
s	21
t	23

VLAN3 bridging table

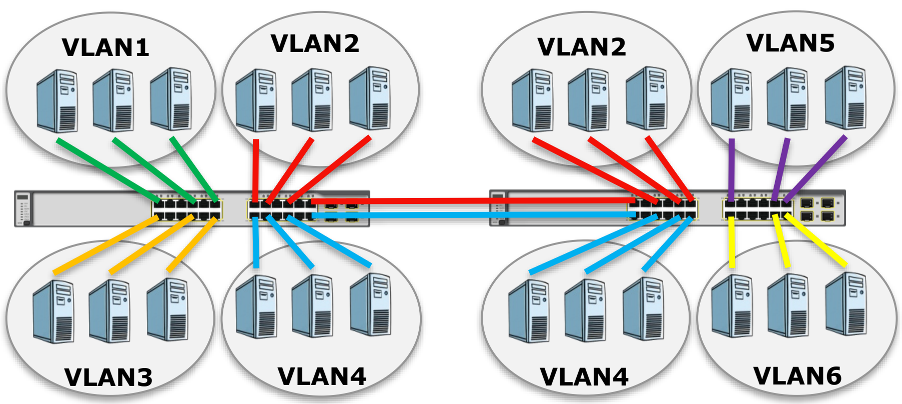
MAC address	port
a	2
b	8
c	12

VLAN4 bridging table

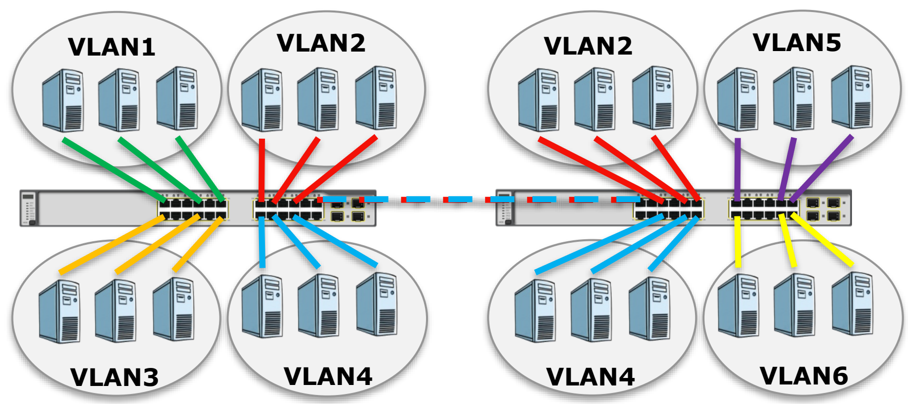
MAC address	port
i	14
j	22
k	24

VLANs spanning multiple switches

- Problem: how to extend multiple VLANs over two distinct switches ?
- Solution #1
  - one link connecting the two switches for each VLAN that needs to be extended
  - costly and inefficient
- Solution #2 – *port trunking*
  - a single link (*trunk*) connects the two switches and carries traffic for all the VLANs that live in both switches
  - To associate each frame to the corresponding VLAN, a special tag is required in the frame header (*VLAN tagging*)
- In general, a *trunk* is a link carrying traffic for several VLANs and a switch may have several trunking ports



Two pairs of ports dedicated to extend VLANs, one for VLAN2 and another for VLAN4



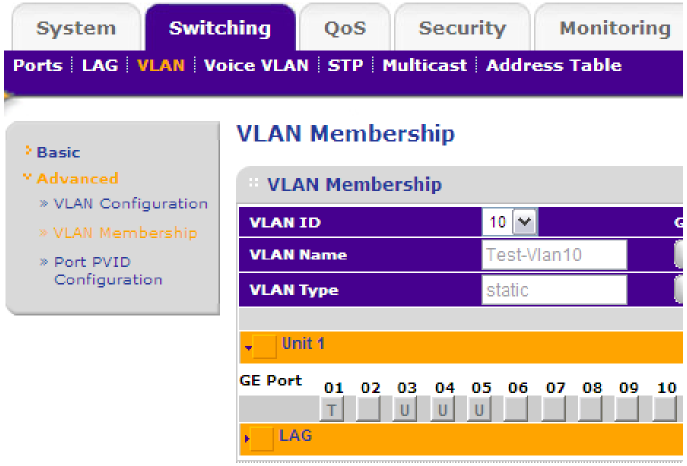
VLANs extended by means of port trunking

VLAN tagging

- VLAN Tagging is used when a link connecting two different switches needs to carry traffic for more than one VLAN
- A unique packet identifier is added within each header to designate the VLAN membership of each packet
- When a packet enters a trunk port with a given VLAN ID:
  - VLAN ID is removed from the packet
  - Packet is forwarded to the appropriate port based on the VLAN ID and destination MAC address
  - If the destination MAC address is FF:FF:FF:FF:FF:FF, the packet is forwarded to all the VLAN ports
- 2 major methods of VLAN tagging: Cisco proprietary Inter-Switch Link (ISL) and IEEE 802.1Q
- IEEE 802.1Q inserts VLAN ID (12 bits) in a new header field

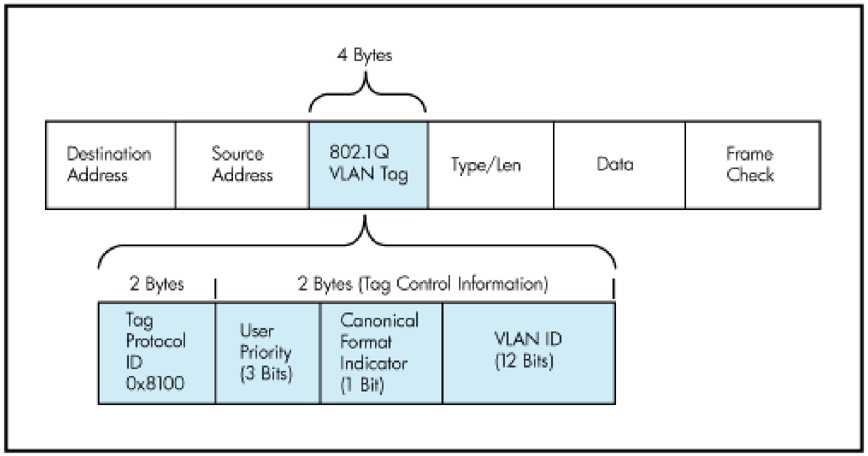
Port 01 is configured as a trunk port for VLAN 10
(T stands for Tagged)

Ports 03, 04 and 05 are statically associated to VLAN 10
without any tagging (U stands for Untagged)



IEEE 802.1Q header

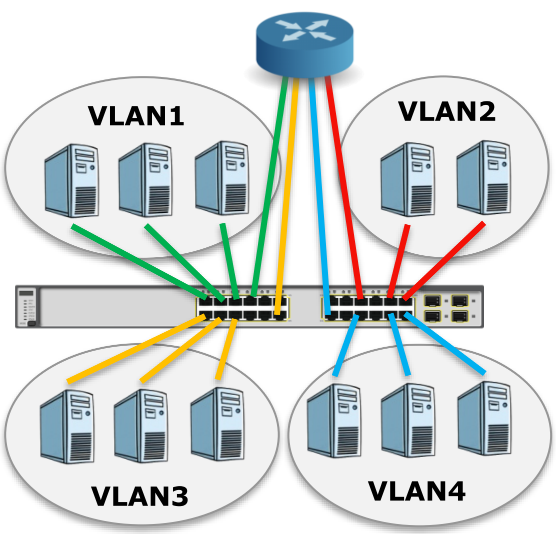
- IEEE 802.1Q adds a 4-byte header field:
- 2-byte tag protocol identifier (TPID) with a fixed value of 0x8100
- 2-byte tag control information (TCI) containing the following elements:
  - Three-bit user priority (8 priority levels, 0 thru 7)
  - One-bit canonical format (CFI indicator), 0 = canonical, 1 = noncanonical, to signal bit order in the encapsulated frame (see IETF RFC2469)
  - Twelve-bit VLAN identifier (VID) - Uniquely identifies the VLAN to which the frame belongs
    - defining 4,096 VLANs, with 0 and 4095 reserved values

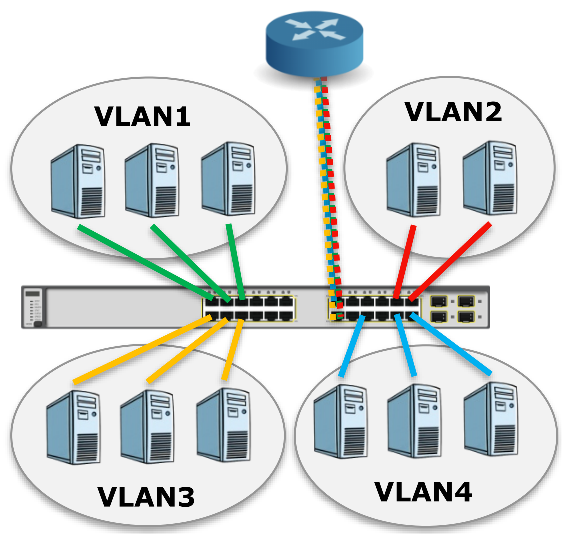


Inter-VLAN routing

- When a node in one VLAN needs to communicate with a node in another VLAN, a router is necessary to route the traffic between VLANs
  - Without the routing device, inter-VLAN traffic would not be possible
- The routing function may be external or internal to the switch
  - In the latter case, the switch itself acts as a router (so called *multilayer switches* or L3 switches)
- External router
  - Approach #1: the router is connected to the switch by one link per VLAN
  - Approach #2: the router is connected to the switch by one trunk link for all the VLANs
    - Also known as "router on a stick"
    - Possible only if the router supports sub-interfaces
      to divide a single physical interface into multiple logical interfaces

Router connected
by as many links
as the VLANs
to be connected



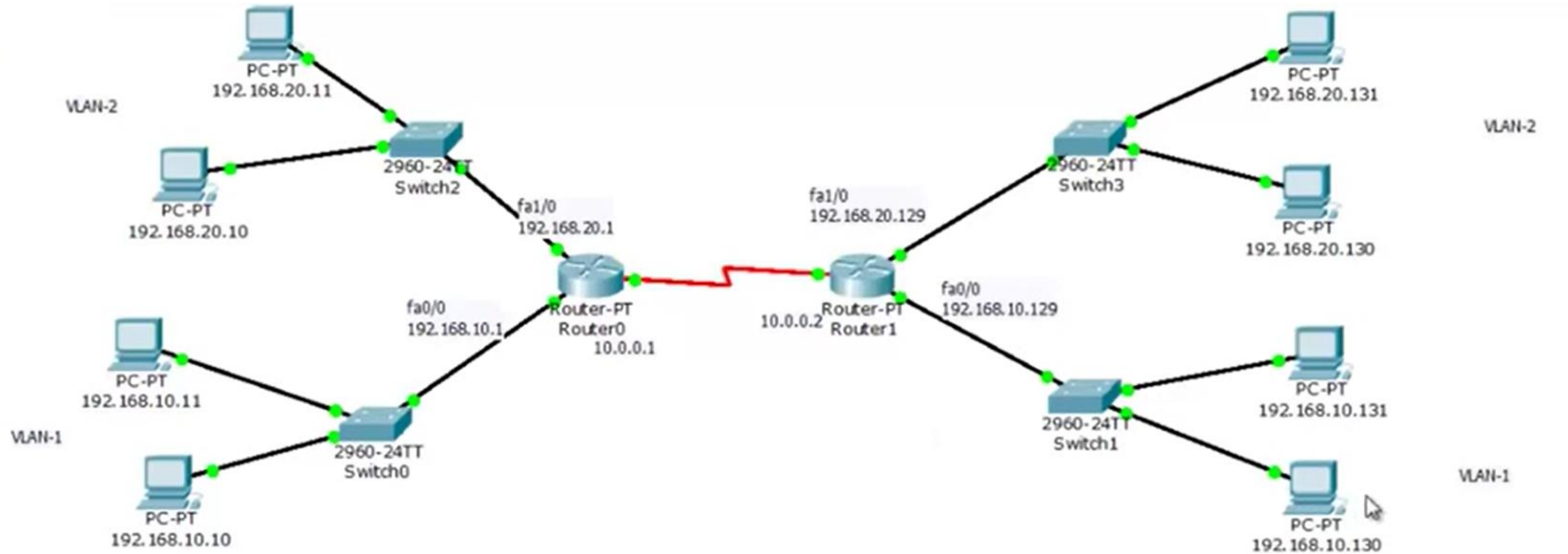


Router-on-a-stick

Router connected
by one trunk link
for all the VLANs

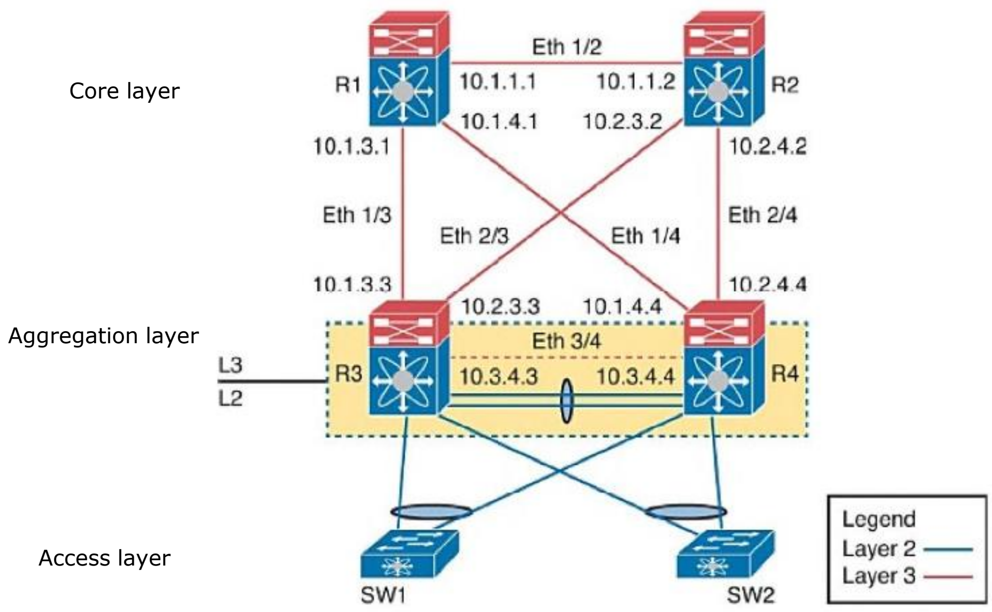
Inter-VLAN routing across different switches

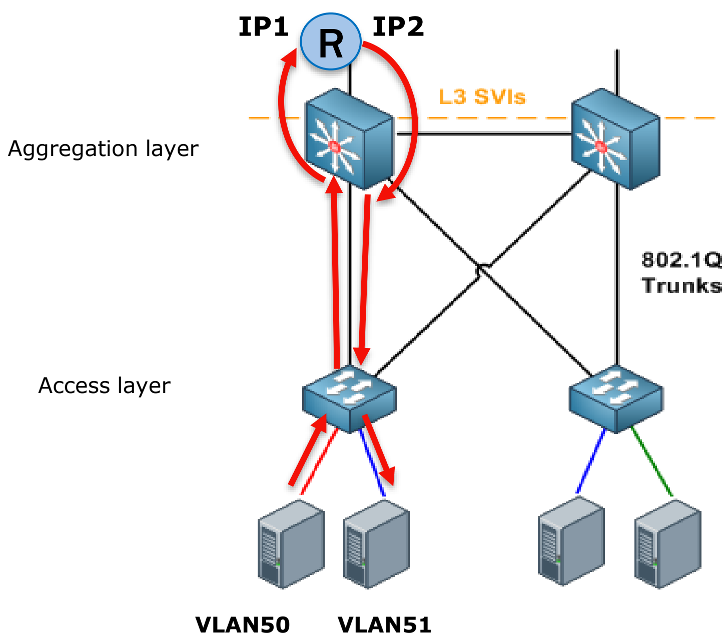
- This scenario is an enteprise network, does not fit a datacenter
- Two VLANs, spread across two distinct switches connected by routers
- In fact, these are four VLANs, each associated to a /25 subnet
- Communication between host 192.168.20.10 (on the left) and 192.168.10.10 (on the left) is routed by Router0
- Communication between host 192.168.10.11 (on the left) and 192.168.10.130 (on the right) is routed by Router0 and Router1



Multilayer switches in a datacenter

- A multilayer switch is able to perform both kinds of packet forwarding: *bridging* at Layer 2 and *routing* at Layer 3
- Layer 3 routing in an aggregation switch can be used to route traffic among different VLANs without the need for an external router by means of so-called “Virtual Switch Interfaces” (SVIs)
    - An SVI should be configured with the VLAN’s default gateway IP address
- In a typical datacenter networks, aggregation layer switches are multilayer switches
- If one needs to exchange traffic among 2 servers (or 2 VMs) associated to 2 different VLANs, this machine-to-machine traffic would traverse the network hierarchy up to the aggregation switch even though the communicating hosts (or VMs) are physically located in the same rack



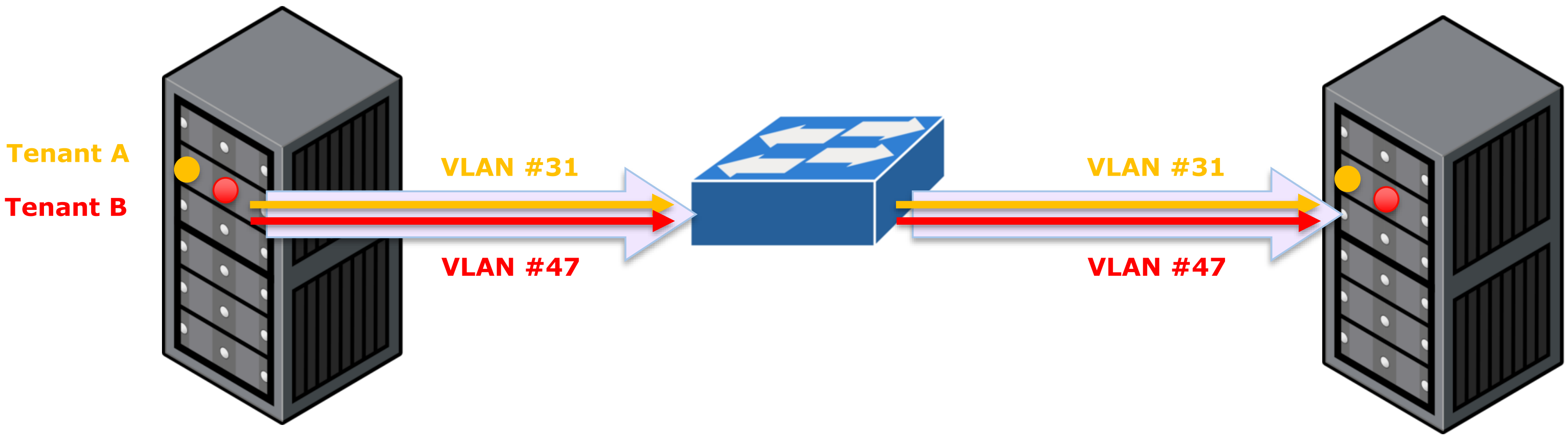


Multi-tenancy: approaches

- Network virtualization techniques allow to map logical tenant networks onto a common shared physical substrate
- Most common network virtualization approaches are based on traffic encapsulation (a.k.a. *tunneling*) and creation of *overlays*
- VLANs is a form of layer 2 encapsulation natively supported by Ethernet switches
- Other forms of encapsulation:
    - Q-in-Q
    - VXLAN: Virtual Extensible LAN
    - NVGRE: Network Virtualization using Generic Routing Encapsulation
    - MPLS

Tenant isolation via VLANs

- Within an L2 island, VLANs can be used to isolate tenants' traffic
- Limitations:
  - Only 4096 VLAN IDs available in IEEE 802.1q
  - Tenants are not allowed to choose VLAN IDs to preserve uniqueness



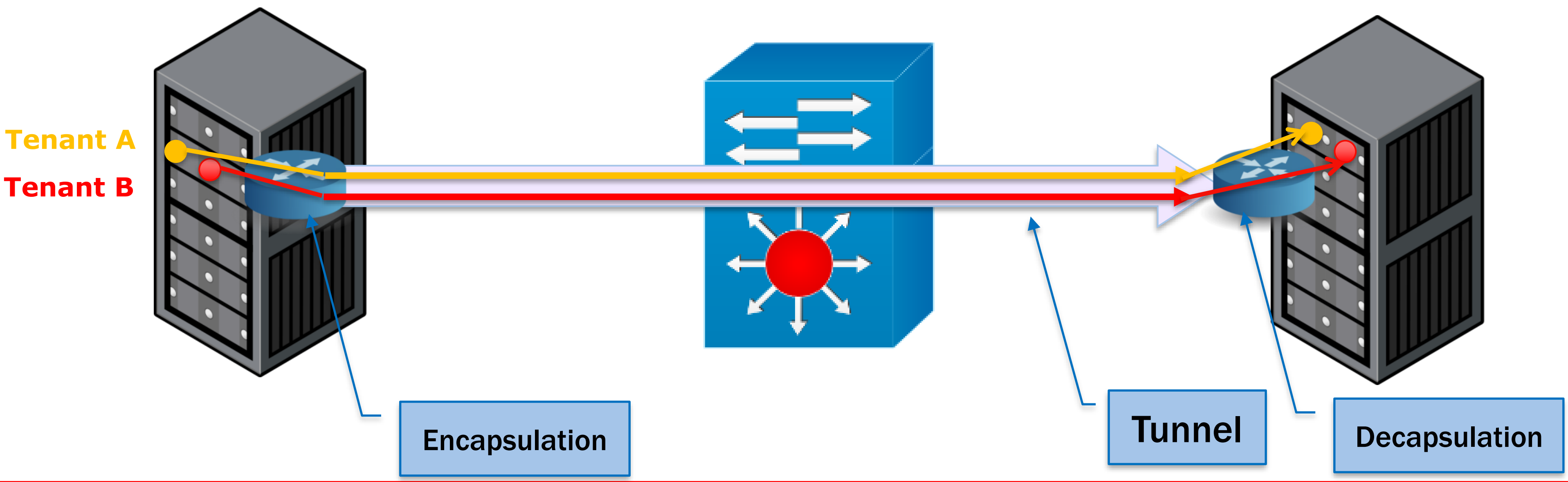
IEEE 802.1ad: Q-in-Q

- IEEE 802.1ad allows to wrap an 802.1q VLAN-tagged packet with an outer VLAN tag (*Q-in-Q*)
- This technique is used to carry proprietary VLAN-tagged traffic on a shared service provider network where the outer 12-bit VLAN ID is used to identify the customer traffic in the provider network
  - Mainly adopted in Metro Ethernet services
- The 3-bit VLAN priority field may be used to provide different classes of service in the provider network
- The inner VLAN ID is left untouched and can be used by the customer for their own purposes
- The 12-bit limit of the VLAN ID severely limits the usability of this technique in large scale provider networks and datacenters

DMAC	SMAC	Outer VLAN	Inner VLAN	Type	IP	TCP/UDP	Data

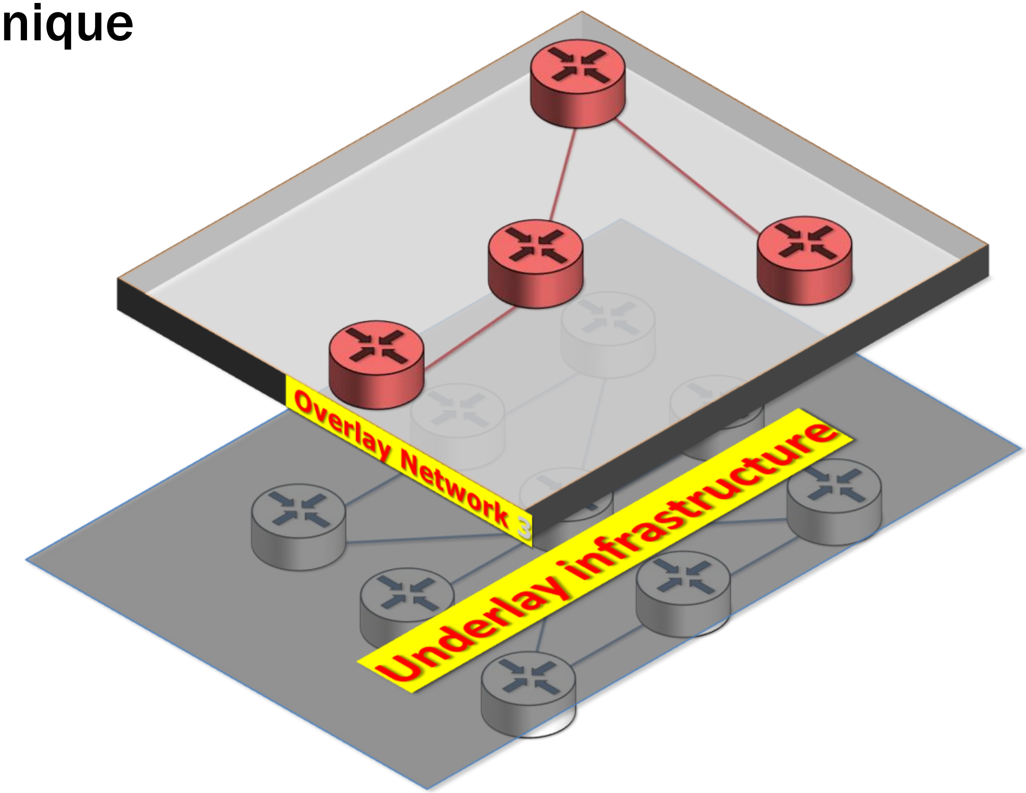
Network Virtualization using encapsulation

- VXLAN and NVGRE are two different network virtualization methods that use encapsulation and tunneling to create large numbers of virtual LANs for subnets that can extend across layer 2 and 3
- Encapsulation/decapsulation is performed by entities that could reside either in End Devices or in ToR edge switches (or in both)
- VXLAN is supported by Cisco and Vmware
- NVGRE was proposed by Microsoft, Intel, HP and Dell



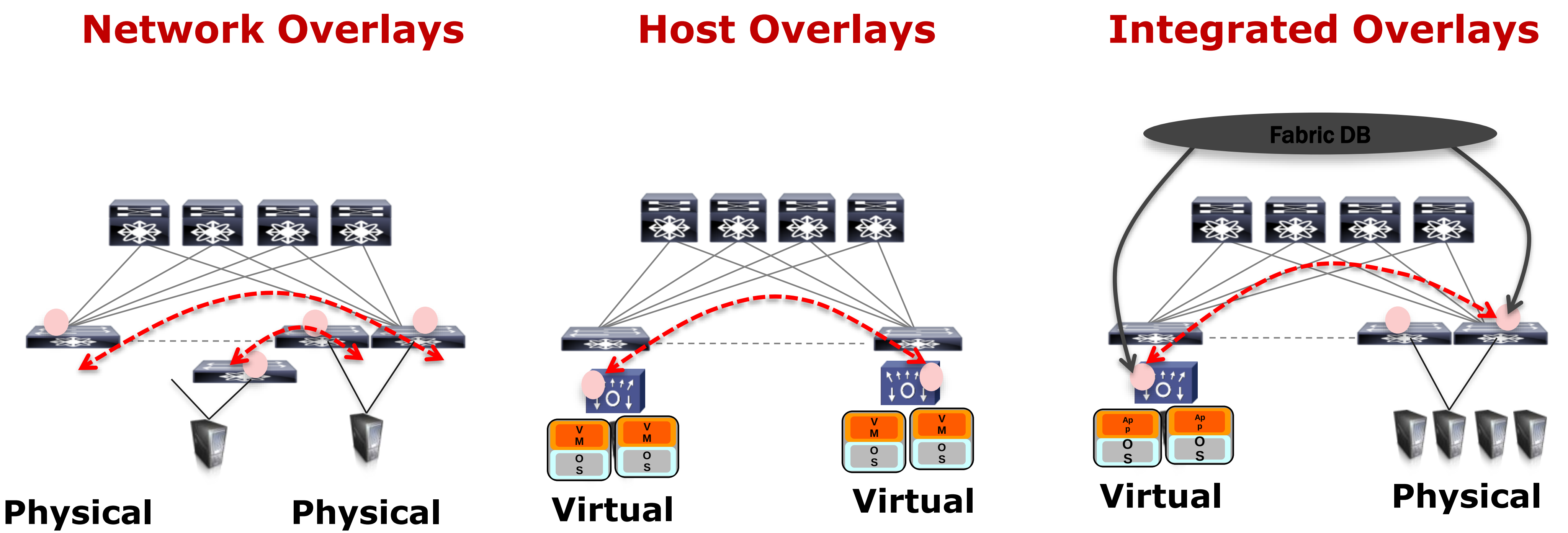
Overlay networks

- Encapsulation technologies all build an overlay network "on top of" a shared underlay infrastructure
- Hence, overlay networking is a sort of "network virtualization" technique
- In cloud computing environments, overlays are built to separate different tenants traffic
- Overlays may be built in different ways and may be deployed both in LAN and WAN contexts



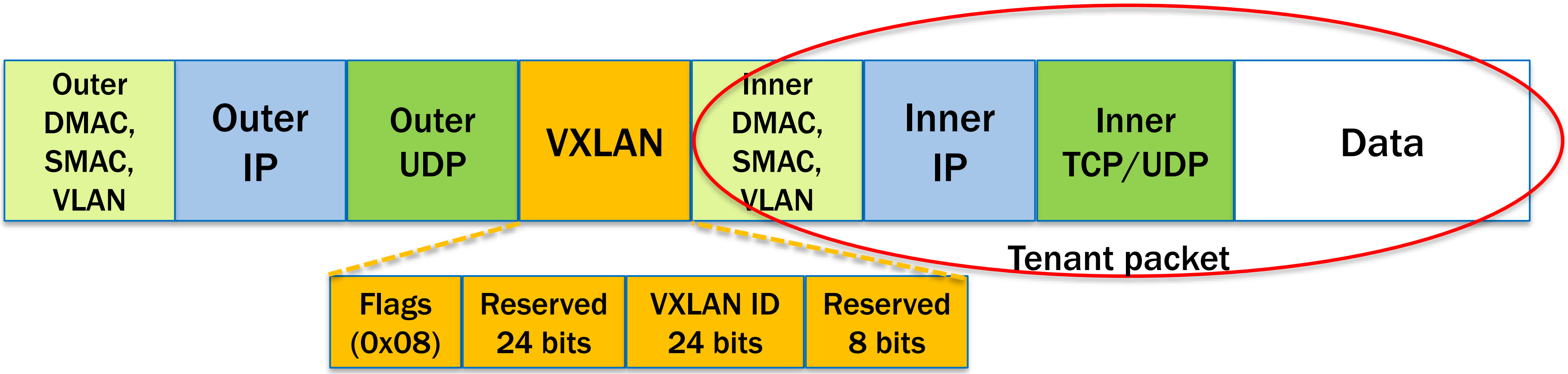
Types of overlays

- The endpoints of tunnels may be either physical devices or virtual network functions
- This leads to thre different types of overlays



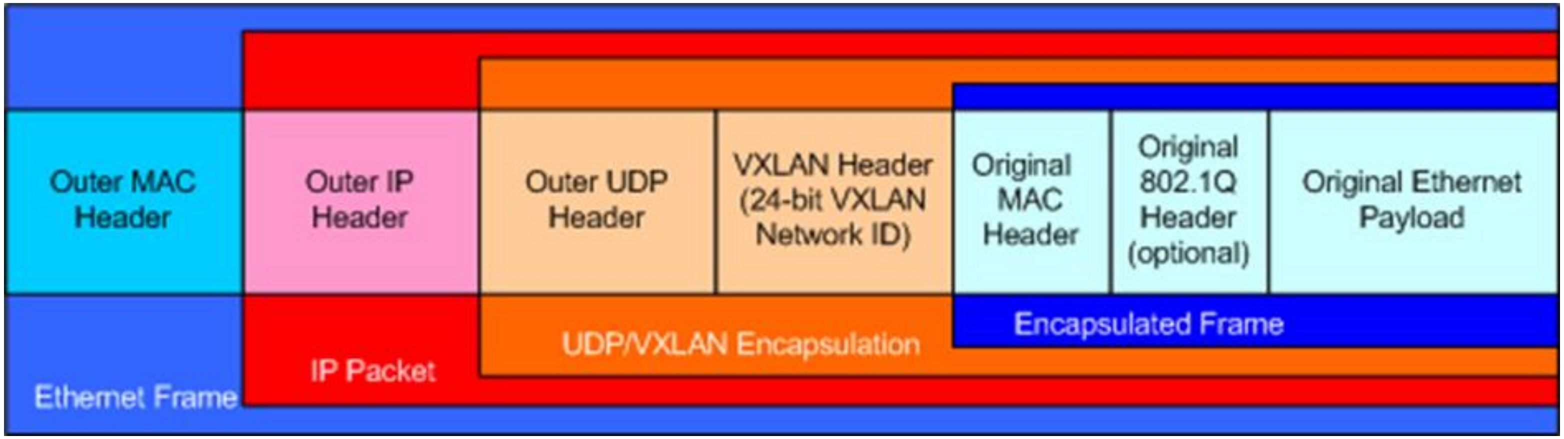
VXLAN (RFC 7348)

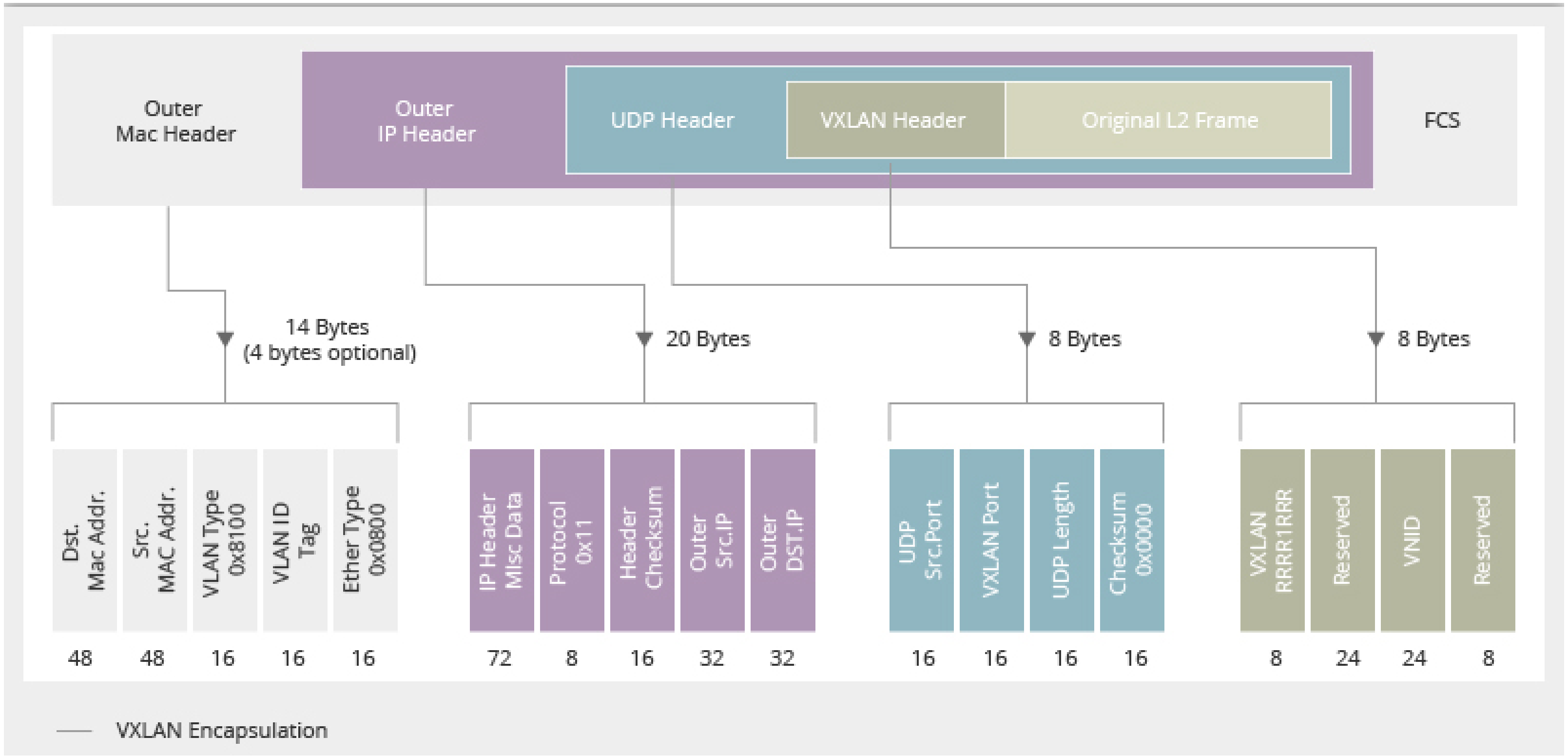
- Virtual eXtensible LAN (VXLAN) was originally proposed by Cisco and VMware to tunnel virtual layer 2 networks on a substrate layer 3 physical network



- VXLAN encapsulate packets in UDP tunnels with destination port number 4789
- In the shared L3 infrastructure, packets are identified by outer MAC addresses imposed by the infrastructure provider
- Tenants free to choose their own MAC addresses and VLAN IDs with no conflicts
- To avoid packet fragmentation in the shared infrastructure, it must support larger MTU values
- Encapsulation/decapsulation is performed at *VXLAN Tunnel End Points* (VTEPs)
- VXLAN ID allows to identify up to 2^{24} distinct virtual networks

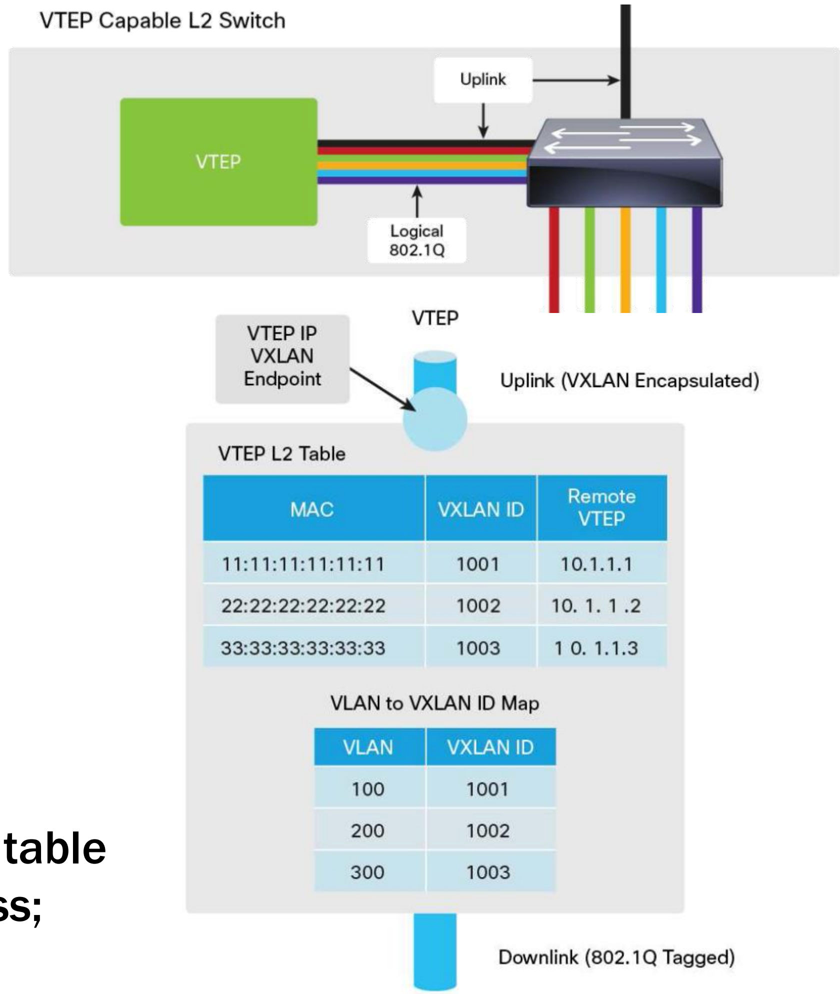
VxLAN packet





VXLAN: VTEP encapsulation & decapsulation

- A VTEP has two logical interfaces: an uplink and a downlink
  - Uplink to encapsulate
  - Downlink to decapsulate
- The VTEP can be located either on a physical switch (e.g. a ToR) or within the hypervisor's virtual switch
- The *outer IP destination* address is that assigned to the destination VTEP
- The *outer IP source* address is that assigned to the VTEP sending the frame
- Packets received from a tenant's VM on the downlink are mapped to a VXLAN ID
  - A lookup is then performed in the VTEP Layer 2 table using the VXLAN ID and destination MAC address; this lookup provides the IP address of the destination VTEP
- Packets received from a VTEP on the uplink are mapped from the VXLAN ID to an IEEE 802.1Q VLAN ID and sent as Ethernet frames on the downlink to the VM

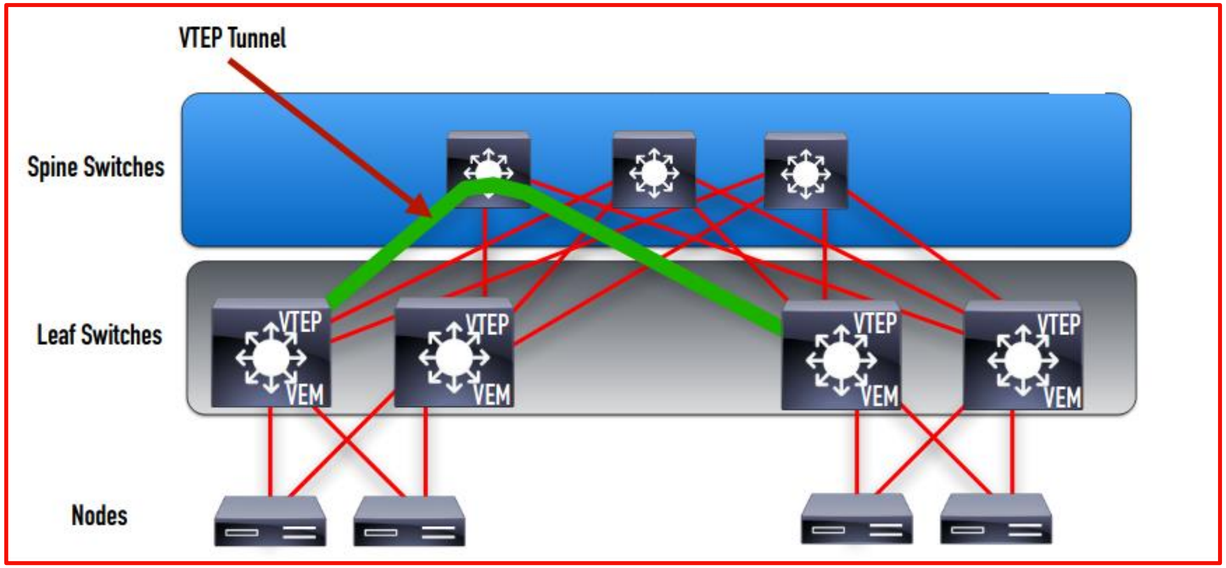


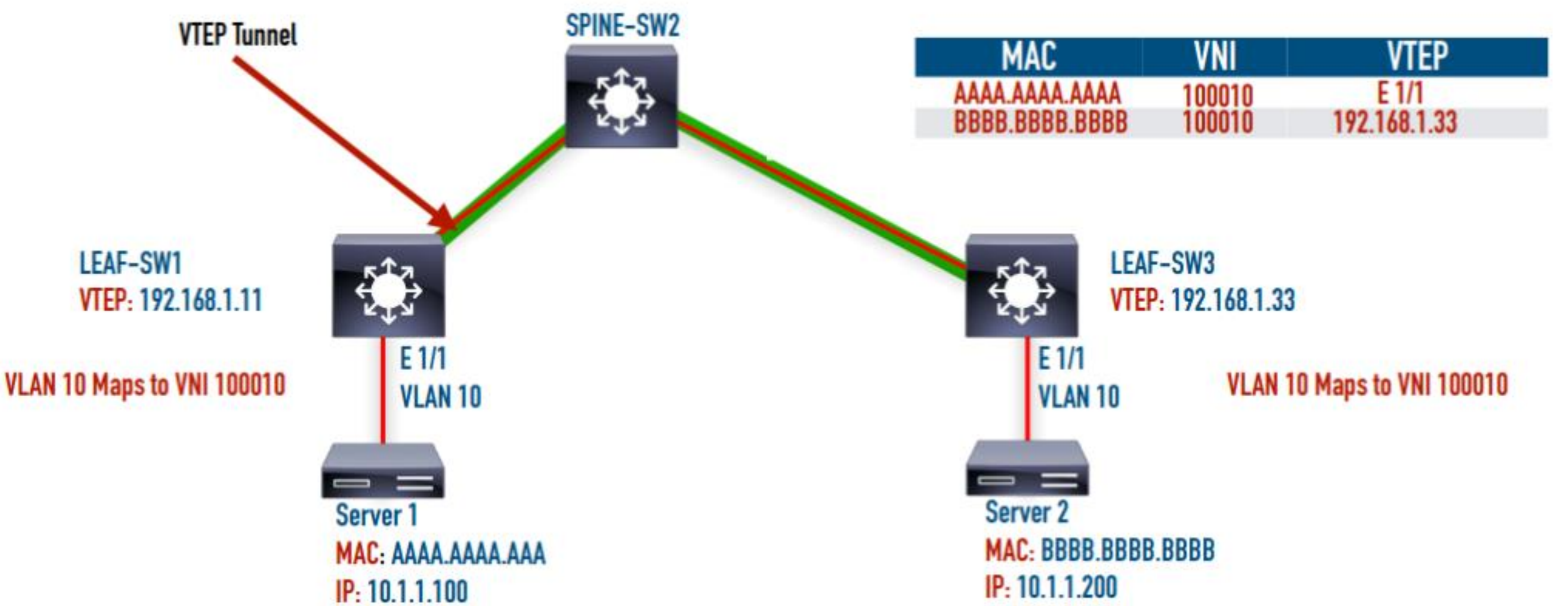
MAC	VXLAN ID	Remote VTEP
11:11:11:11:11:11	1001	10.1.1.1
22:22:22:22:22:22	1002	10. 1. 1 .2
33:33:33:33:33:33	1003	1 0. 1.1.3

VLAN	VXLAN ID
100	1001
200	1002
300	1003

VxLAN in the datacenter

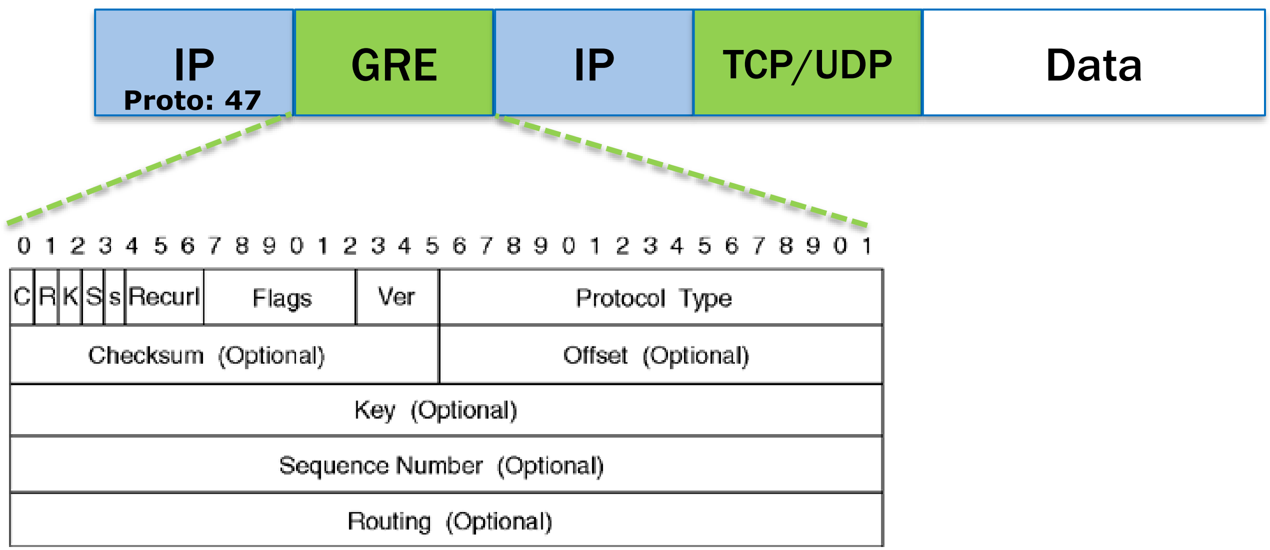
- A typical use of VxLAN in a leaf-spine datacenter network





GRE: Generic Routing Encapsulation (RFC 2784)

- Generic Routing Encapsulation (GRE) is a protocol that encapsulates packets in order to route other protocols over IP networks
- GRE was developed as a tunneling tool meant to carry any OSI Layer 3 protocol over an IP network
- GRE works by encapsulating an *inner packet* (*payload*) that needs to be delivered to a destination network inside an *outer IP packet*

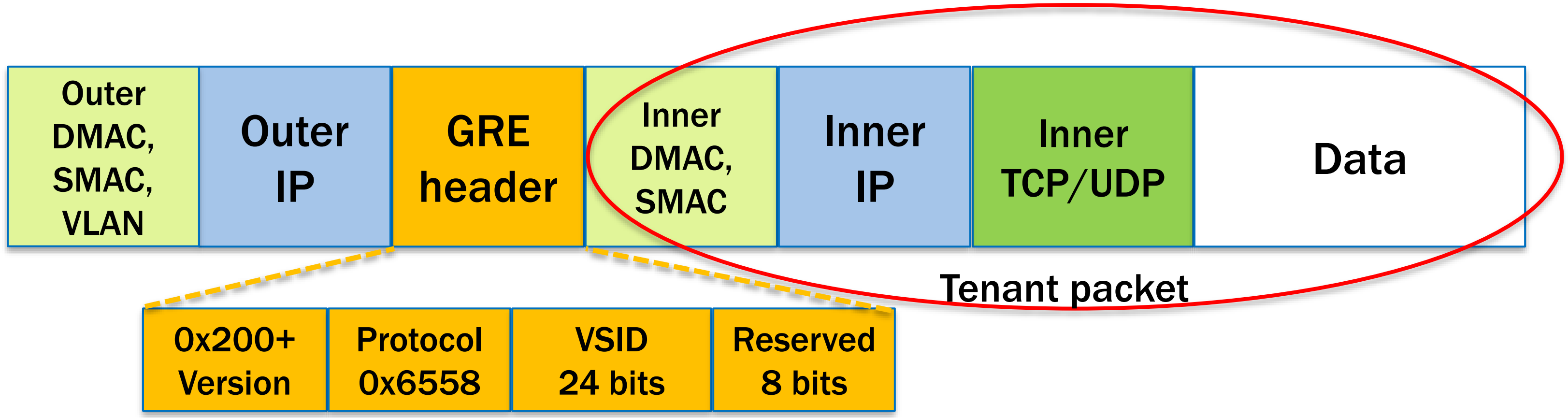


How GRE works

- GRE creates point-to-point connections as those used to create Virtual Private Networks (VPNs)
- IP routers along the way do not parse the payload
- Upon reaching the tunnel endpoint, GRE header is removed and the payload is forwarded along to its ultimate destination
- GRE tunneling can transport multicast and IPv6 traffic as payload but it does not use encryption like the IPsec Encapsulating Security Payload (ESP) as defined in RFC 2406

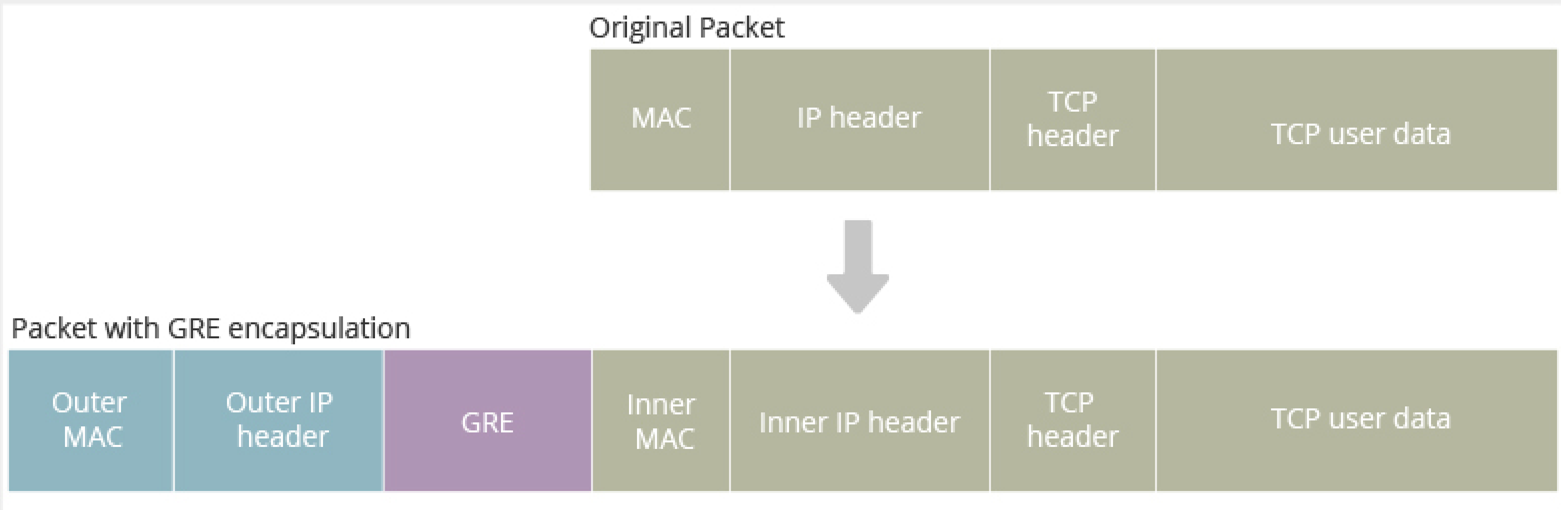
Network Virtualization using Generic Routing Encapsulation

- NVGRE (*Network Virtualization using Generic Routing Encapsulation*) is a network virtualization method that uses encapsulation and tunneling to create large numbers of virtual LANs for subnets that can extend across layer 2 and 3



- VSID is a 24 bits Virtual Segment Identifier
- The inner packet does not contain a VLAN ID as in VXLAN
  - If a tenant needs multiple VLANs, it must be assigned different VSIDs
- Encapsulation/decapsulation is performed by *Network Virtual Endpoints* (NVEs)
- Which NVE is associated to a given DMAC is through mechanisms not in NVGRE specs

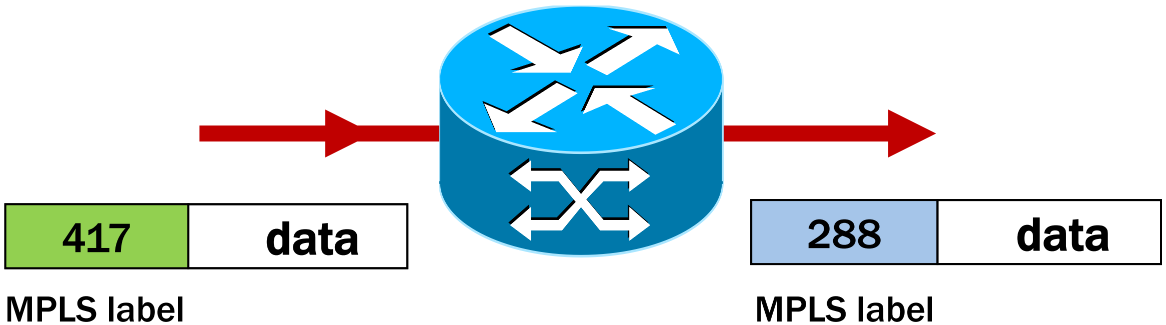
NVGRE Encapsulation

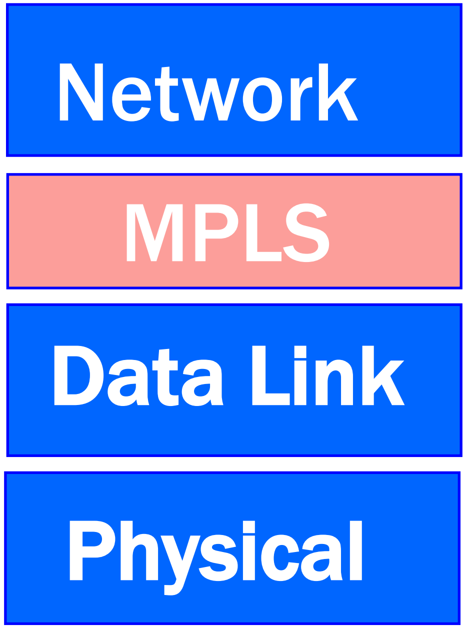


NVGRE Encapsulation

MPLS

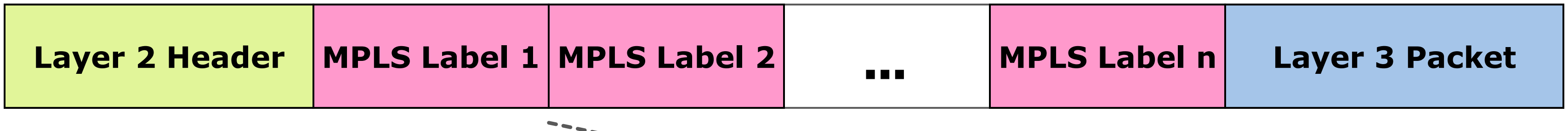
- A "Layer 2.5" tunneling protocol based on ATM-like notion of "label swapping"
  - A simple way of labeling each network layer packet
  - Independent of Link Layer
  - Independent of Network Layer
- Used to set up "Label-switched paths" (LSP), similar to ATM PVCs, to carry L3 packets (e.g. IP datagrams) on virtual circuits
- RFC 3031: Multiprotocol Label Switching Architecture
- An MPLS switch forwards packets according to labels





MPLS encapsulation

- RFC 3032. MPLS Label Stack Encoding

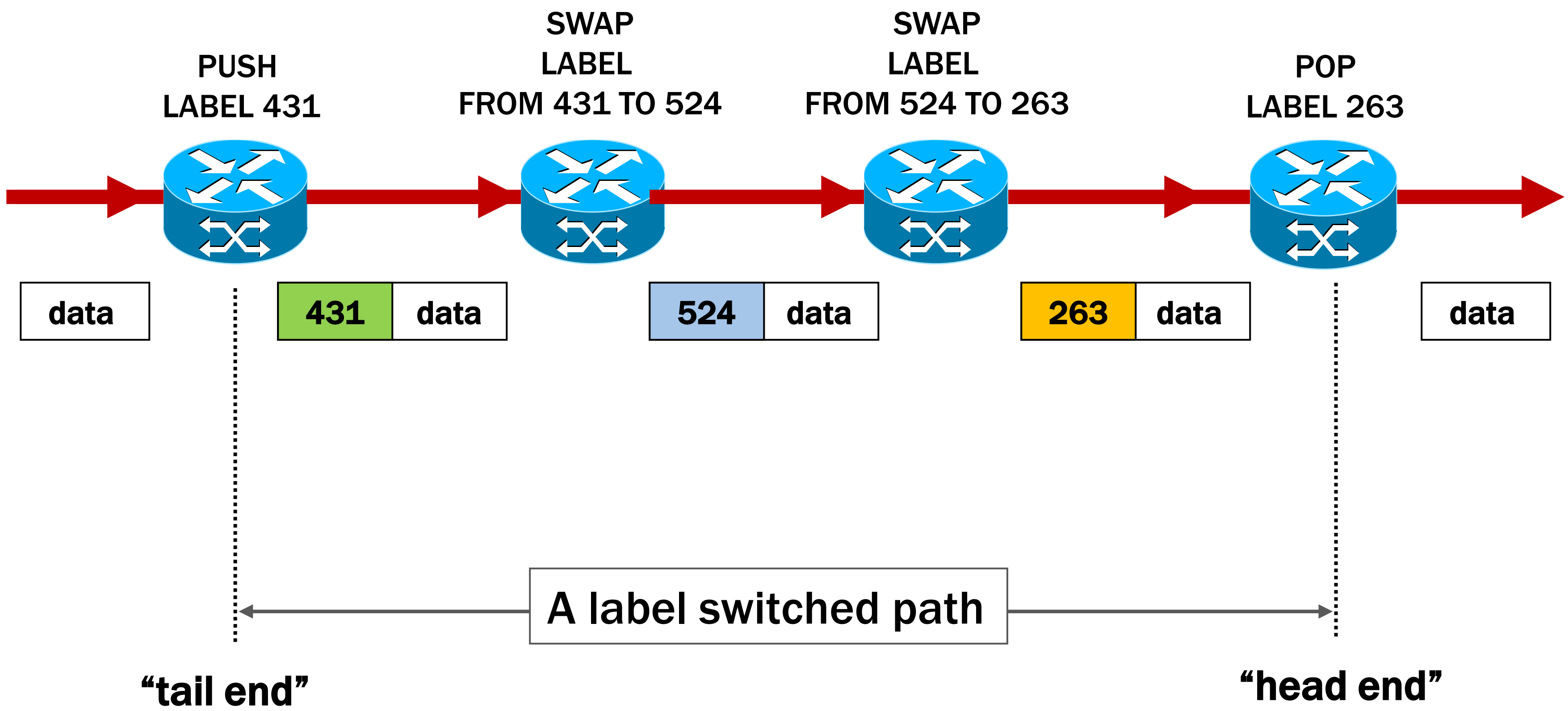



```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Label                  | Exp |S|       TTL     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- **Label: Label Value, 20 bits**
- **Exp: Experimental, 3 bits**
- **S: Bottom of Stack, 1 bit**
- **TTL: Time to Live, 8 bits**

LSP: Label Switched Path

- Also called an MPLS tunnel: payloads (*data*) are not inspected inside an LSP
- MPLS can carry any traffic, not only IP



MPLS label distribution

- Label distribution protocols are needed to
  1. create labels associated to an LSP
  2. distribute bindings to neighbors
  3. maintain consistent label swapping tables
- Two different approaches
  - "Piggyback" label information on top of existing IP routing protocol
    - Allows only traditional destination-based, hop-by-hop forwarding paths
  - Create new label distribution protocol(s)
    - Not limited to destination-based, hop-by-hop forwarding paths
    - E.g. LDP (IETF) and TDP (Cisco proprietary)
- Combine resource reservation with label distribution; two approaches:
  - Add label distribution and explicit routes to a resource reservation protocol
    - RSVP-TE
  - Add explicit routes and resource reservation to a label distribution protocol
    - CR-LDP