

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

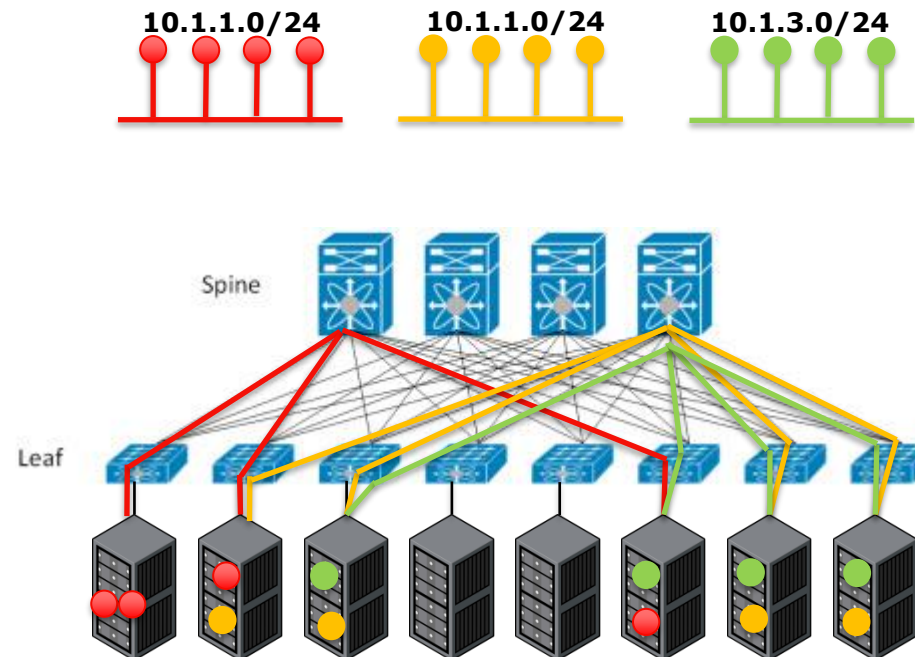


- ▶ Multitenancy
- ▶ Virtual networking techniques in a datacenter
- ▶ Tunneling protocols

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 independence
 - ▶ Challenges: address collisions, partitioning, mapping, ...

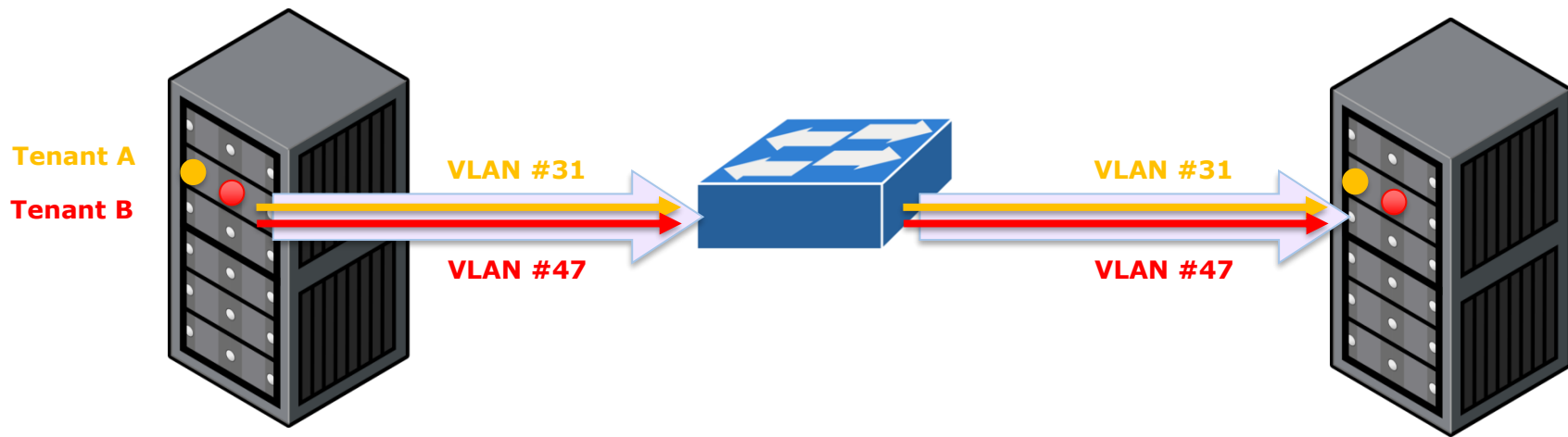


- ▶ Network virtualization techniques allow to map logical tenant networks onto a common shared physical substrate
- ▶ Most common network virtualization approaches are based on traffic encapsulation (a.k.a. *tunneling*) and creation of *overlays*
- ▶ VLANs is a form of layer 2 encapsulation natively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

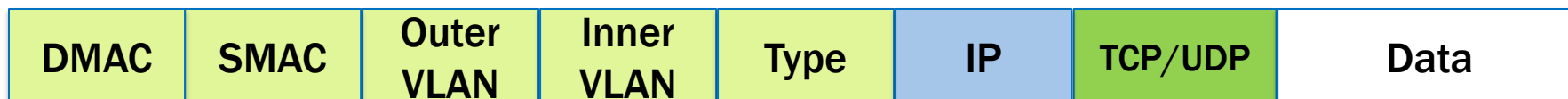
Tenant isolation via VLANs



- ▶ 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



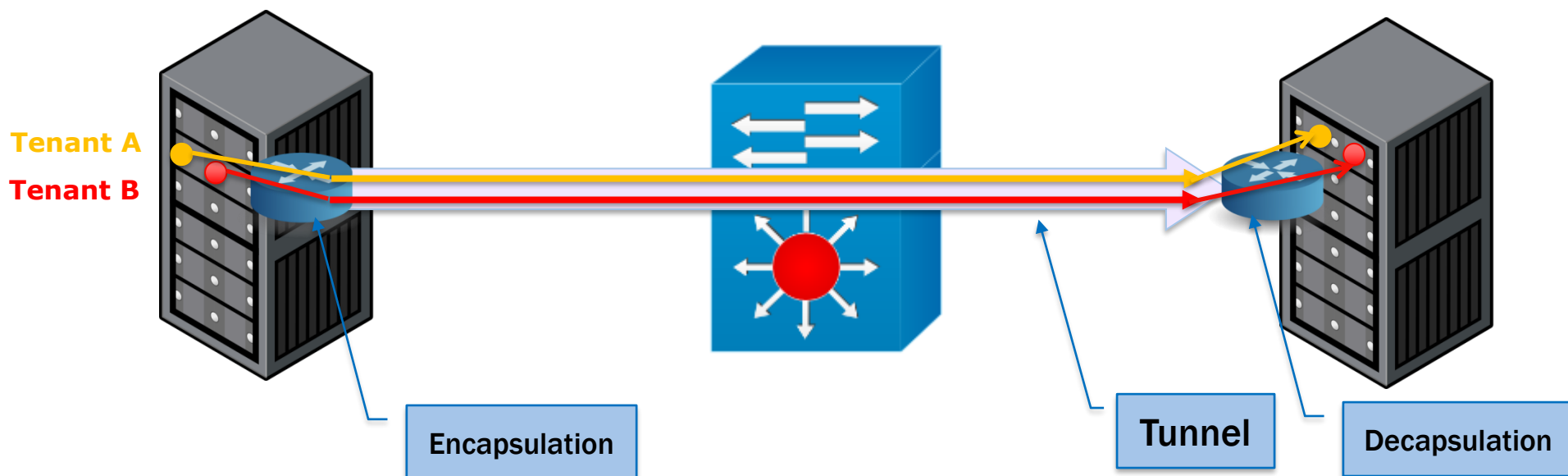
- ▶ IEEE 802.1ad allows to wrap an 802.1q VLAN-tagged packet with an outer VLAN tag (*Q-in-Q*)
- ▶ This technique is used to carry proprietary VLAN-tagged traffic on a shared service provider network where the outer 12-bit VLAN ID is used to identify the customer traffic in the provider network
 - ▶ 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



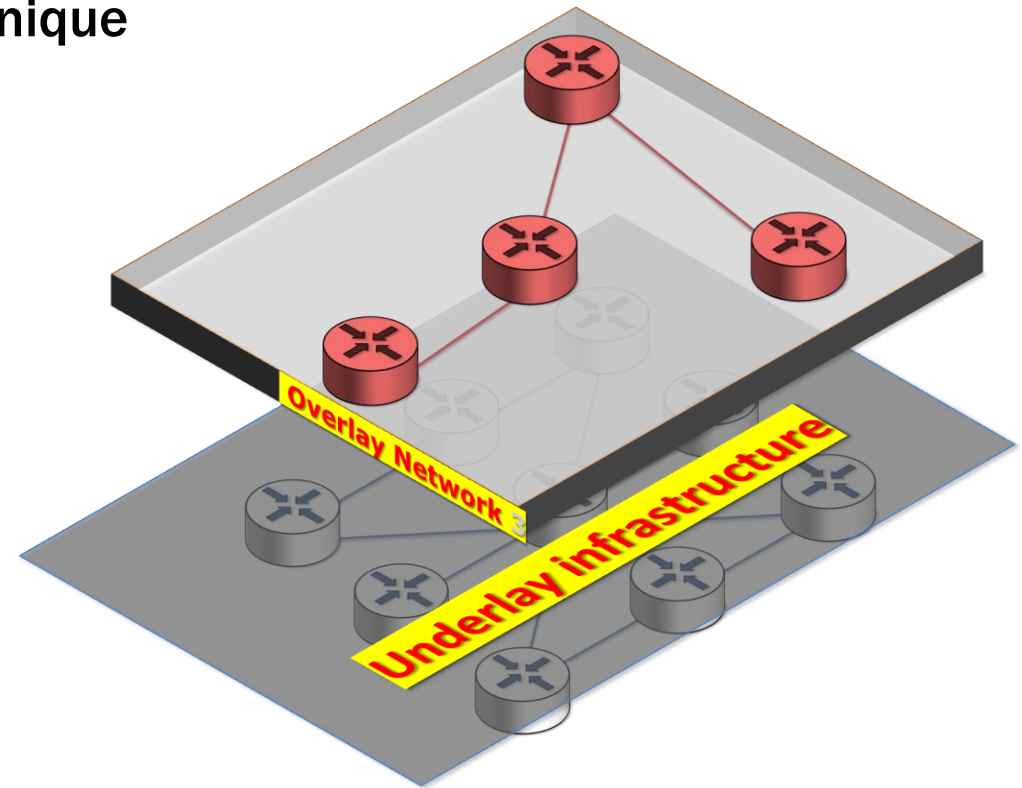
Network Virtualization using encapsulation



- ▶ 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



- ▶ 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

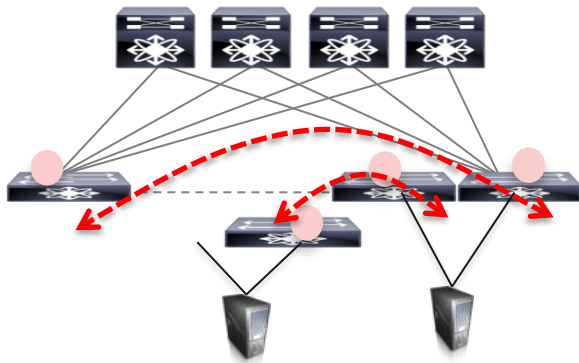


Types of overlays



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

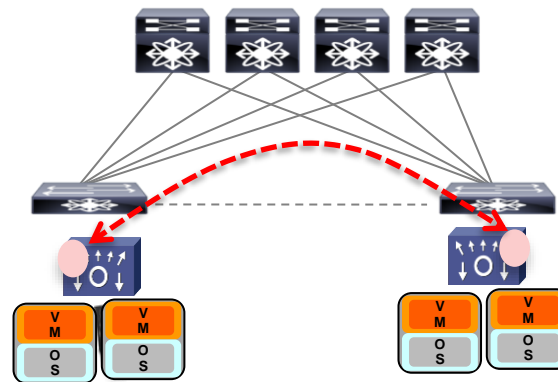
Network Overlays



Physical

Physical

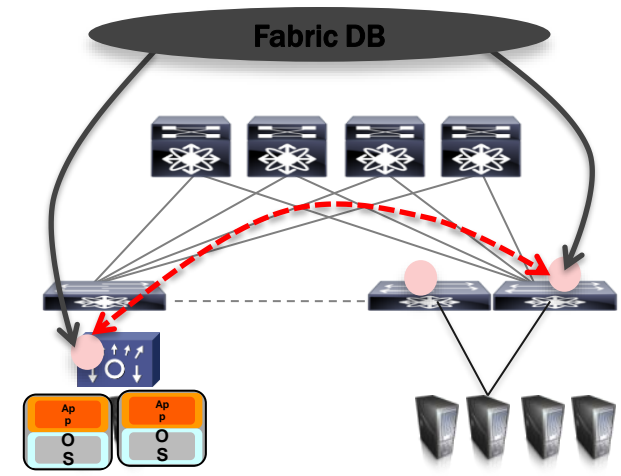
Host Overlays



Virtual

Virtual

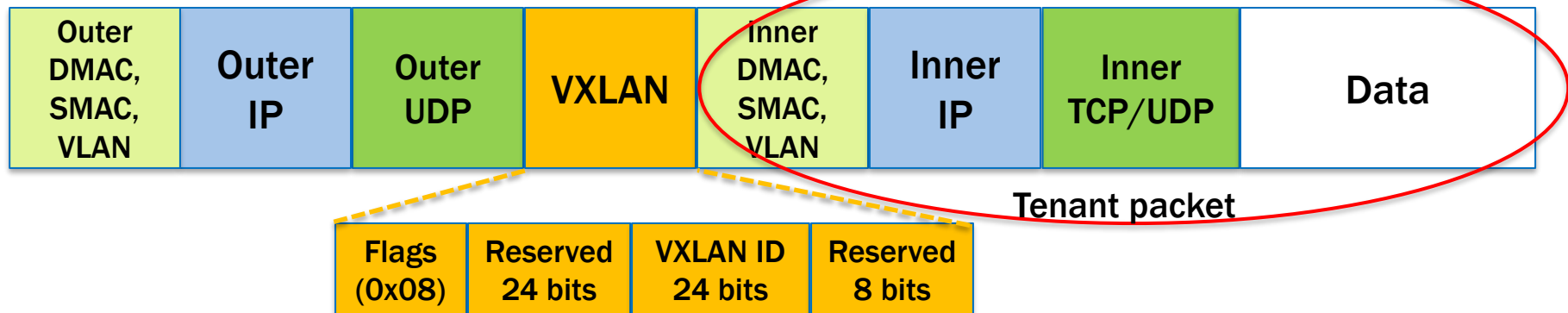
Integrated Overlays



Virtual

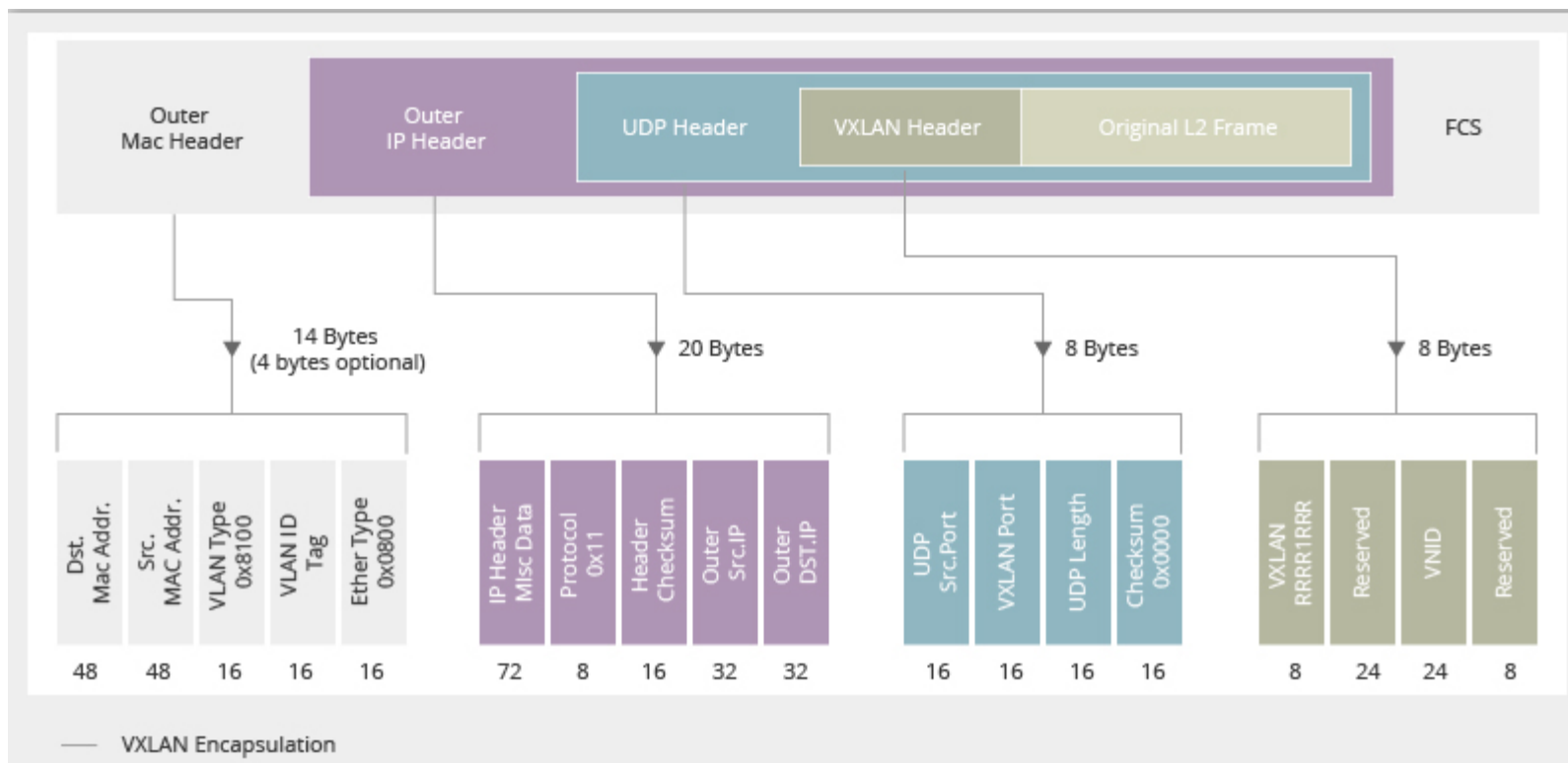
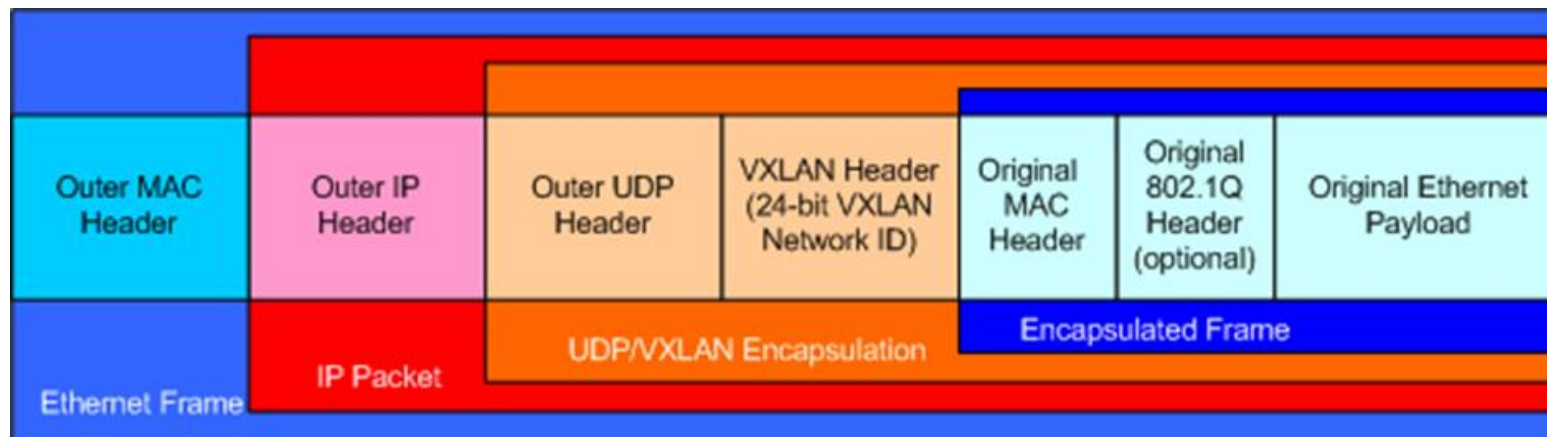
Physical

- ▶ 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

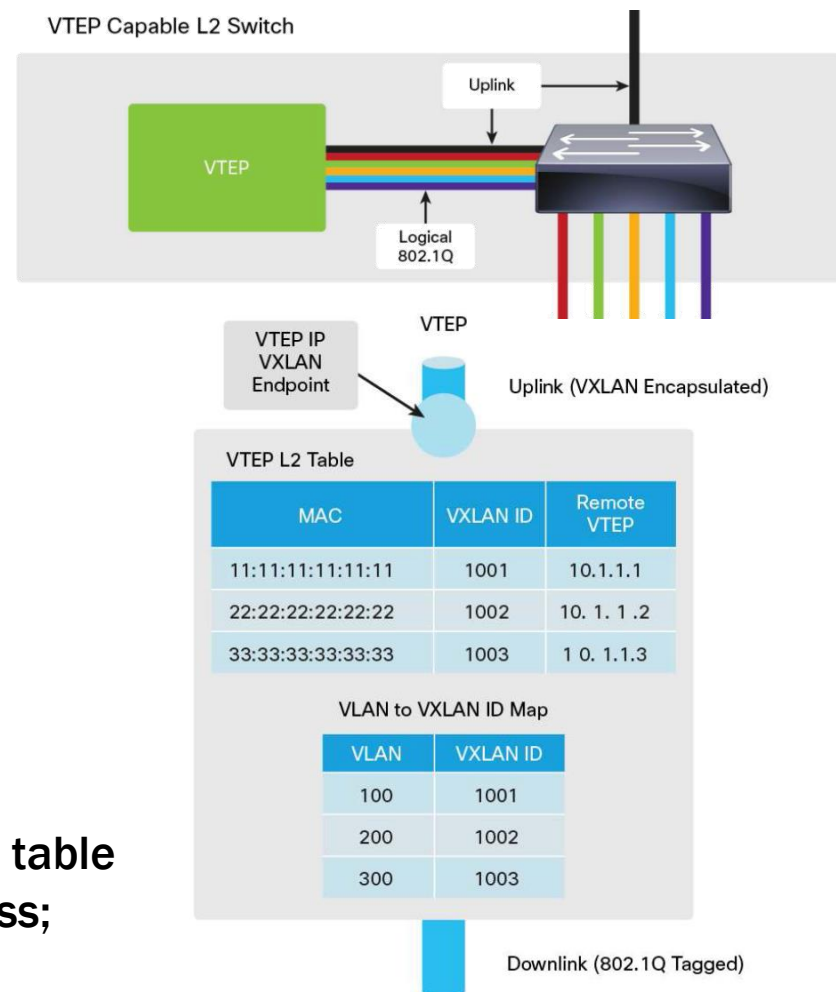
VxLAN packet



VXLAN: VTEP encapsulation & decapsulation



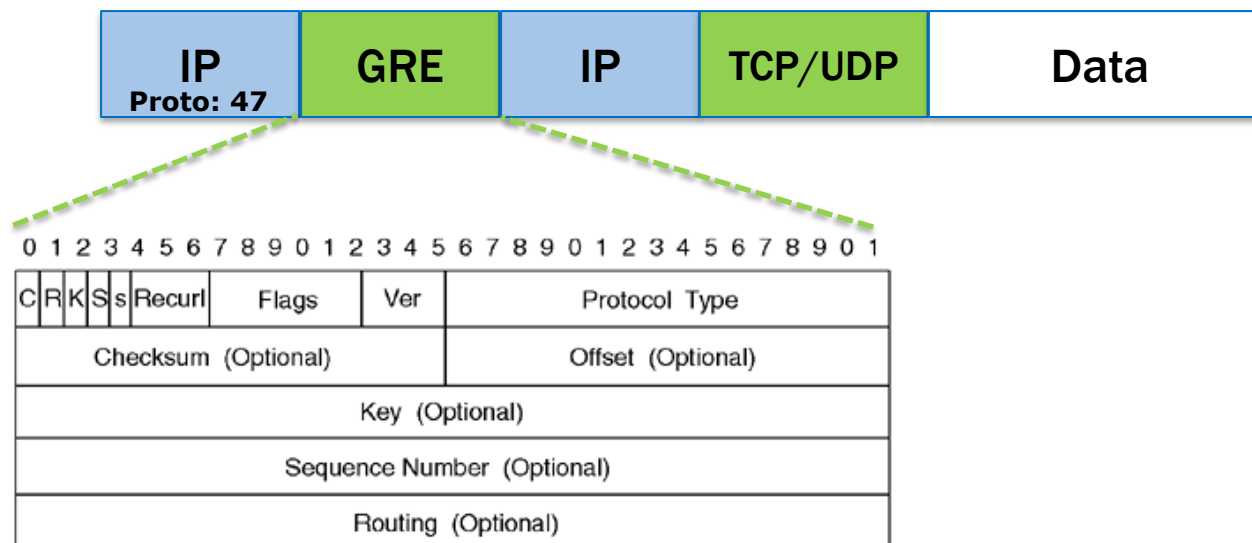
- ▶ A VTEP has two logical interfaces: an uplink and a downlink
 - ▶ Uplink to encapsulate
 - ▶ Downlink to decapsulate
- ▶ The VTEP can be located either on a physical switch (e.g. a ToR) or within the hypervisor's virtual switch
- ▶ The *outer IP destination* address is that assigned to the destination 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



GRE: Generic Routing Encapsulation (RFC 2784)

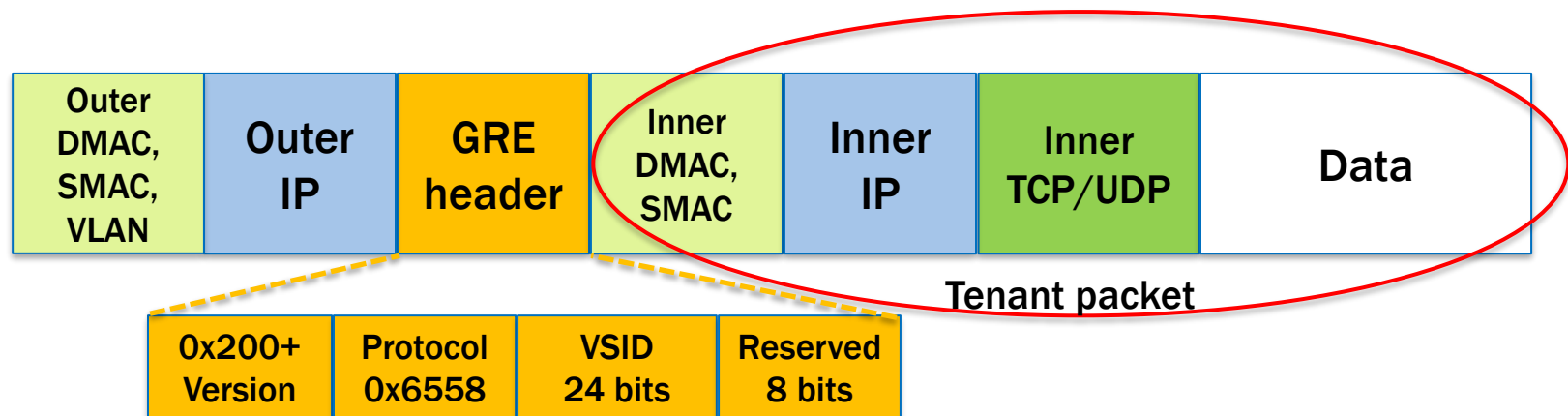


- ▶ Generic Routing Encapsulation (GRE) is a protocol that encapsulates packets in order to route other protocols over IP networks
- ▶ GRE was developed as a tunneling tool meant to carry any OSI Layer 3 protocol over an IP network
- ▶ GRE works by encapsulating an *inner packet (payload)* that needs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

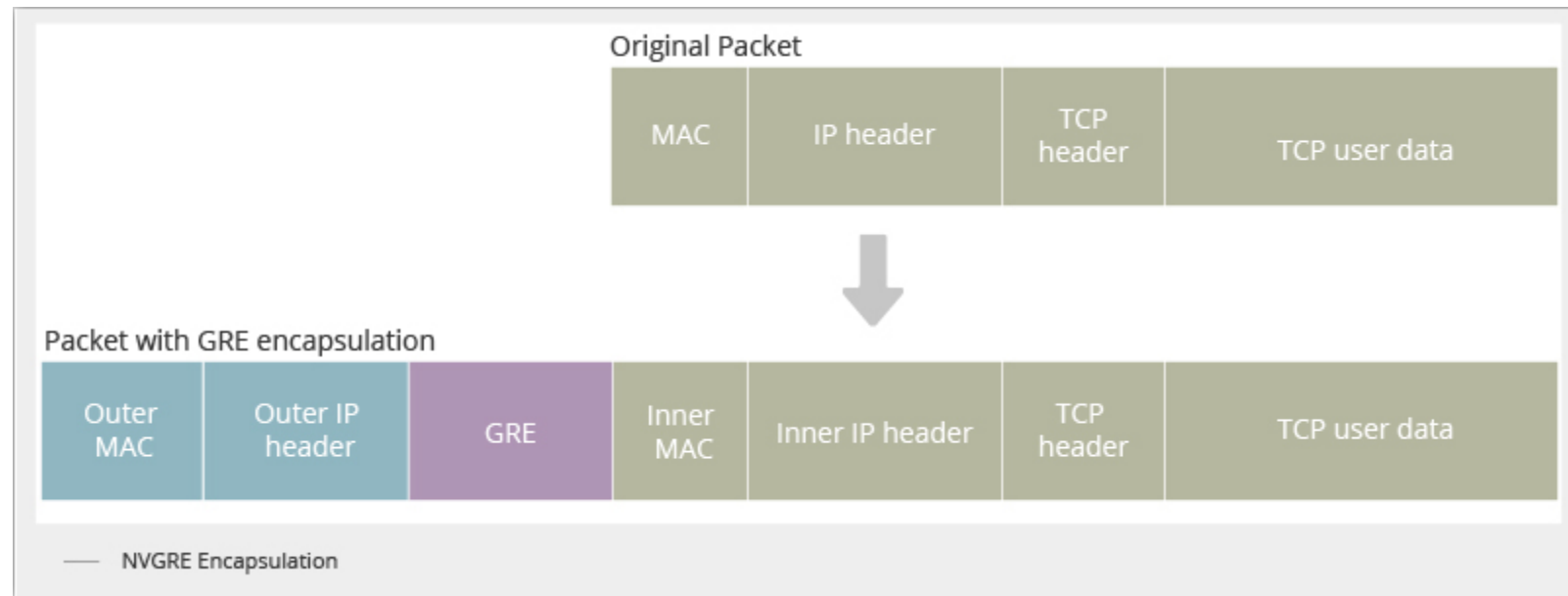


- ▶ GRE creates point-to-point connections as those used to create Virtual Private Networks (VPNs)
- ▶ IP routers along the way do not parse the payload
- ▶ Upon reaching the tunnel endpoint, GRE header is removed and the payload is forwarded along to its ultimate destination
- ▶ GRE tunneling can transport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

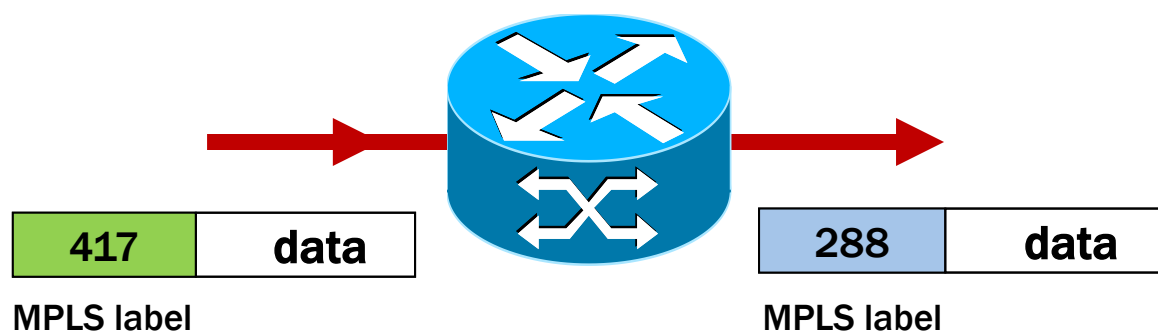
- ▶ NVGRE (*Network Virtualization using Generic Routing Encapsulation*) is a network virtualization method that uses encapsulation and tunneling to create large numbers of virtual LANs for subnets that can extend across layer 2 and 3



- ▶ 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



- ▶ 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



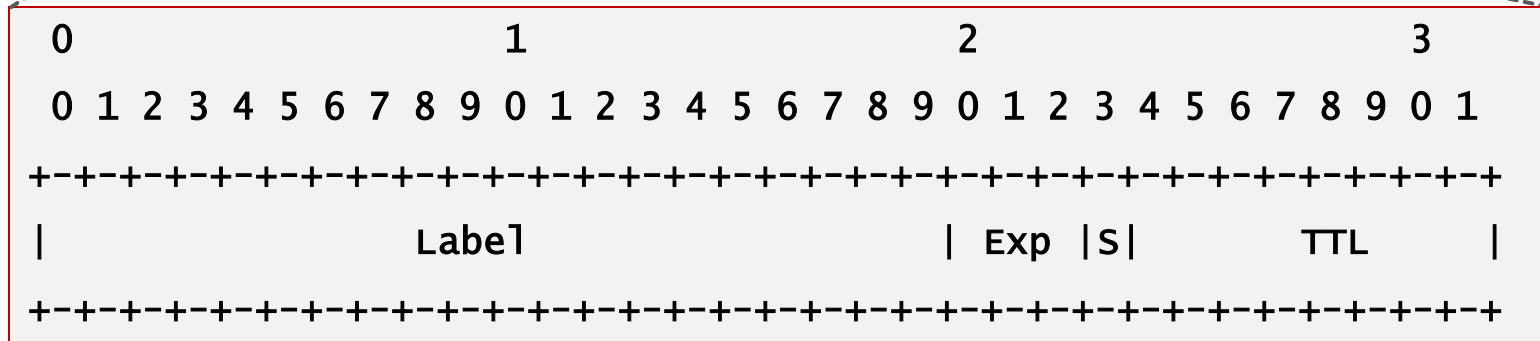
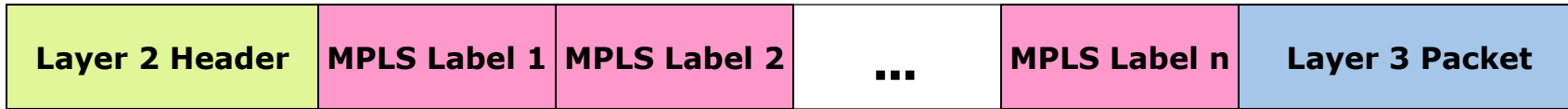
Network

MPLS

Data Link

Physical

► RFC 3032. MPLS Label Stack Encoding

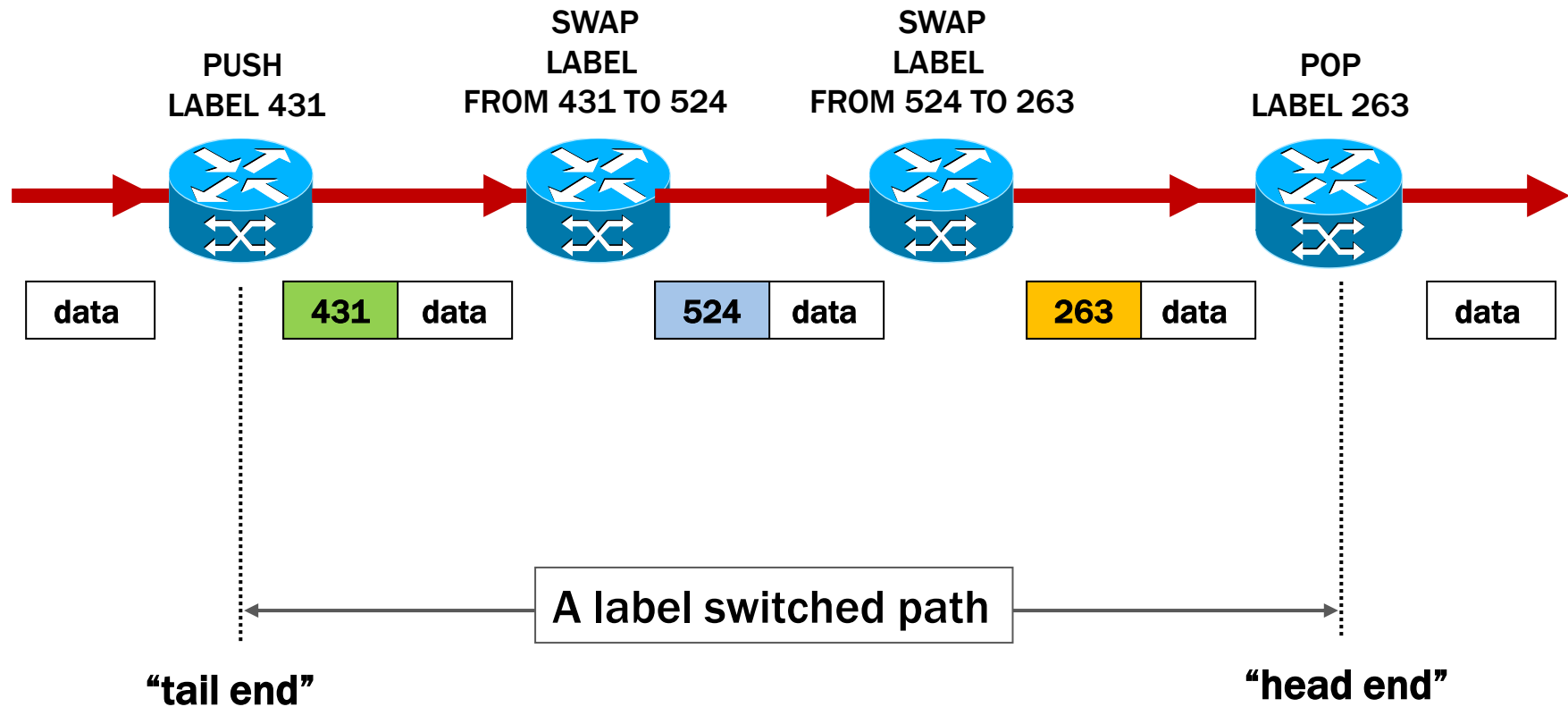


- **Label:** Label Value, 20 bits
- **Exp:** Experimental, 3 bits
- **S:** Bottom of Stack, 1 bit
- **TTL:** Time to Live, 8 bits

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



- ▶ 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