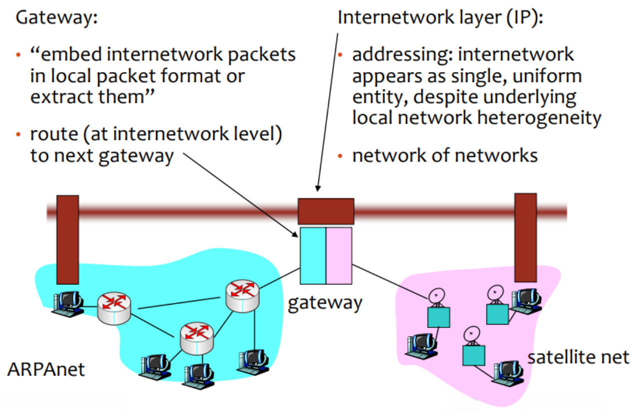
Multi-Protocol Label Switching (MPLS)

Index

Virtualization of networks

- Virtualization of resources is a powerful abstraction in systems engineering
- Computing examples: virtual memory, virtual devices
- Layering of abstractions
  - Don't sweat the details of the lower layer, only deal with lower layers abstractly
- The Internet: virtualizing networks

1974: Multiple Unconnected Nets	...differing in
ARPAnet	Addressing conventions
Data-over-cable networks	Packet formats
Packet satellite network (Aloha)	Error recovery mechanisms
Packet radio network	Routing



Cerf & Kahn's Internetwork Architecture

- Two layers of addressing: local network and internetwork
- New layer (IP) makes everything homogeneous at internetwork layer
- Underlying local network technology (Ethernet, satellite, ATM, **MPLS**) becomes "invisible" at internetwork layer. Looks like a link layer technology to IP!

Virtual Circuit vs. Datagram Networks

Virtual Circuit Networks	Datagram Networks

Virtual Circuit Networks	**Datagram Networks**
• VC establishment prior to data transmission, first packets delayed • All packets follow the same path • In-order delivery • Failures must be explicitly handled • Exact matching of VC identifier • Packets contain VC identifier • Routers maintain per-VC info • Easy to combine with resource reservation • Traffic engineering easy	• No VC establishment, data may be sent immediately • Packets forwarded independently • Packets may be reordered in transit • Robust to link or node failures • Longest prefix matching of addresses • Packets contain source & destination addresses • Routers maintain only aggregate destination info • Resource reservation hard, requires additional protocols • Traffic engineering harder

Multiprotocol Label Switching (MPLS)

- Initial goal: speed up IP forwarding by forwarding based on a fixed length label (instead of IP address)
    - Borrowing ideas from Virtual Circuit (VC) approach
    - IP datagram still keeps IP address!

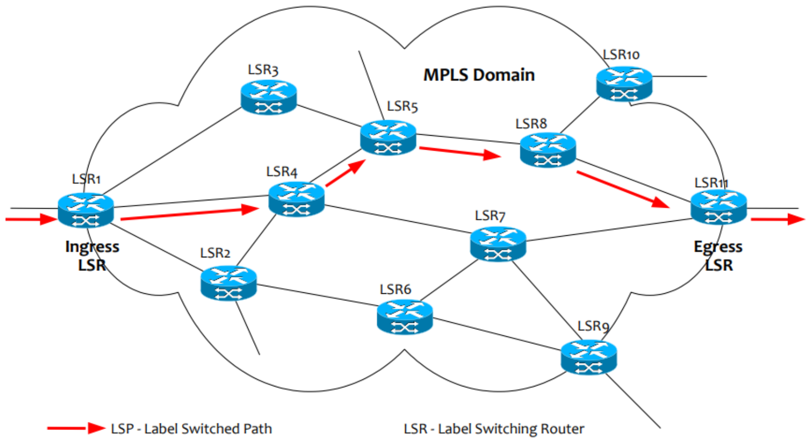
Objectives and Advantages

1. Flow detection and routing based on labels (simpler and faster decision process)
    - Greater scalability
    - Better performance (main reason at the beginning, not significant nowadays...)
    - Separation of Routing and Forwarding
        - Routing: how to send packets from source to destination - global action
        - Forwarding: transfer a packet from an entry port to an exit port - local action
2. Enable establishing VPNs across telecom operator's network
    - Interconnection simplicity for clients that want to use different sites as if they were a single network
3. Enable traffic engineering
    - Allows going beyond the routing protocols when deciding the path for a packet

MPLS supports

- Integration with routing protocols (BGP, OSPF, etc.), unicast, multicast, source routing, route pinning, QoS

MPLS - Global View

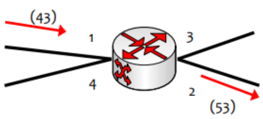


MPLS - Architecture

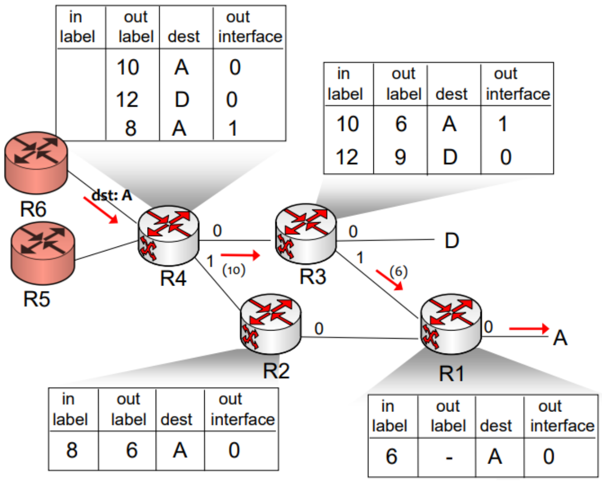
- Switching based on labels (labels change at every node - label-swapping)
- Labels transported in frames or shim header
- **MPLS node**
  - Supports MPLS, forwards based on labels, and supports one or more L3 routing protocols
- **Label Switching Routers (LSR)**
  - MPLS Nodes capable of forwarding native L3 packets
- **Edge routers**
  - MPLS nodes at the border of MPLS domains
  - Ingress
    - Decides Forwarding Equivalence Class (FEC)
    - Transmits packet with label corresponding to FEC
  - Egress
    - Removes label*
- MPLS Routers
  - Search the label on the Label Information Base (LIB)
  - New label - Next-Hop Label Forwarding Entry (NHLFE)
  - Transmit packet in out interface with new label
- Requires a mechanism for label distribution

Itf In	Label In	Itf Out	Label Out
1	43	2	53
4	4	1	16

LIB

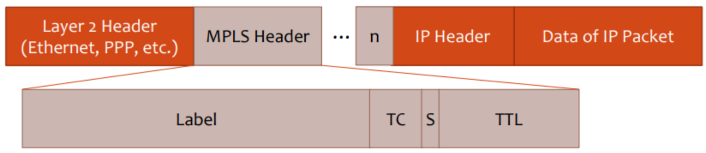


MPLS Forwarding Tables



MPLS - Labels

Shim Header	Label
• Generic	• Small
• "Layer 2.5"	• Fixed size
• Stackable	• Local meaning



Label: 20 bits
TC: 3 bits (traffic class information)
S: 1 bit (1 means bottom of stack)
TTL: 8 bits

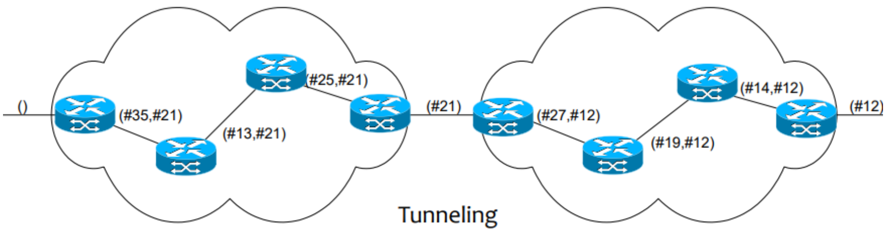
No need for shim header when the layer 2 technology supports labels (e.g., with ATM maps to VPI/VCI)

MPLS - Need for a TTL field

- Nodes after the MPLS domain must see the same TTL as if MPLS were not used
- TTL in shim header is set from the IP header
- TTL in the shim header is decremented in each MPLS node the packet goes through
- When removing the label, TTL in the IP header should be set to the value in the shim header

Label stacking

- Non-hierarchical
  - Different labels added at the ingress LSR
  - Each router removes a label from the stack
  - More overhead, but even faster forwarding performance
- Hierarchical (tunneling)
  - Intra-domain and inter-domain



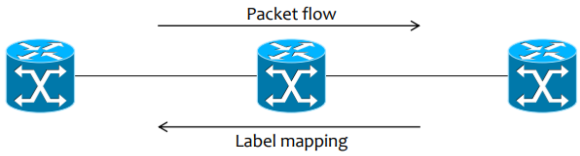
Forwarding Equivalence Class (FEC)

- Subset of packets handled similarly by the router (same Next Hop, interface, treatment)
- The FEC is determined only at the ingress LSR and determines the output label at that router
- Criteria for setting the FEC
  - IP prefix, aggregating

- Egress edge router of domain
- By flow, end-to-end
- QoS / Traffic Engineering
- Other criteria

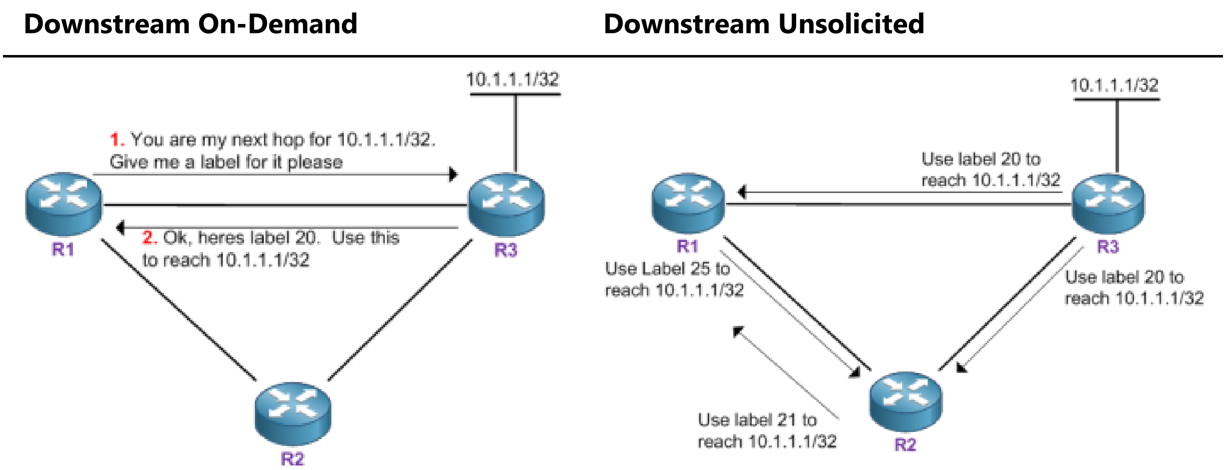
Label distribution

- Routing information used to distribute labels
  - Piggyback on routing protocols
- MPLS nodes
  - Receive mapping from nodes "down" the path
  - Allocate and distribute labels for nodes "up" the path



Label Distribution Protocol (LDP)

- Assign labels to routing table entries and define LDP adjacencies (Hello, etc.)
- Distribute info whether node
  - Is an egress LSR
  - Has an exit label for the FEC
- Two modes for label distribution:



- Two modes for label control
  - <u>Ordered Control</u>
    - LSR advertises FEC if it is the egress LSR for the FEC or has received an advertisement from the donwstream peer
    - Non-egress LSRs must wait for their downstream peers before advertising the FEC
  - <u>Independent Control</u>
    - LSRs advertise independently
    - May lead to (temporary) blackholing of some traffic
- Applicable when FECs are associated with destination address
- Alternatives to LDP
  - CR-LDP (Constraint-based Routing LDP)
  - RSVP-TE (Extensions to RSVP for LSP Tunnels)

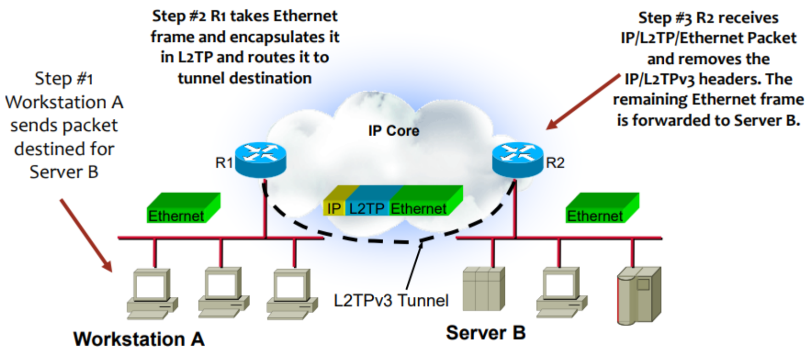
Explicit Routing in MPLS

- Two options for route selection: Hop by hop routing or Explicit routing
- Explicit Routing (Source Routing) is a very powerful technique
  - With pure datagram routing, overhead of carrying complete explicit route is prohibitive
  - MPLS allows explicit route to be carried only at the time the LSP is setup, not with each packet
  - MPLS makes explicit routing practical ☺
- In an explicitly routed LSP
  - LSP next hop is not chosen by the local node
  - It is selected by a single node, usually the ingress
- The sequence of LSRs may be chosen by
  - Configuration (administrator or centralized server)
  - An algorithm (e.g., the ingress node may use topological information learned from a link state routing protocol)

VPNs

Motivation

Layer2 Example



Overlay Model

- Service Provider provides PtP links to customer routers on other sites
- Connectivity
  - Fully connected
  - Hub-and-spoke

Limitations of Overlay

- Customers need to manage the backbones
- Mapping between Layer 2 QoS and IP QoS
- Scaling problems

The Peer Model

- Service provider and customer exchange Layer 3 routing information
  - Provider relays data between customer sites using best path

- Goal: provide a large-scale VPN service
- Key technologies
  - Constrained distribution of routing info (do not mix routes from different customers)
  - Multiple forwarding tables (one per VPN)
  - VPN-IP addresses (combine VPN info and IP prefix)
  - MPLS switching

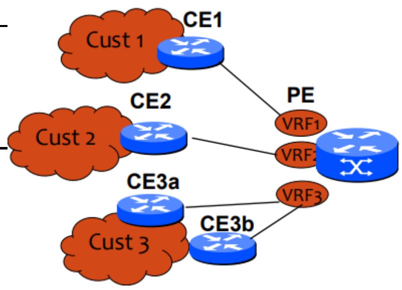
Layer 2 vs Layer 3 VPNs

Layer 2 VPNs	**Layer 3 VPNs**
• Provider devices forward customer packets based on Layer 2 information • Tunnels, circuits, LSPs, MAC address • "Pseudo-wire" concept • VPLS - Virtual Private LAN Service Using (LDP) Signaling	• Provider devices forward customer packets based on Layer 3 information (e.g., IP) • Service Provider involvement in routing • MPLS/BGP VPNs, GRE, virtual router approaches

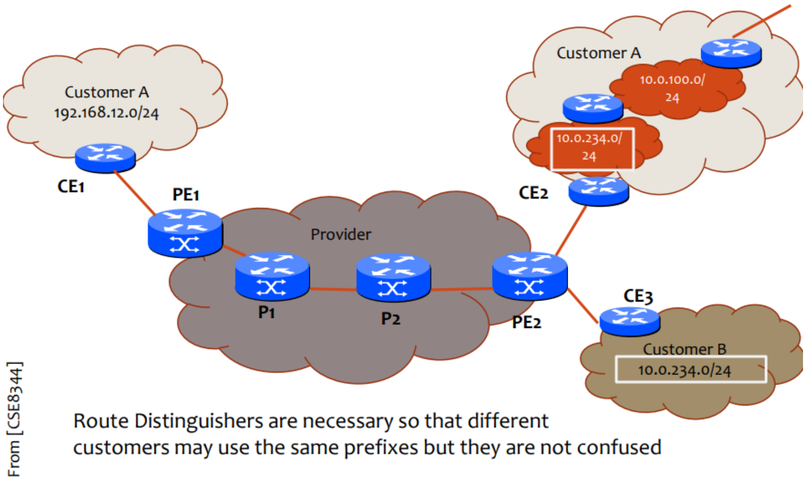
> The following discussion will concern Layer 3 VPNs

Terminology

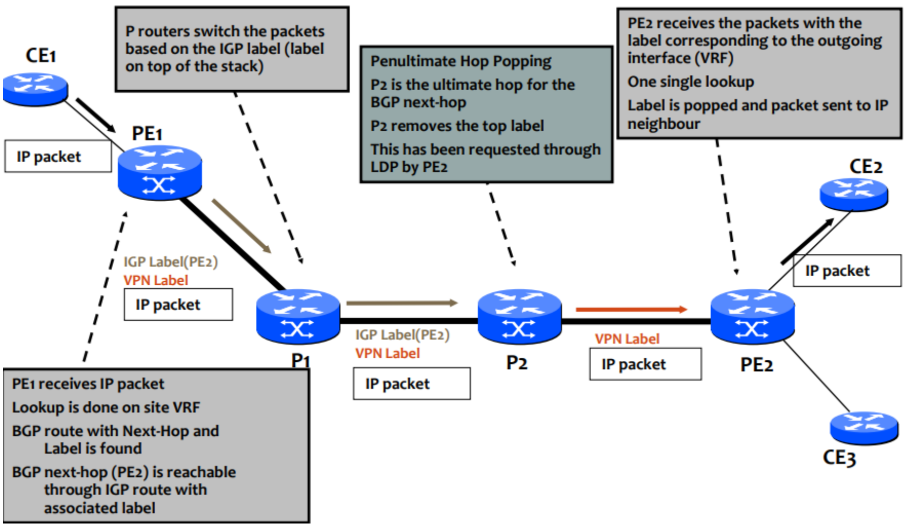
CE router 	• Customer Edge router
PE router 	• Provider Edge router • Interfaces to CE routers • Is a Label Edge Router (LER)
P router 	• Provider (core) router, without knowledge of VPN or customer routes • Is a Label Switching Router (LSR)
Route Distinguisher (aka route target)	• Attribute of each route used to uniquely identify prefixes among VPNs (64 bits)
VPN-IPv4 addresses	• Address including the 64 bits Route Distinguisher and the 32 bits IP address
VRF	• VPN Routing and Forwarding Instance • Routing table and FIB table



Network Example



Forwarding Example



Connection Model

- The VPN backbone is composed by MPLS LSRs
  - PE routers (edge LSRs)
  - P routers (core LSRs)
- PE routers are faced to CE routers and distribute VPN information through MBGP to other PE routers
- P routers do not run MBGP and do not have any knowledge of VPNs
  - Complexity kept at the edges
- P and PE routers share a common IGP
  - Routing for destinations in the provider network
- PE and CE routers exchange routing information through some means
  - eBGP, OSPF, RIP, static routing
  - Protocol may be different in different sites of the same customer
- CE routers run standard routing software

Routing

- Routes PE receives from CE are installed in the appropriate VRF
  - Assigned according to the incoming interface

- VRF necessary to segregate customers
- By using separate VRFs, addresses need NOT be unique among VPNs
    - Useful with private addressing
- Routes PE receives through the backbone IGP are installed in the global routing table
    - Routes in the provider network

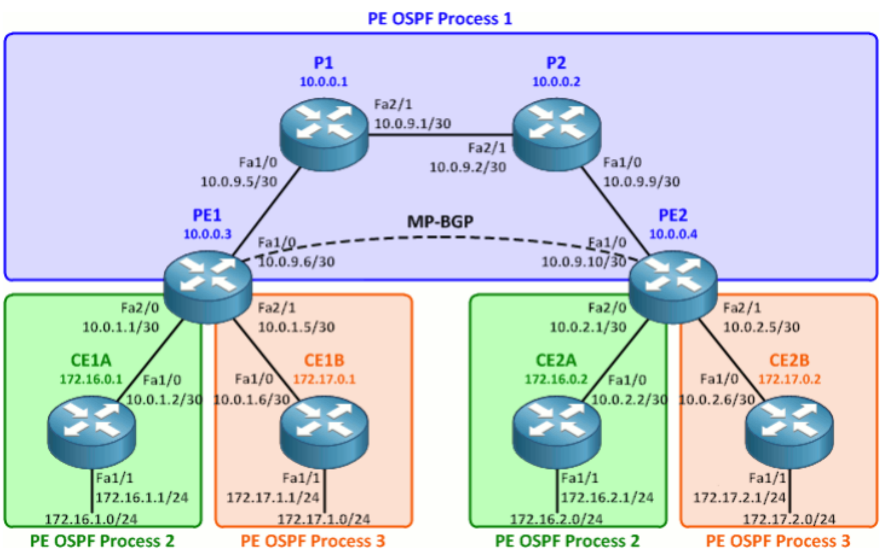
Forwarding

- PE routers use MBGP to exchange reachability information and learn the BGP Next-Hop
- PE and P routers use IGP to establish the IGP NextHop towards the BGP Next-Hop
    - The BGP Next-Hop is the egress PE router
- Labels corresponding to BGP Next-Hops are distributed through LDP (hop-by-hop)
- Label Stack is used for packet forwarding
    - Top (outer) label indicates IGP Next-Hop
    - Bottom (inner) label indicates outgoing interface or VRF
- The upstream LDP peer of the BGP next-hop (PE router) will pop the first level label
    - Penultimate Hop Popping
    - Avoid double processing at the egress PE Router
- The egress PE router will forward the packet based on the bottom label
    - The only one it receives
    - Determines the outgoing VPN and interface

Scalability

- Existing BGP techniques can be used to scale the route distribution (e.g, use of route reflectors)
- Each edge router needs only the information for the VPNs it supports
    - Directly connected VPNs
- Easy to add new sites
    - Configure the site on the PE connected to it, the network automatically does the rest ☺

Demo



- In this demo, OSPF is used in provider network and customer networks
- In general, different protocols can be used

- In the provider and customer networks
- In the networks of different customers
- In different sites of the same customer

Routes on PE1

- Global routing table

  10.0.0.0/8 is variably subnetted, 7 subnets, 2 masks

  O 10.0.9.0/30 [110/2] via 10.0.9.5, 00:23:17, FastEthernet1/0

  C 10.0.9.4/30 is directly connected, FastEthernet1/0

  O 10.0.0.2/32 [110/3] via 10.0.9.5, 00:23:17, FastEthernet1/0

  C 10.0.0.3/32 is directly connected, Loopback0

  O 10.0.9.8/30 [110/3] via 10.0.9.5, 00:23:17, FastEthernet1/0

  O 10.0.0.1/32 [110/2] via 10.0.9.5, 00:23:17, FastEthernet1/0

  O 10.0.0.4/32 [110/4] via 10.0.9.5, 00:23:17, FastEthernet1/0

- Routing table for vrf Customer_A

  172.16.0.0/16 is variably subnetted, 4 subnets, 2 masks

  O 172.16.1.0/24 [110/2] via 10.0.1.2, 00:23:35, FastEthernet2/0

  O 172.16.0.1/32 [110/2] via 10.0.1.2, 00:23:35, FastEthernet2/0

  B 172.16.2.0/24 [200/2] via 10.0.0.4, 00:22:48

  B 172.16.0.2/32 [200/2] via 10.0.0.4, 00:22:48

  10.0.0.0/30 is subnetted, 2 subnets

  B 10.0.2.0 [200/0] via 10.0.0.4, 00:22:48

  C 10.0.1.0 is directly connected, FastEthernet2/0

- Global routing table

  172.16.0.0/16 is variably subnetted, 4 subnets, 2 masks

  C 172.16.1.0/24 is directly connected, FastEthernet1/1

  C 172.16.0.1/32 is directly connected, Loopback0

  O IA 172.16.2.0/24 [110/3] via 10.0.1.1, 00:23:39, FastEthernet1/0

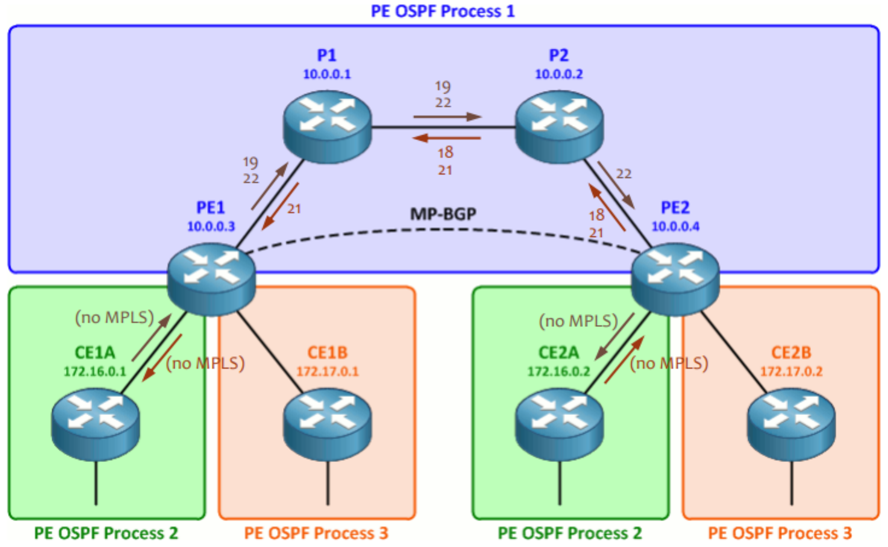
  O IA 172.16.0.2/32 [110/3] via 10.0.1.1, 00:23:39, FastEthernet1/0

  10.0.0.0/30 is subnetted, 2 subnets

  O IA 10.0.2.0 [110/2] via 10.0.1.1, 00:23:39, FastEthernet1/0

  C 10.0.1.0 is directly connected, FastEthernet1/0

MPLS Labels



NOTE: P1 and P2 assigned the same label to 10.0.0.3/32 (19) and to 10.0.0.4/32 (18) due to the symmetry of the topology / by coincidence; usually, they will all be different.

BGP VPNV4 Labels

PE1

```
  Network          Next Hop       In label/Out label
Route Distinguisher: 65000:1 (Customer_A)
   10.0.1.0/30      0.0.0.0            21/aggregate(Customer_A)
   10.0.2.0/30      10.0.0.4           nolabel/21
   172.16.0.1/32    10.0.1.2           22/nolabel
   172.16.0.2/32    10.0.0.4           nolabel/22
   172.16.1.0/24    10.0.1.2           23/nolabel
   172.16.2.0/24    10.0.0.4           nolabel/23
Route Distinguisher: 65000:2 (Customer_B)
   10.0.1.4/30      0.0.0.0            24/aggregate(Customer_B)
   10.0.2.4/30      10.0.0.4           nolabel/24
   172.17.0.1/32    10.0.1.6           25/nolabel
   172.17.0.2/32    10.0.0.4           nolabel/25
   172.17.1.0/24    10.0.1.6           26/nolabel
   172.17.2.0/24    10.0.0.4           nolabel/26
```

MPLS Fwd Table

PE1

```
Local  Outgoing    Prefix            Bytes tag  Outgoing   Next Hop
tag    tag or VC   or Tunnel Id      switched   interface
16     Pop tag     10.0.9.0/30       0          Fa1/0      10.0.9.5
17     16          10.0.9.8/30       0          Fa1/0      10.0.9.5
18     Pop tag     10.0.0.1/32       0          Fa1/0      10.0.9.5
19     17          10.0.0.2/32       0          Fa1/0      10.0.9.5
20     19          10.0.0.4/32       0          Fa1/0      10.0.9.5
21     Aggregate   10.0.1.0/30[V]    0
22     Untagged    172.16.0.1/32[V]  1140       Fa2/0      10.0.1.2
23     Untagged    172.16.1.0/24[V]  0          Fa2/0      10.0.1.2
24     Aggregate   10.0.1.4/30[V]    0
25     Untagged    172.17.0.1/32[V]  570        Fa2/1      10.0.1.6
26     Untagged    172.17.1.0/24[V]  0          Fa2/1      10.0.1.6
```

P1

```
Local  Outgoing    Prefix            Bytes tag  Outgoing   Next Hop
tag    tag or VC   or Tunnel Id      switched   interface
16     Pop tag     10.0.9.8/30       0          Fa2/1      10.0.9.2
17     Pop tag     10.0.0.2/32       0          Fa2/1      10.0.9.2
18     Pop tag     10.0.0.3/32       29201      Fa1/0      10.0.9.6
19     19          10.0.0.4/32       20005      Fa2/1      10.0.9.2
```

Other uses of MPLS

- Traffic engineering, IPv6 over MPLS, QoS, Pseudowire - IETF WG, Virtual Private LAN Service (VPLS) IETF WG