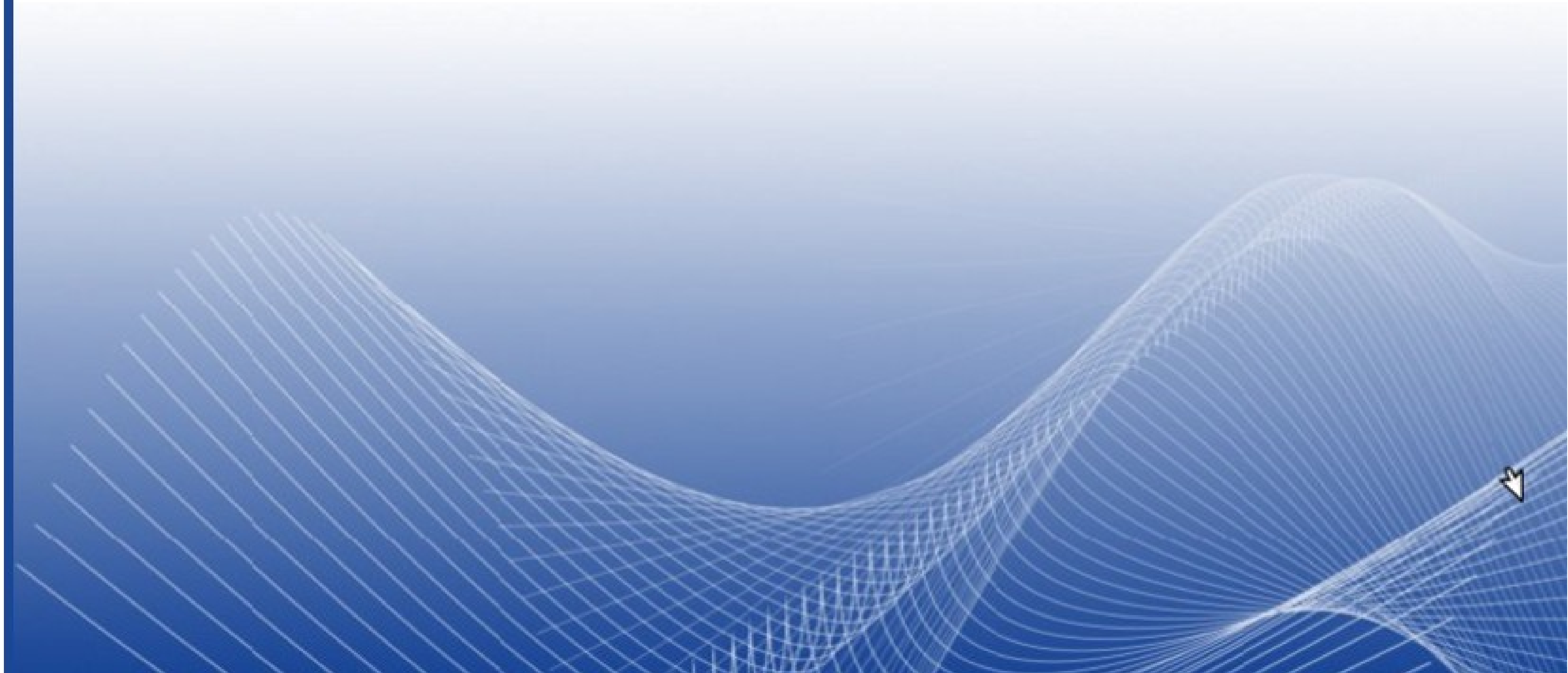
Multiprotocol Label Switching (MPLS)

Arquitetura e Gestão de Redes



universidade de aveiro

deti.ua.pt

Introduction

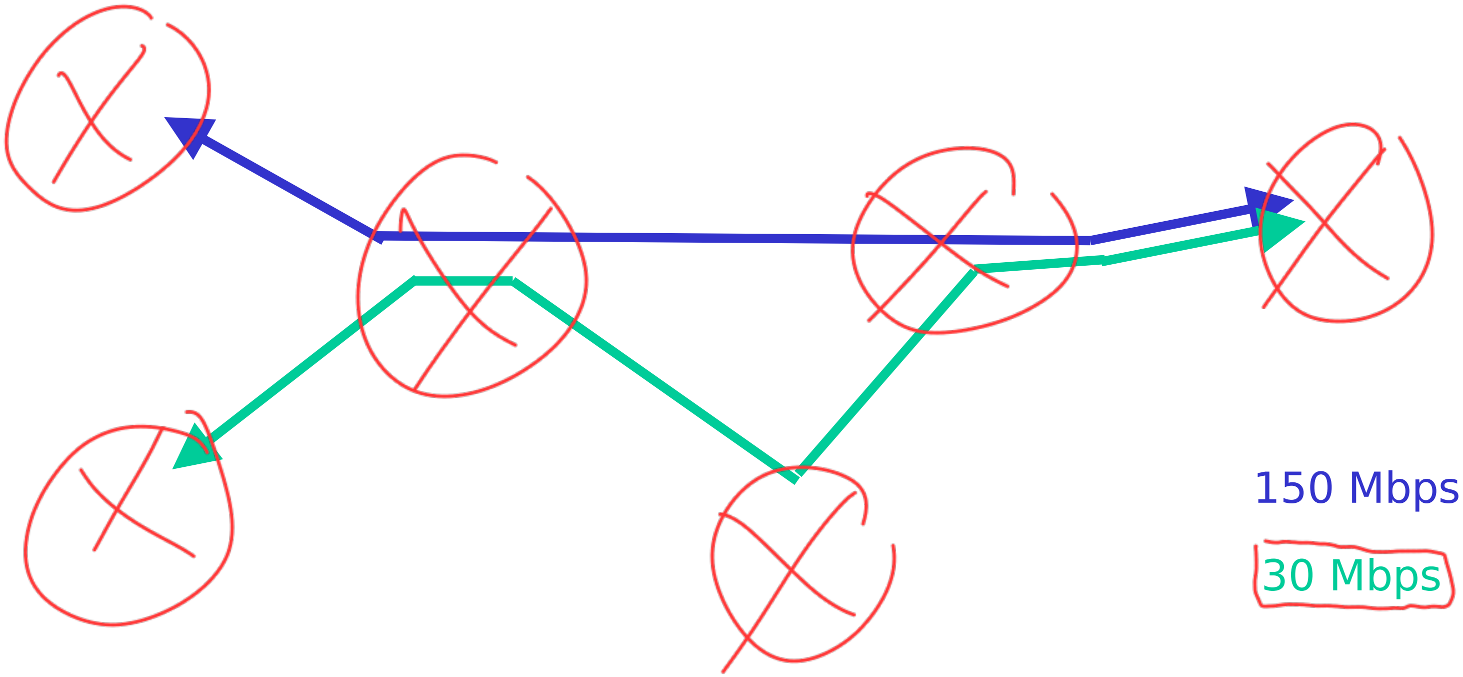
- On IP networks, *IntServ* and *DiffServ* are routing independent architectures
- IP network routing is based on the destination and does not allow to take the maximum possible advantage of the network resources

Example: flows between R1 and R3 (30 Mbps) and between R2 and R3 (150 Mbps)

- With RIP or OSPF or ANY OTHER IGP it is not possible to condition both flows.

Source-based routing

- Packets transport, from their source, a list of routers' addresses that define their path to the destination (*Options* field of the IP datagram header)



IP networks over ATM

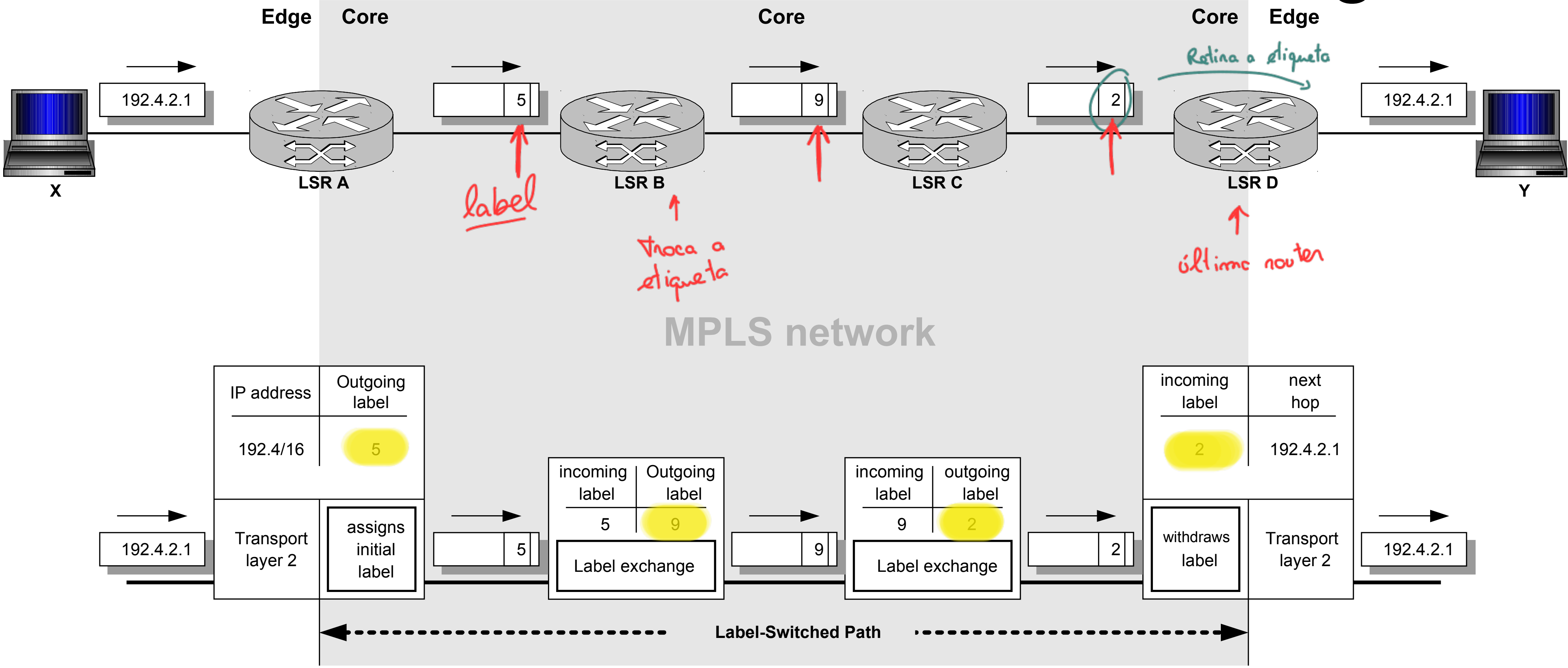
- IP routers are interconnected by an ATM network
- Connections between IP routers are implemented through virtual circuits (VCCs) or virtual paths (VPCs) on the ATM network
- It is necessary to manage two protocol layers

MPLS networks

↳ está por cima do RIP e OSPF,

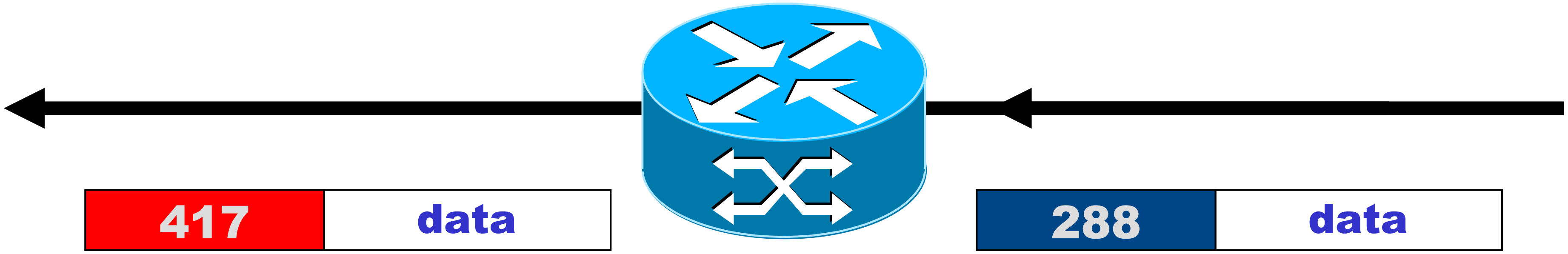
- Packets are labeled at the source with the label of the first hop
- Routers route packets based on their labels, just like ATM does with the VCI and VPI fields
- Advantages
  - Simplification of the packet routing process on routers
  - Traffic engineering capability equivalent to ATM
  - Simplification of the network management (a single protocol layer)

Label Switched Path: virtual circuit with label switching



- Networks are organized in domains
- Border routers insert/withdraw labels
- Labels have local meaning (can be reused on other links)
- Label distribution is made by an appropriate protocol

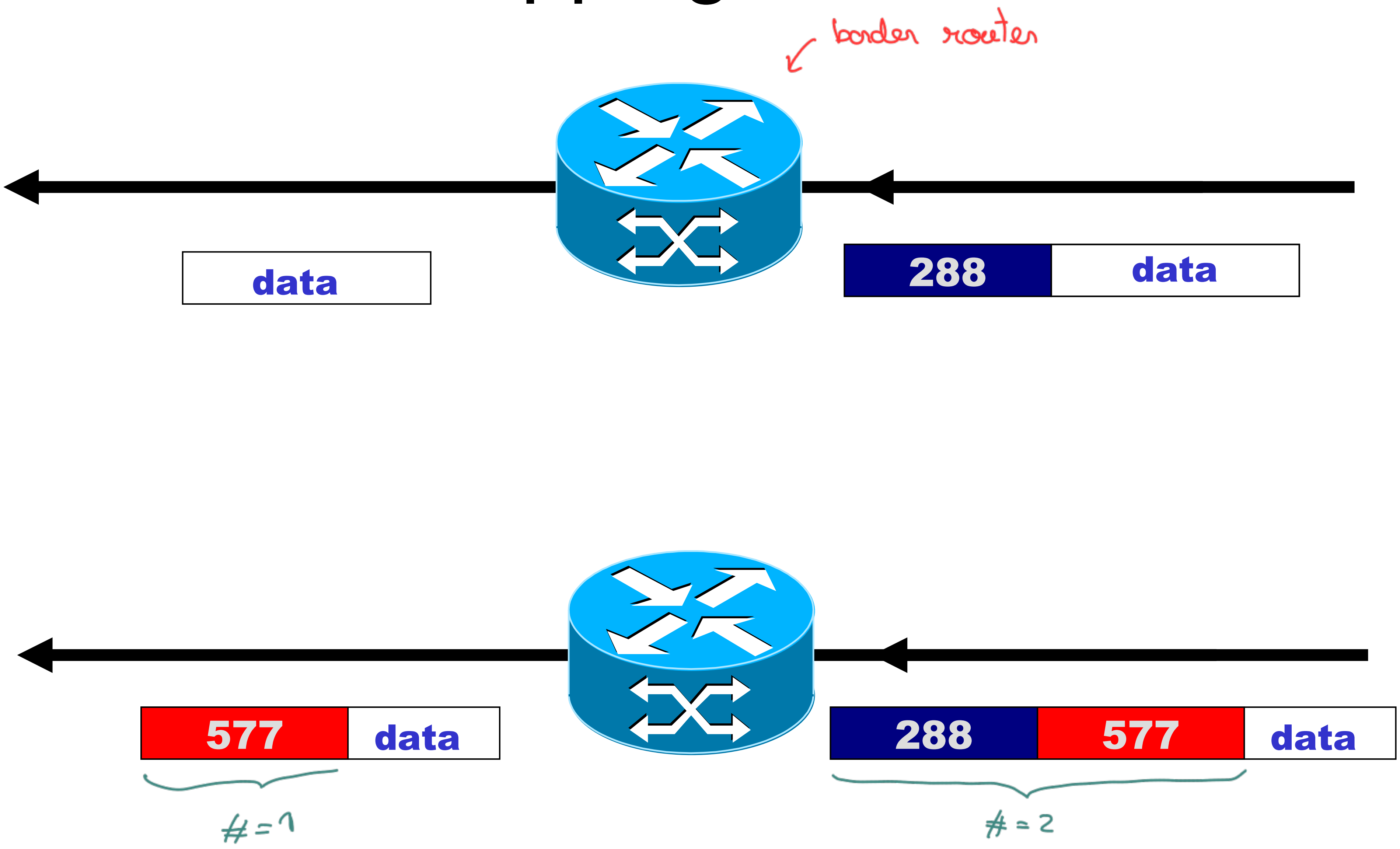
Forwarding via Label Swapping



Labels are short, fixed-length values.

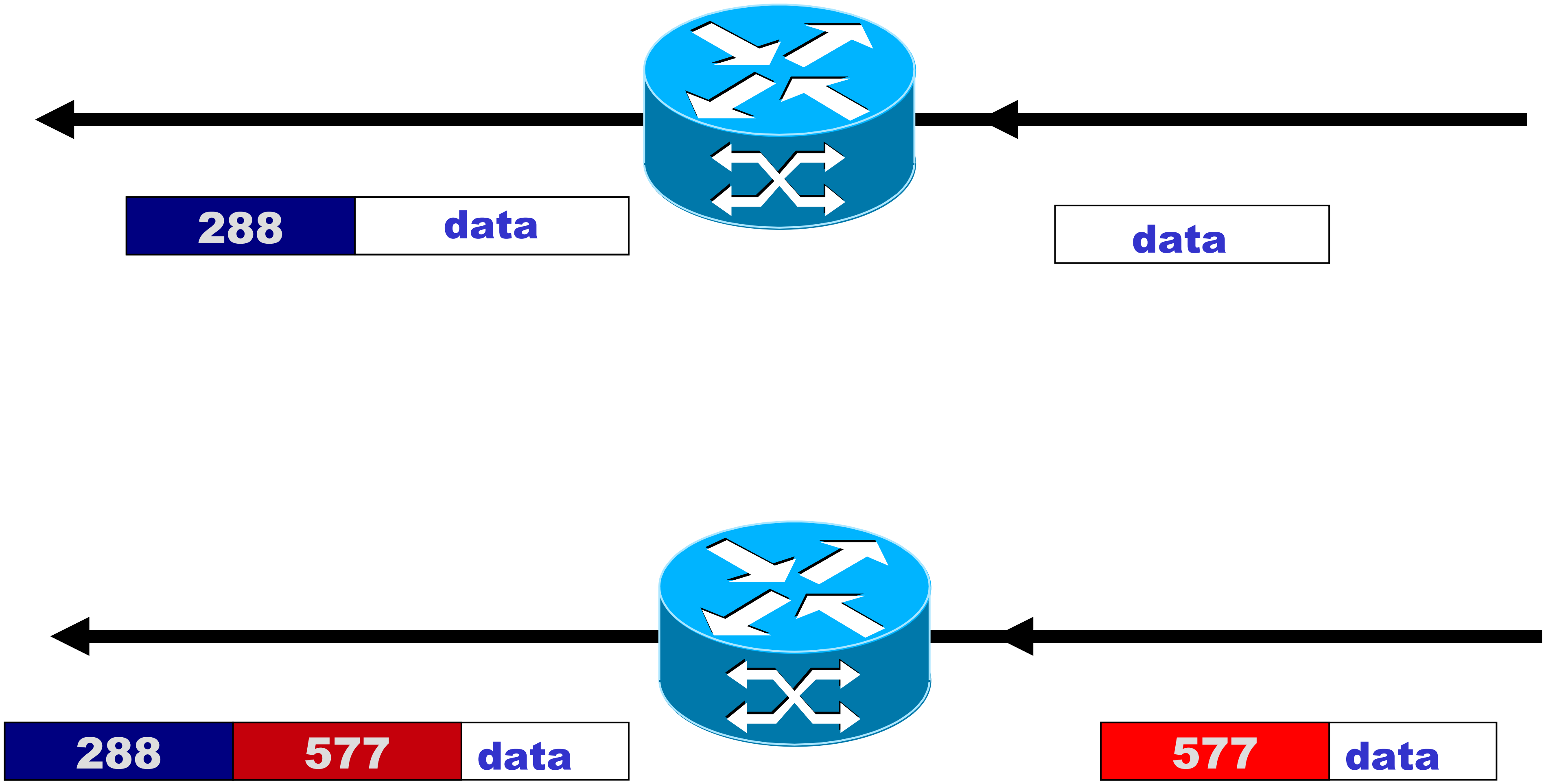
Popping Labels

Para sair

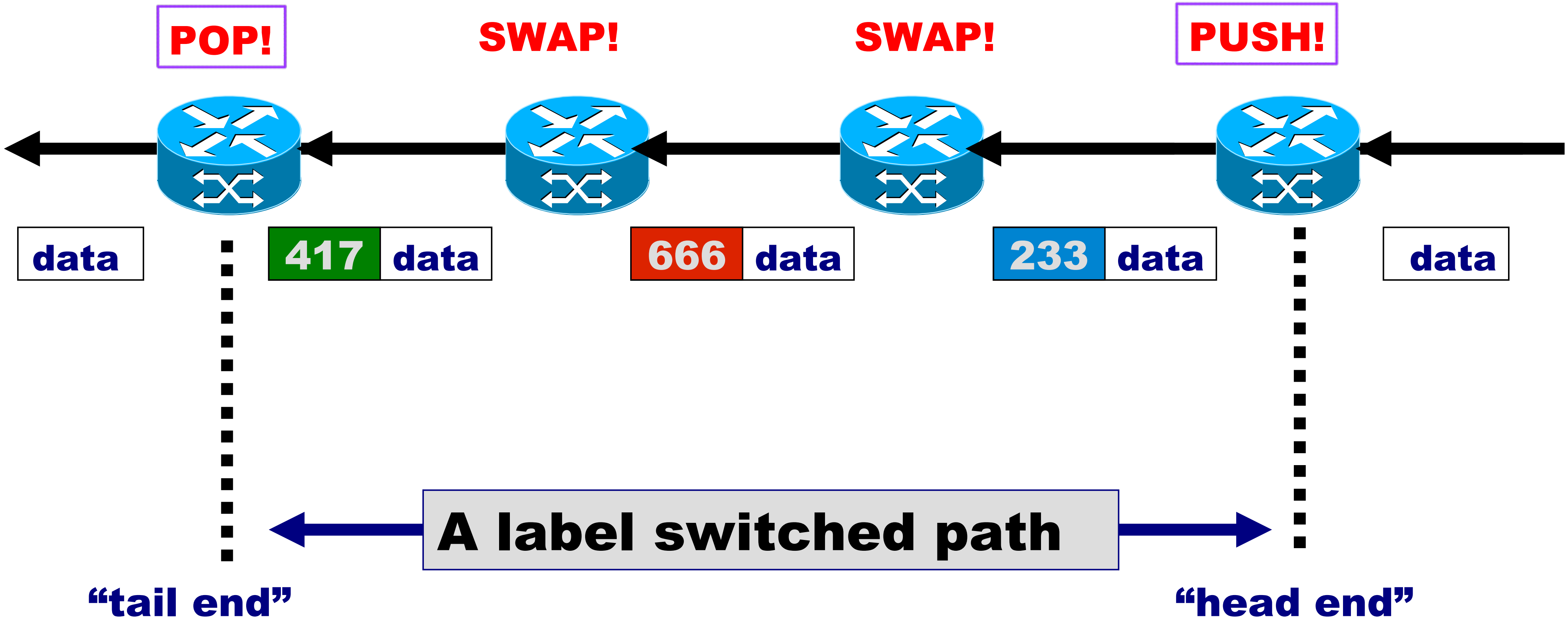


Pushing Labels

Por entrar

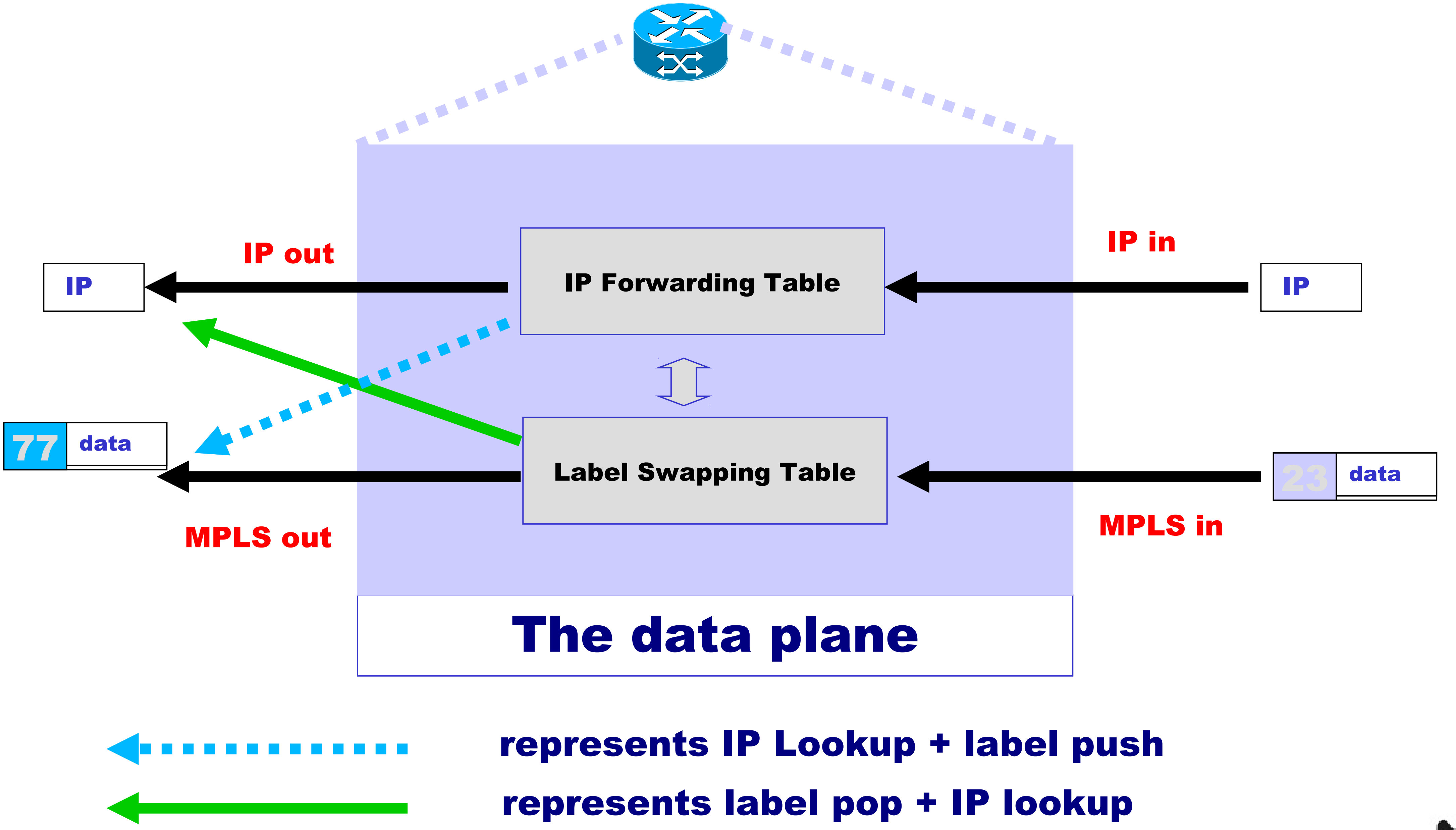


A Label Switched Path (LSP)



Often called an MPLS tunnel: payload headers are not Inspected inside of an LSP. Payload could be MPLS …

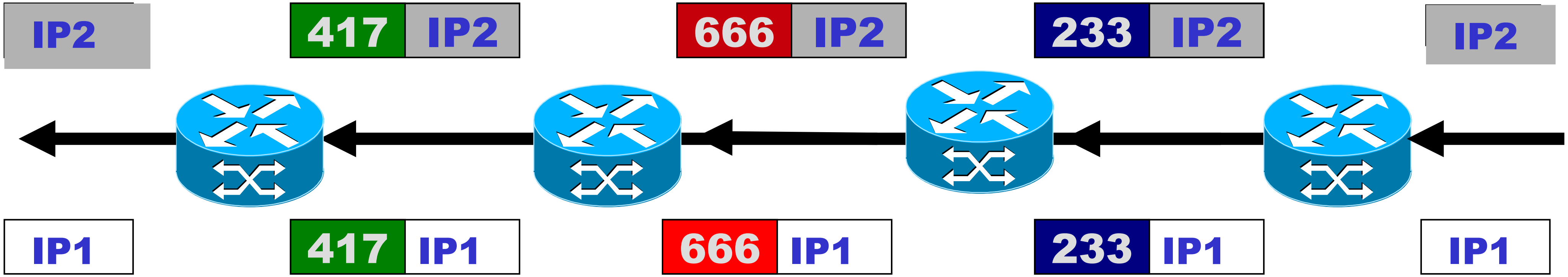
Label Switched Routers



represents IP Lookup + label push

represents label pop + IP lookup

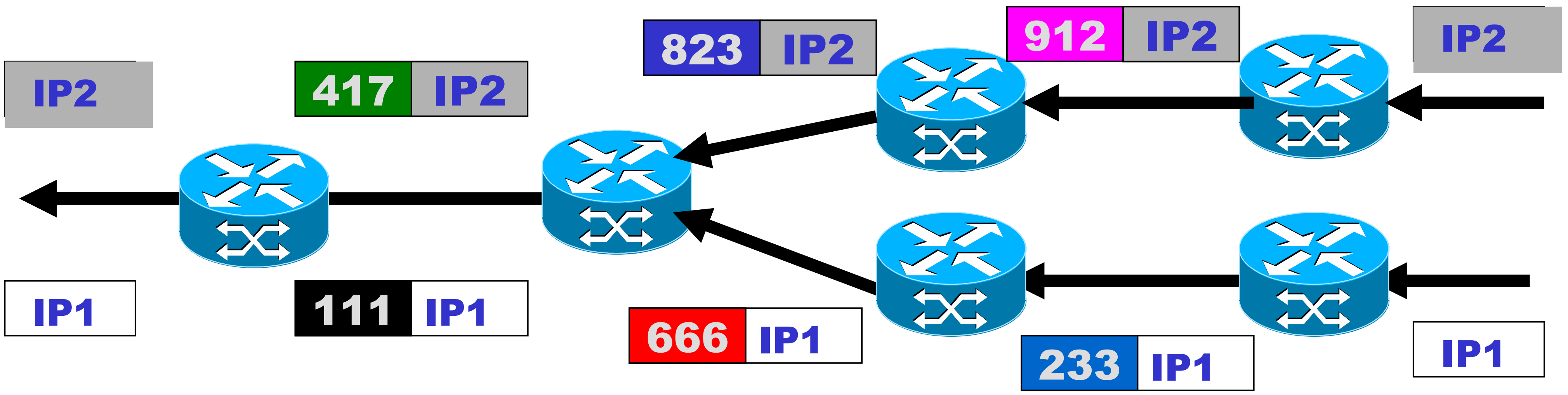
Forwarding Equivalence Class (FEC)



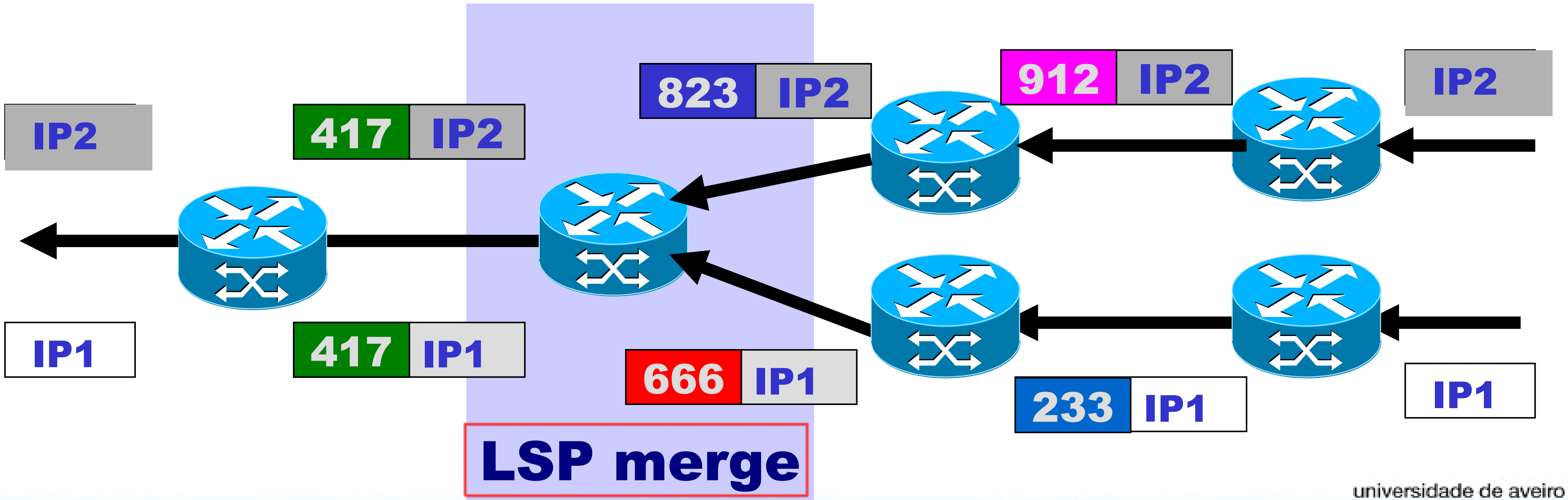
Packets IP1 and IP2 are forwarded in the same way --- they are in the same FEC.

Network layer headers are not inspected inside an MPLS LSP. This means that inside of the tunnel the LSRs do not need full IP forwarding table.

LSP Merge

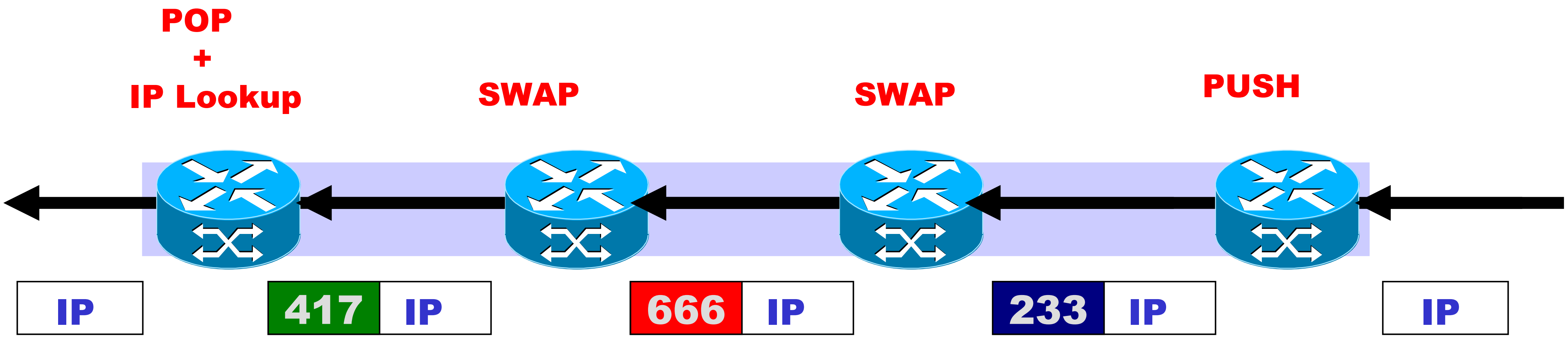


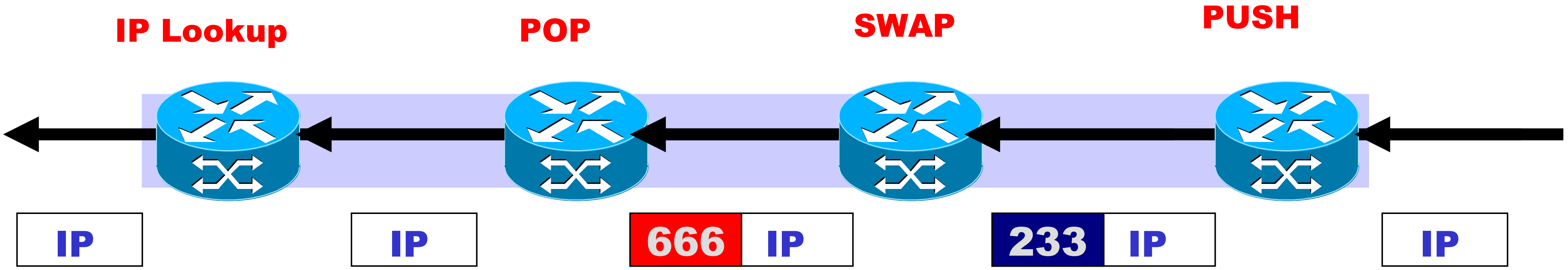
IP2
417 IP2
823 IP2
912 IP2
IP2
IP1
111 IP1
666 IP1
233 IP1
IP1



IP2
417 IP2
823 IP2
912 IP2
IP2
IP1
417 IP1
666 IP1
233 IP1
IP1
LSP merge

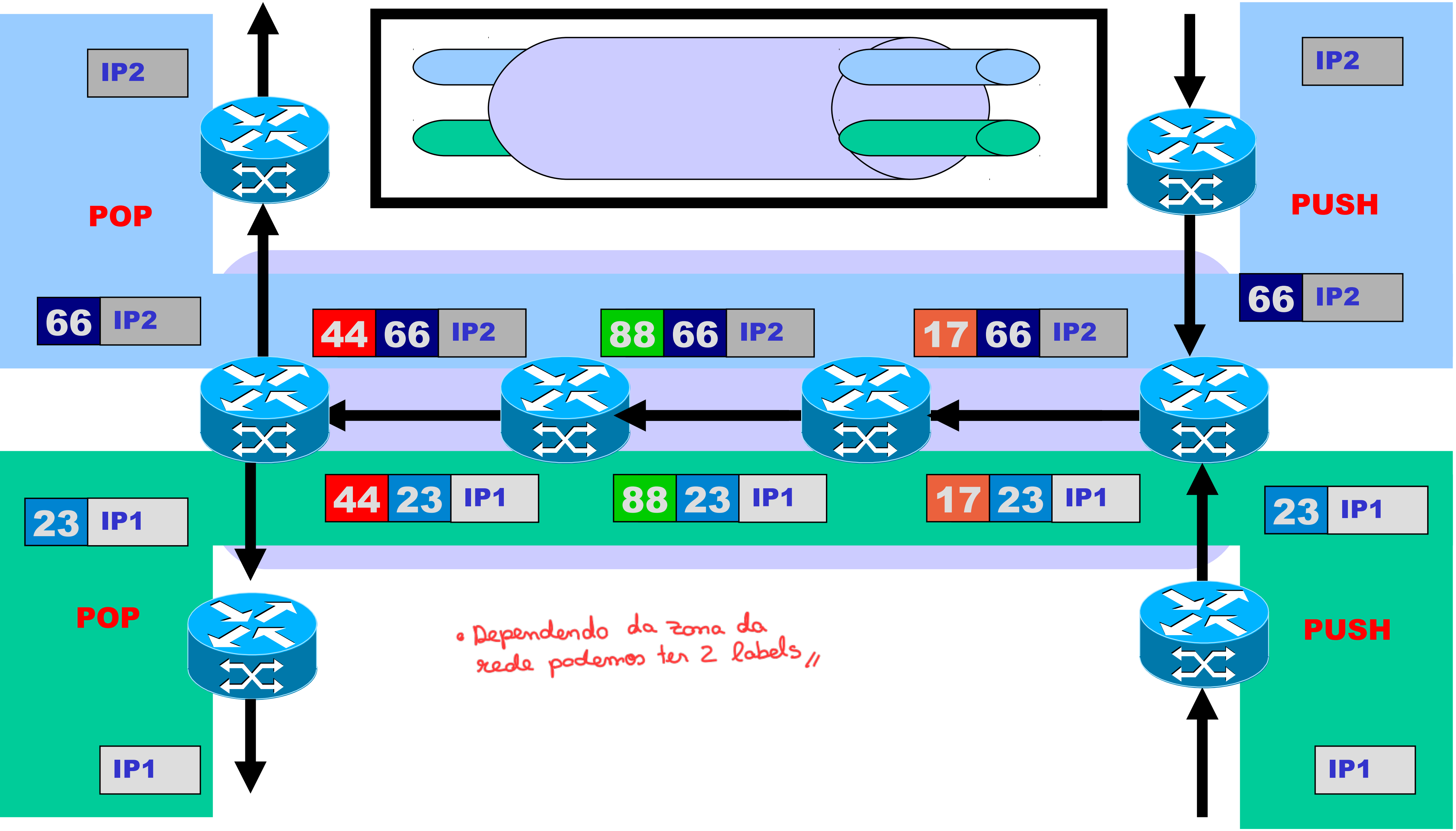
Penultimate Hop Popping





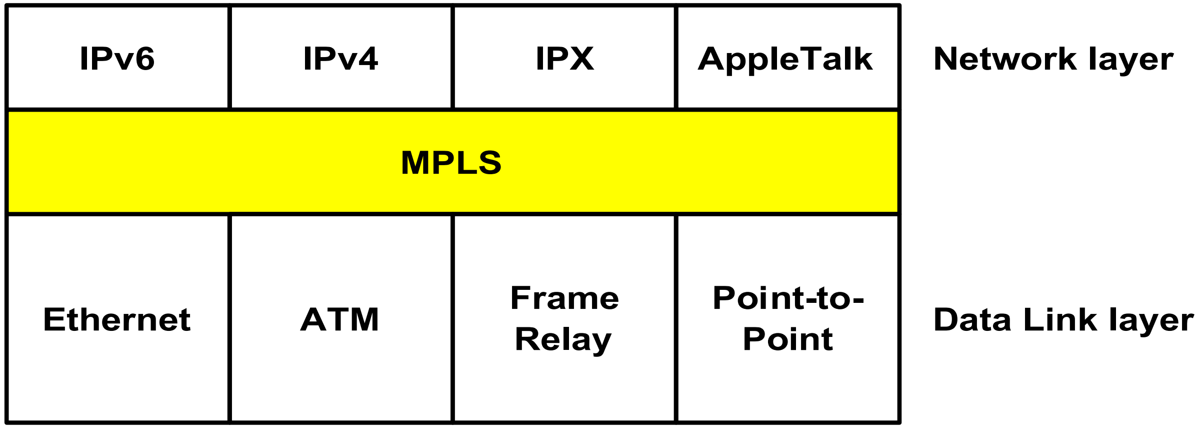
To reduce Label Edge Router overload

LSP Hierarchy via Label Stacking



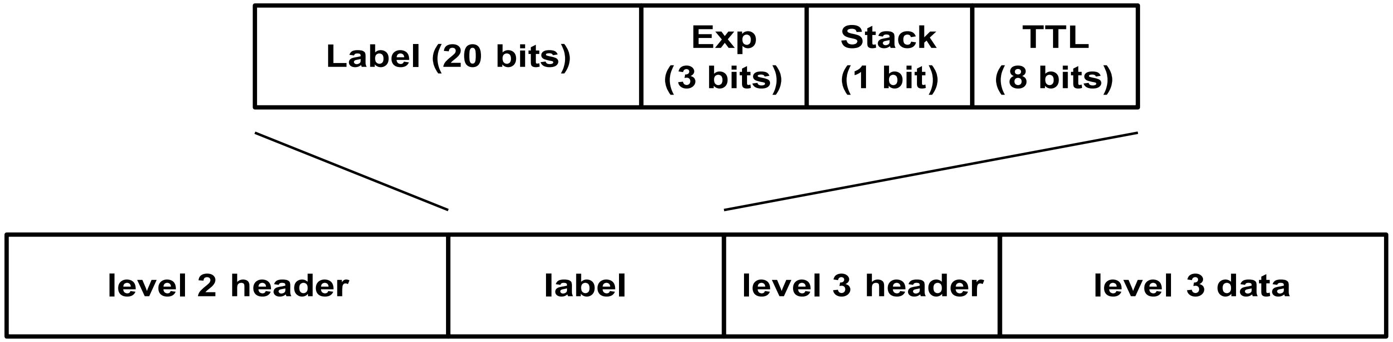
• Dependendo da zona da rede podemos ter 2 labels //

Labels



IPv6	IPv4	IPX	AppleTalk	Network layer
MPLS				
Ethernet	ATM	Frame Relay	Point-to-Point	Data Link layer

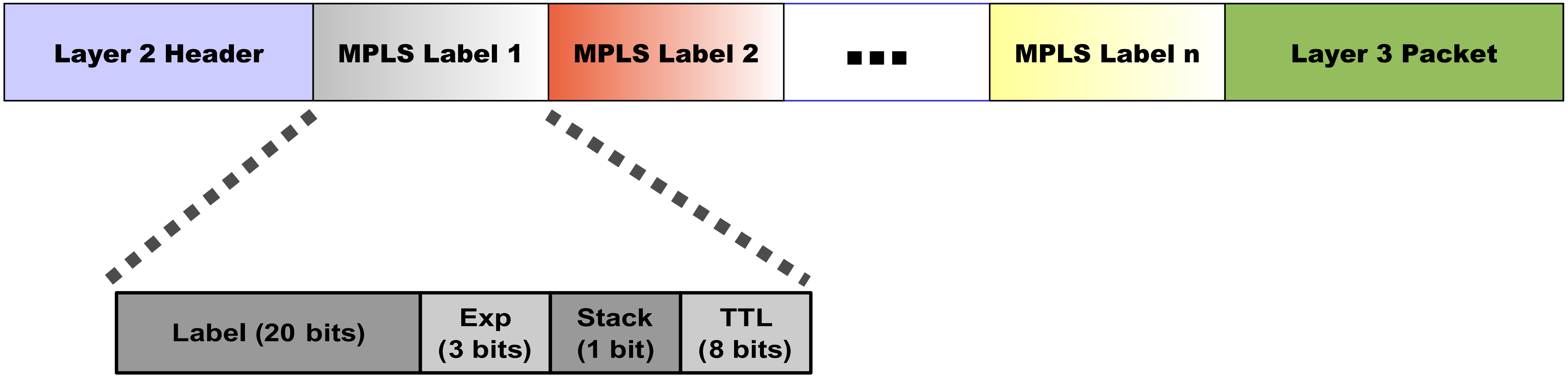
- On some Data Link (level 2) technologies, label is given by the appropriate fields of their header
  - ATM technology : VPI (Virtual Path ID) and VCI (Virtual Channel ID) fields
  - Frame Relay technology: DLCI (Data Link Connection Identifier) field
- On other Data Link technologies (Point-to-Point, Ethernet), the label is inserted between layer 2 and layer 3 headers



Label (20 bits)	Exp (3 bits)	Stack (1 bit)	TTL (8 bits)

level 2 header	label	level 3 header	level 3 data

Generic MPLS Encapsulation



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**

Constrained based Routing

- A cost is associated to each link
- Each link has a set of attributes that represent performance metrics
- The routing objective is to determine the lowest cost path that does not violate the restrictions that were assigned
- Restrictions can be associated to a set of performance characteristics, like for example, bandwidth, delay, priority, etc.
  - For the bandwidth case, the restriction that is imposed to the routing algorithm is that the path must have, on each connection it traverses, a bandwidth higher than a certain threshold.
  - In this case, the connection attribute used is the available bandwidth.

Constraint Based Routing

Problem here: OSPF areas hide information for scalability. So these extensions work best only within an area...

Basic components

1. Specify path constraints
2. Extend topology database to include resource and constraint information
3. Find paths that do not violate constraints and optimize some metric
4. Signal to reserve resources along path
5. Set up LSP along path (with explicit route)
6. Map ingress traffic to the appropriate LSPs

Extend Link State Protocols (IS-IS, OSPF)

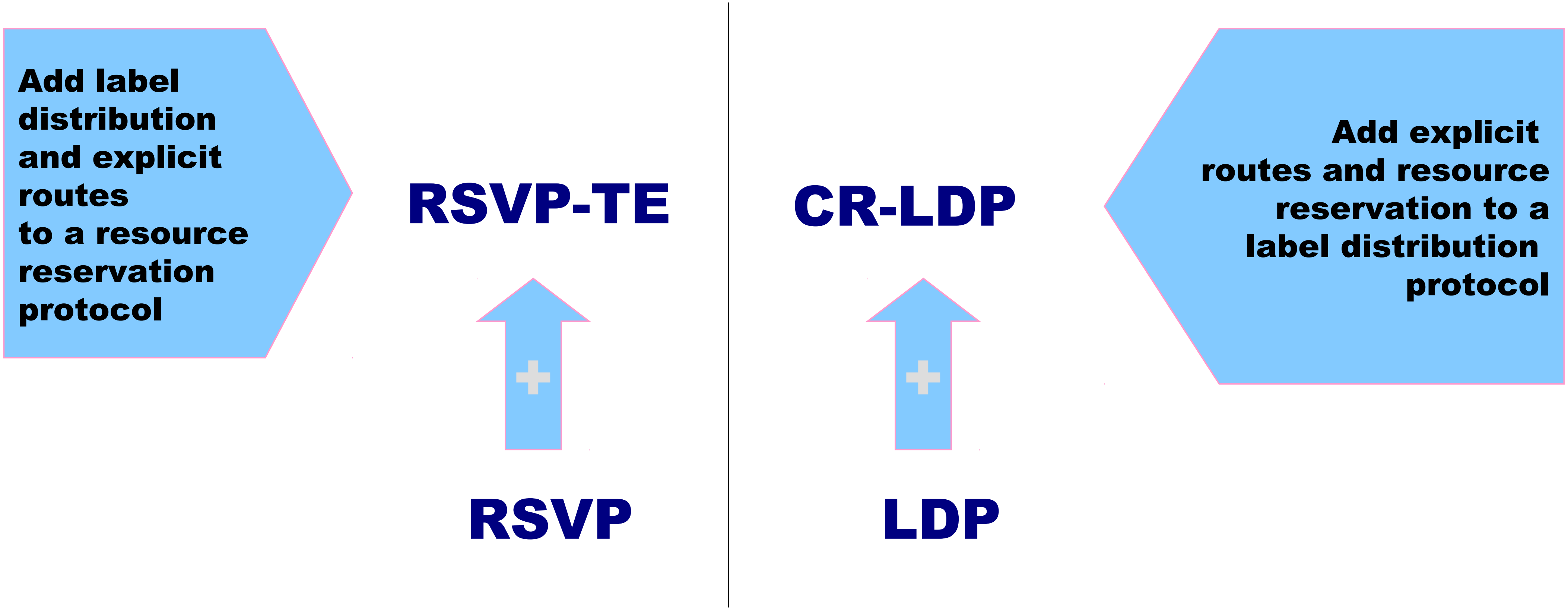
Extend RSVP or LDP or both!

Problem here: what is the “correct” resource model for IP services?

Note: (3) could be offline,
or online (perhaps an extension to OSPF)

Resource Reservation + Label Distribution

Two emerging/competing/dueling approaches:



Add label distribution and explicit routes to a resource reservation protocol

RSVP-TE

RSVP

CR-LDP

LDP

Add explicit routes and resource reservation to a label distribution protocol

RSVP-TE: Extensions to RSVP for LSP Tunnels
draft-ietf-mpls-rsvp-lsp-tunnel-08.txt

Constraint-Based LSP Setup using LDP
draft-ietf-mpls-cr-lpd-05.txt

RSVP-TE vs. CR-LPD

RSVP-TE

- Soft state periodically refreshed
- IntServe QoS model

CR-LDP

- State maintained incrementally
- New QoS model derived from ATM and Frame Relay

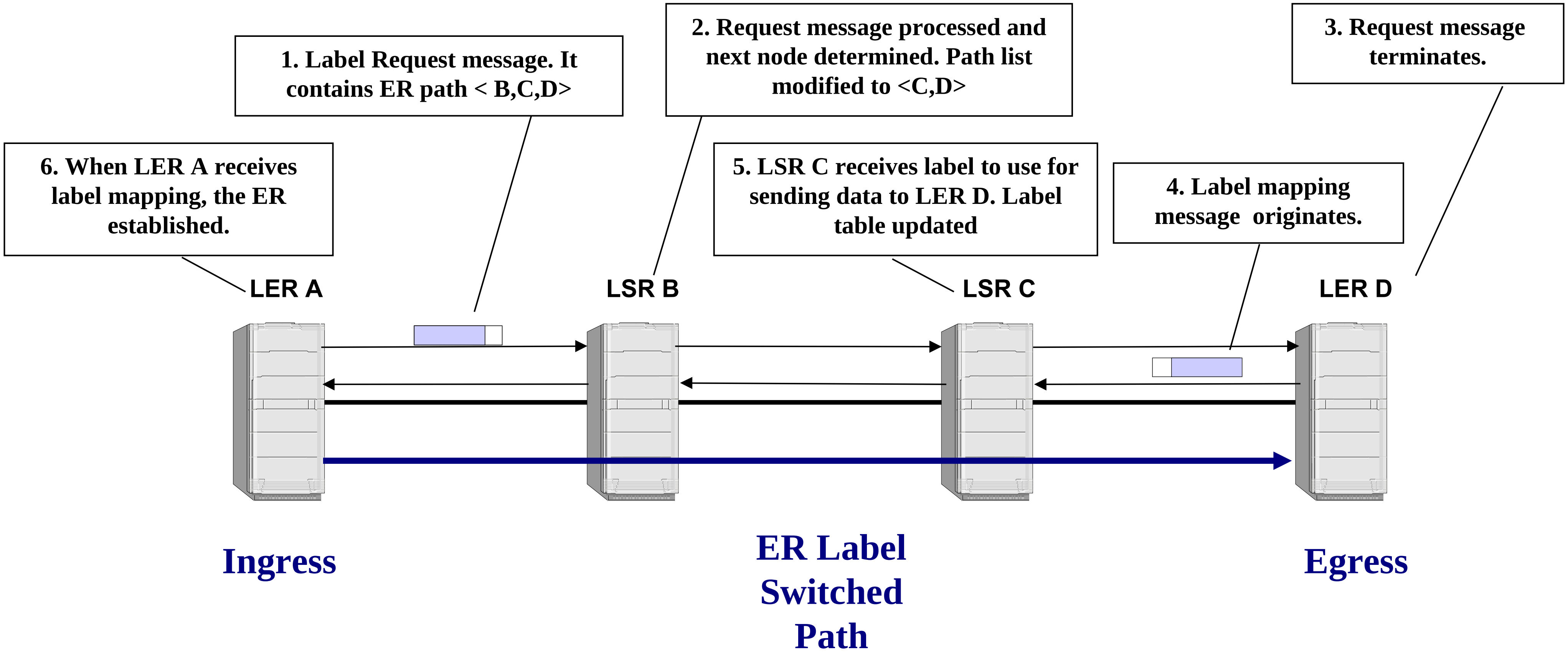
And the QoS model determines the additional information attached to links and nodes and distributed with extended link state protocols…

And what about that other Internet QoS model, diffserve?

LSPs establishing protocols

- RSVP-TE (*Resource Reservation Protocol – Traffic Engineering*)
  - Extension of the RSVP protocol
- CR-LDP (*Constrained based Routing – Label Distribution Protocol*)
  - Extension of the LDP protocol
- Both protocols enable:
  - The specification of a route to a LSP
  - To chose the labels on each link of the route
  - To make resources reservation for the LSP
- Routes are previously determined:
  - By management (Traffic engineering)
  - By a *Constrained based Routing* type protocol

Establishment of a LSP with the CR-LDP protocol



1. Label Request message. It contains ER path < B,C,D>
2. Request message processed and next node determined. Path list modified to <C,D>
3. Request message terminates.
6. When LER A receives label mapping, the ER established.
5. LSR C receives label to use for sending data to LER D. Label table updated
4. Label mapping message originates.
LER A
LSR B
LSR C
LER D
Ingress
ER Label Switched Path
Egress

Label Distribution Protocol (LDP)

RFC 3036. LDP Specification. (1/2001)

- Dynamic distribution of label binding information
- LSR discovery
- Reliable transport with TCP
- Incremental maintenance of label swapping tables (only deltas are exchanged)
- Designed to be extensible with Type-Length-Value (TLV) coding of messages
- Modes of behavior that are negotiated during session initialization
  - Label retention (liberal or conservative)
  - LSP control (ordered or independent)
  - Label assignment (unsolicited or on-demand)

LDP Message Categories

- Discovery messages: used to announce and maintain the presence of an LSR in a network.
- Session messages: used to establish, maintain, and terminate sessions between LDP peers.
- Advertisement messages: used to create, change, and delete label mappings for FECs.
- Notification messages: used to provide advisory information and to signal error information.

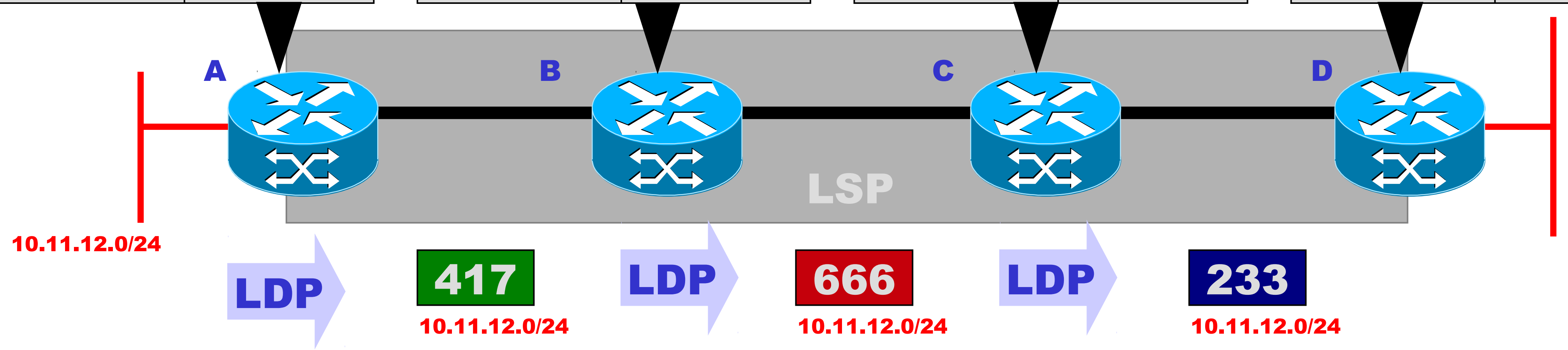
LDP and Hop-by-Hop routing

network	next-hop
10.11.12.0/24	direct

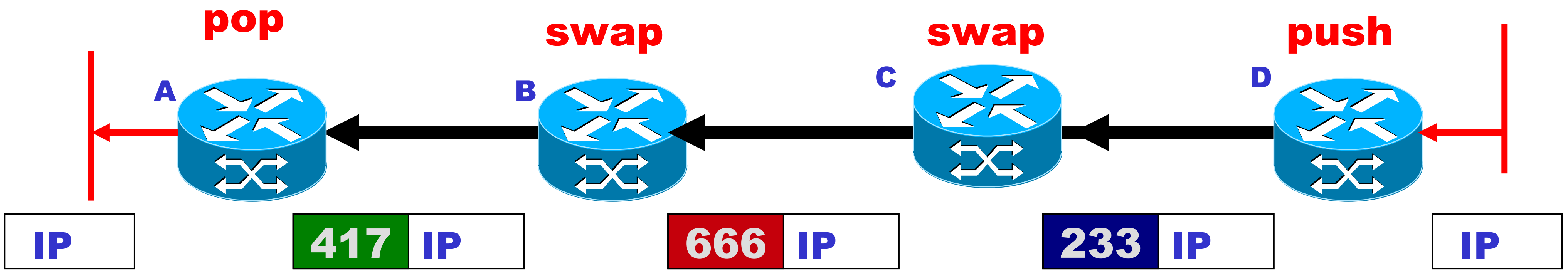
network	next-hop
10.11.12.0/24	A

network	next-hop
10.11.12.0/24	B

network	next-hop
10.11.12.0/24	C



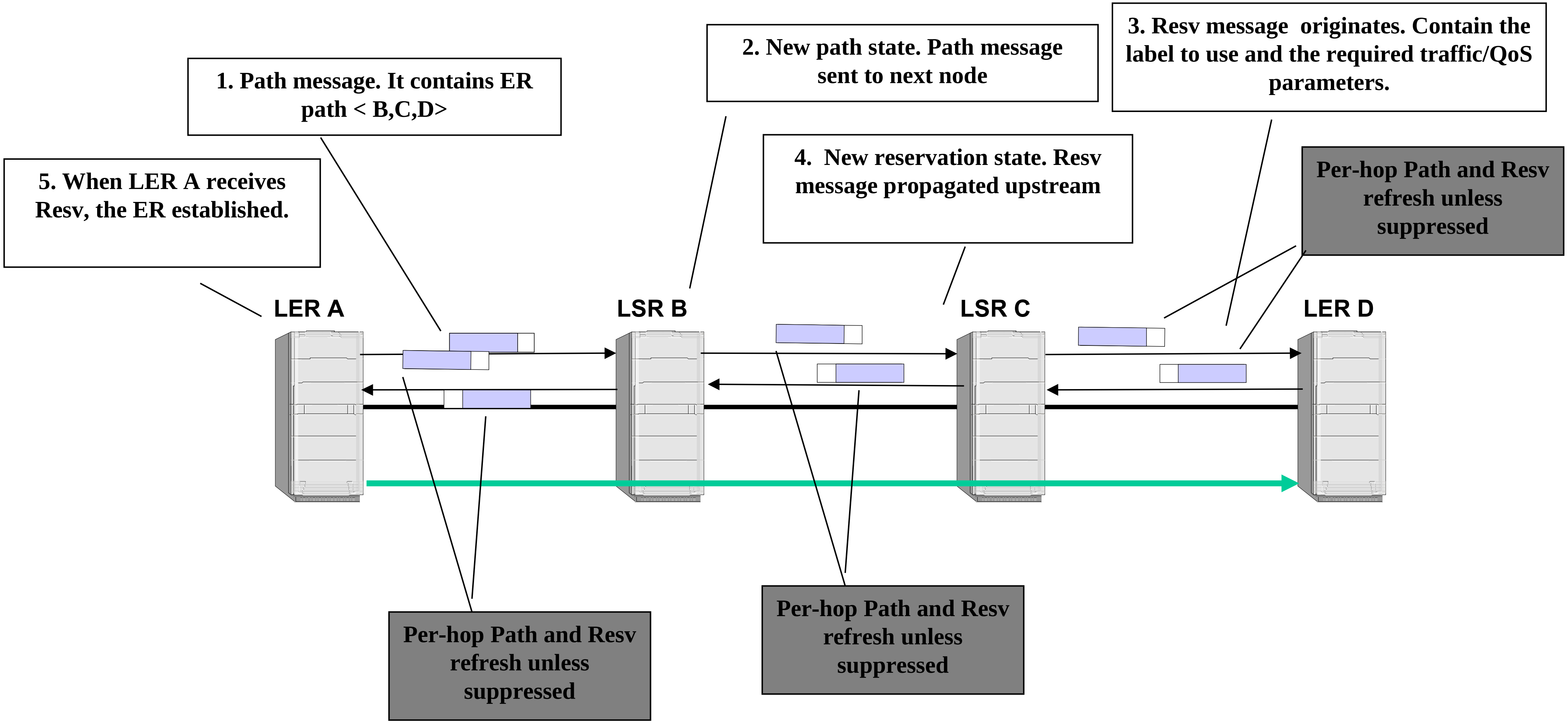
Generate new label
And bind to destination



A closer look at CR-LDP

- Defines new TLV encodings and procedures for
  - Explicit routing (strict and loose)
  - Route pinning (nail down some segments of a loosely routed path)
  - Traffic parameter specification
    - Peak rate
    - Committed rate
    - Weight
    - Resource class or color
  - LSP preemption (reroute existing paths to accommodate a new path)
  - LSP Identifiers (LSPIDs)

Establishment of a LSP with the RSVP-TE protocol



1. Path message. It contains ER path < B,C,D>

2. New path state. Path message sent to next node

3. Resv message originates. Contain the label to use and the required traffic/QoS parameters.

4. New reservation state. Resv message propagated upstream

5. When LER A receives Resv, the ER established.

Per-hop Path and Resv refresh unless suppressed

Per-hop Path and Resv refresh unless suppressed

Per-hop Path and Resv refresh unless suppressed

LSPs priorities

- When:
  - A new LSP requires resources that are not available on the network, or
  - On failure situations (on a link, for example)

- The operator can establish different priorities to avoid the “most important” traffic from becoming blocked by the “less important” traffic.

- Each LSP has two priorities assigned: “*setup Priority*” and “*holding Priority*”

- There are 8 different priority levels

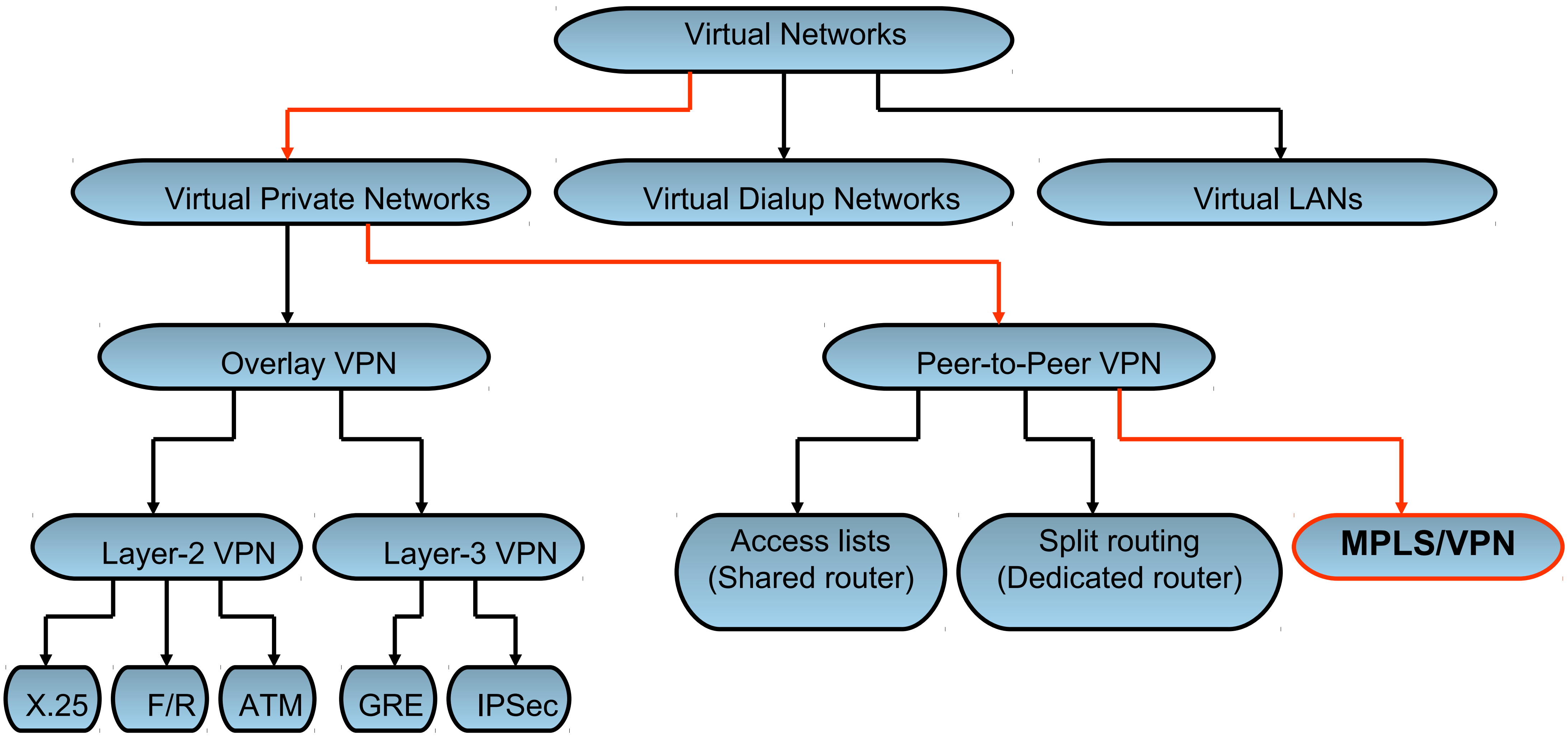
- A established LSP can “steal” network resources from the already established LSPs that have a lower “*holding Priority*” than its “*setup Priority*”

MPLS - Major Drivers

- Provide IP VPN Services
  - Scalable IP VPN service – Build once and sell many
  - Managed Central Services – Building value added services and offering them across VPNs
- Managing traffic on the network using MPLS Traffic Engineering
  - Providing tighter SLA/QoS (Guaranteed BW Services)
  - Protecting bandwidth - Bandwidth Protection Services
- Integrating Layer 2 & Layer 3 Infrastructure
  - Layer 2 services such as Frame Relay and ATM over MPLS
  - Mimic layer 2 services over a highly scalable layer 3 infrastructure

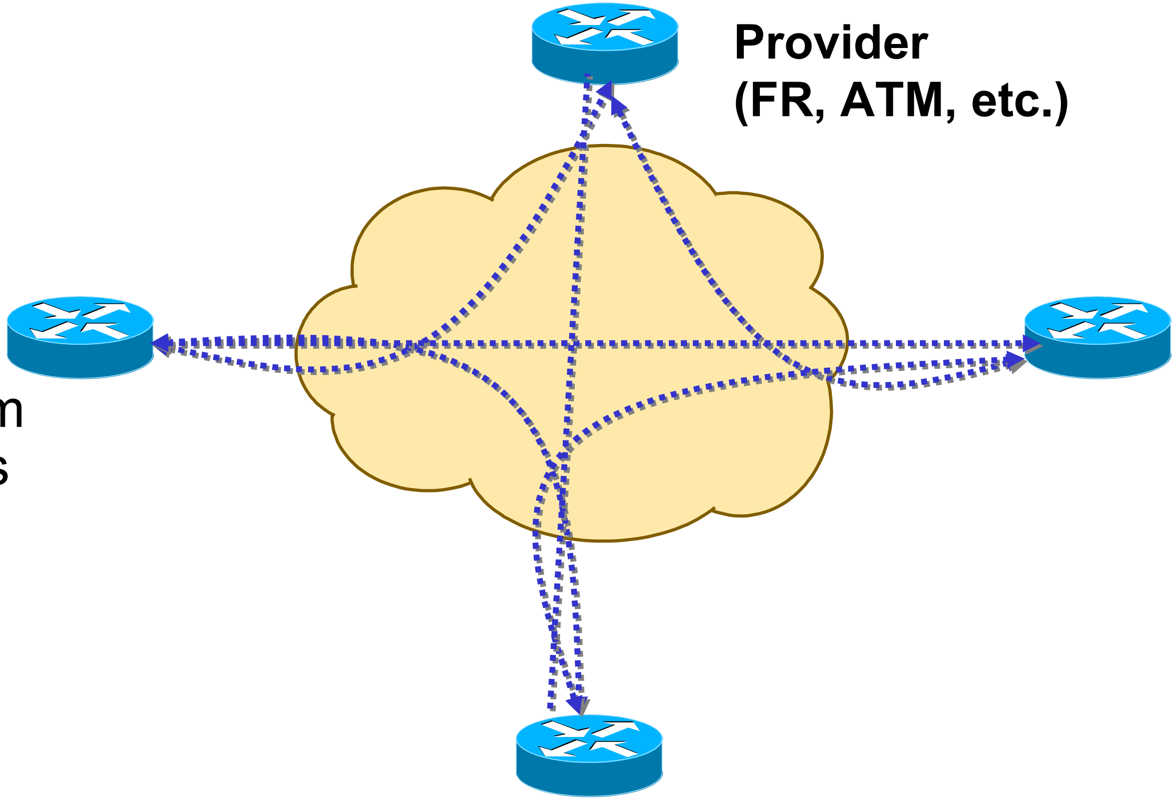
MPLS Layer 3 VPNs

Virtual Network Models



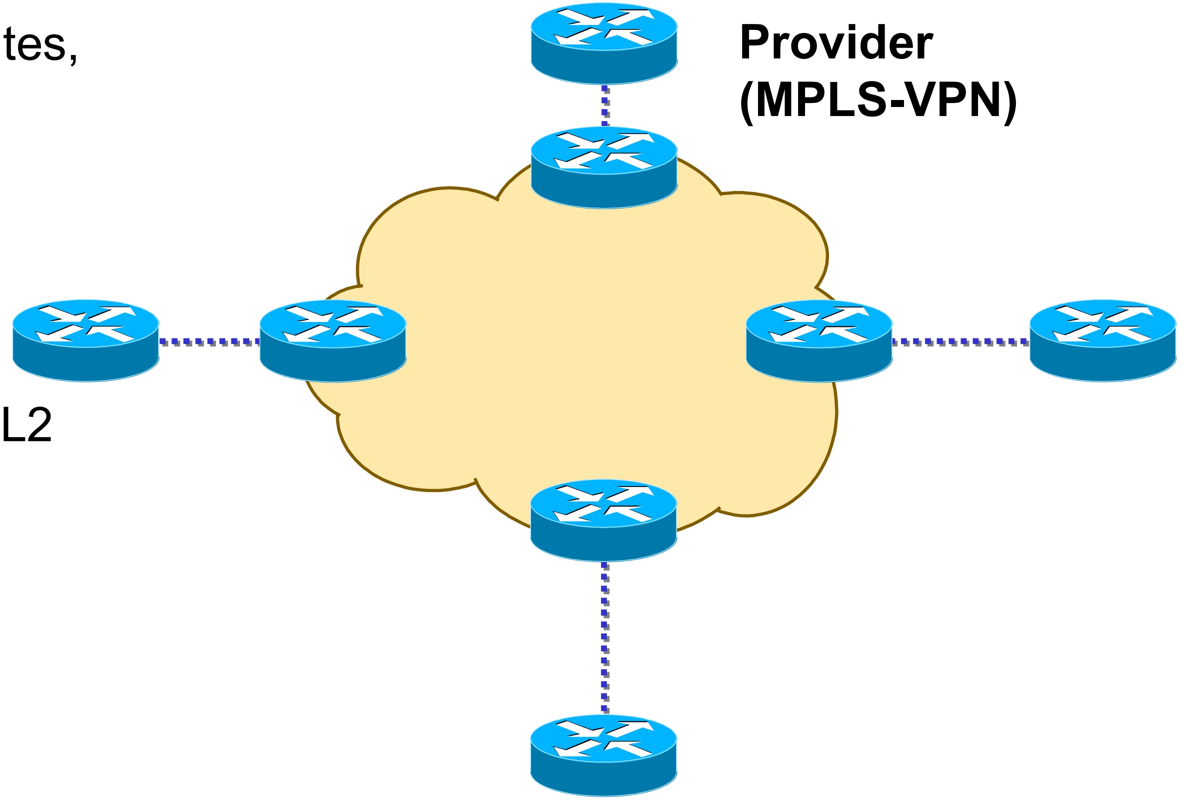
Overlay Network

- Provider sells a circuit service
- Customers purchases circuits to connect sites, runs IP
- N sites, (N*(N-1))/2 circuits for full mesh—expensive
- The big scalability issue here is routing peers— N sites, each site has N-1 peers
- Hub and spoke is popular, suffers from the same N-1 number of routing peers
- Hub and spoke with static routes is simpler, still buying N-1 circuits from hub to spokes
- Spokes distant from hubs could mean lots of long-haul circuits



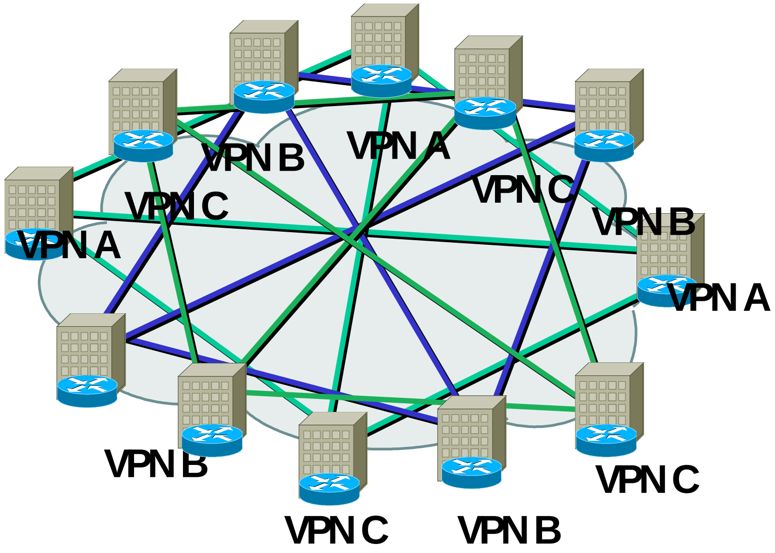
Peer Network

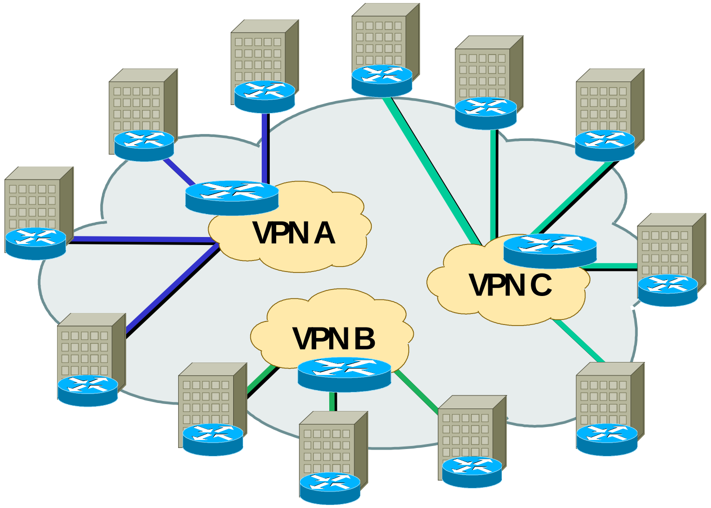
- Provider sells an MPLS-VPN service
- Customers purchases circuits to connect sites, runs IP
- N sites, N circuits into provider
- Access circuits can be any media at any point (FE, POS, ATM, T1, dial, etc.)
- Full mesh connectivity without full mesh of L2 circuits
- Hub and spoke is also easy to build
- Spokes distant from hubs connect to their local provider’s POP, lower access charge because of provider’s size
- The Internet is a large peer network



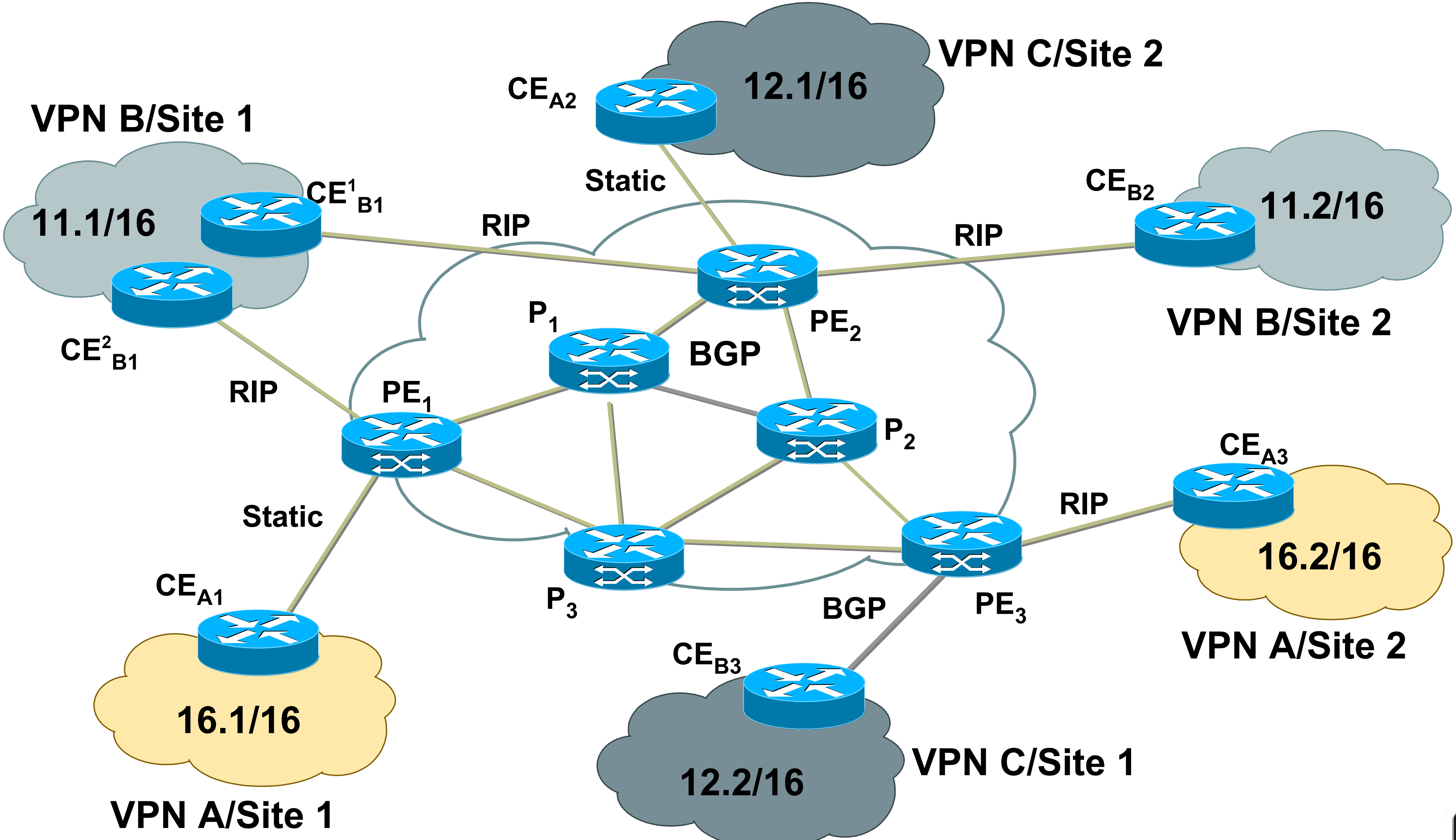
MPLS L3 VPNs using BGP (RFC2547)

- End user perspective
  - Virtual Private IP service
  - **Simple routing – just point default to provider**
  - Full site-site connectivity without the usual drawbacks (routing complexity, scaling, configuration, cost)
- Major benefit for provider – scalability





MPLS VPN Topology



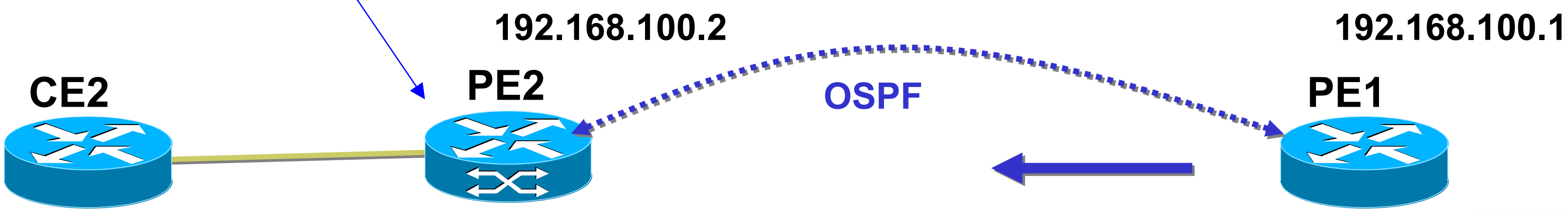
VPN Routing and Forwarding Instance (VRF)

- PE routers maintain separate routing tables
  - Global routing table
    - Contains all PE and P routes (perhaps BGP)
    - Populated by the VPN backbone IGP
  - VRF (VPN routing and forwarding)
    - Routing and forwarding table associated with one or more directly connected sites (CE routers)
    - VRF is associated with any type of interface, whether logical or physical (e.g. sub/virtual/tunnel)
    - Interfaces may share the same VRF if the connected sites share the same routing information

PE Router – Global Routing Table Output

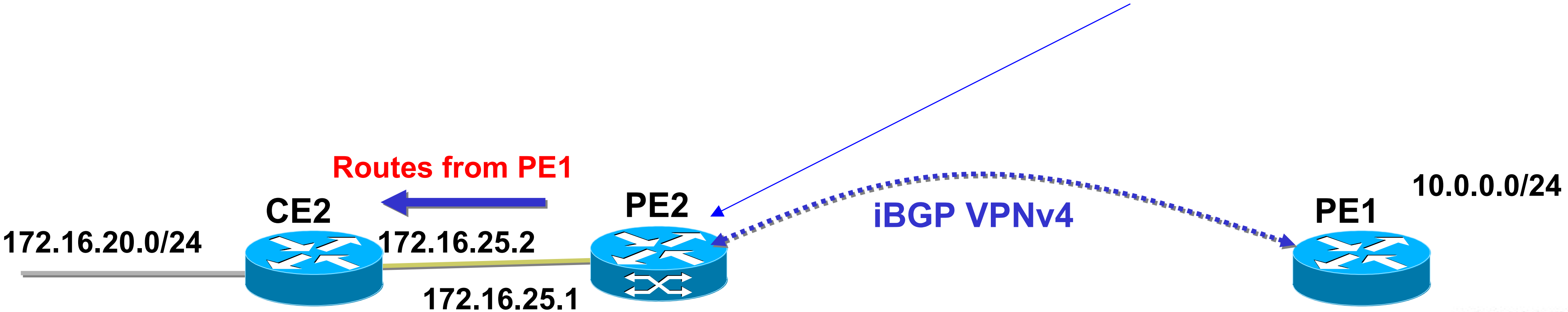
```
PE2#sh ip route

Gateway of last resort is not set

C    192.168.1.0/24 is directly connected, Ethernet0/0
     192.168.100.0/32 is subnetted, 3 subnets
O       192.168.100.1 [110/11] via 192.168.1.1, 00:04:27, Ethernet0/0
C       192.168.100.2 is directly connected, Loopback0
O       192.168.100.3 [110/11] via 192.168.1.3, 00:04:27, Ethernet0/0
```

Routes from PE1’s Global Routing Table



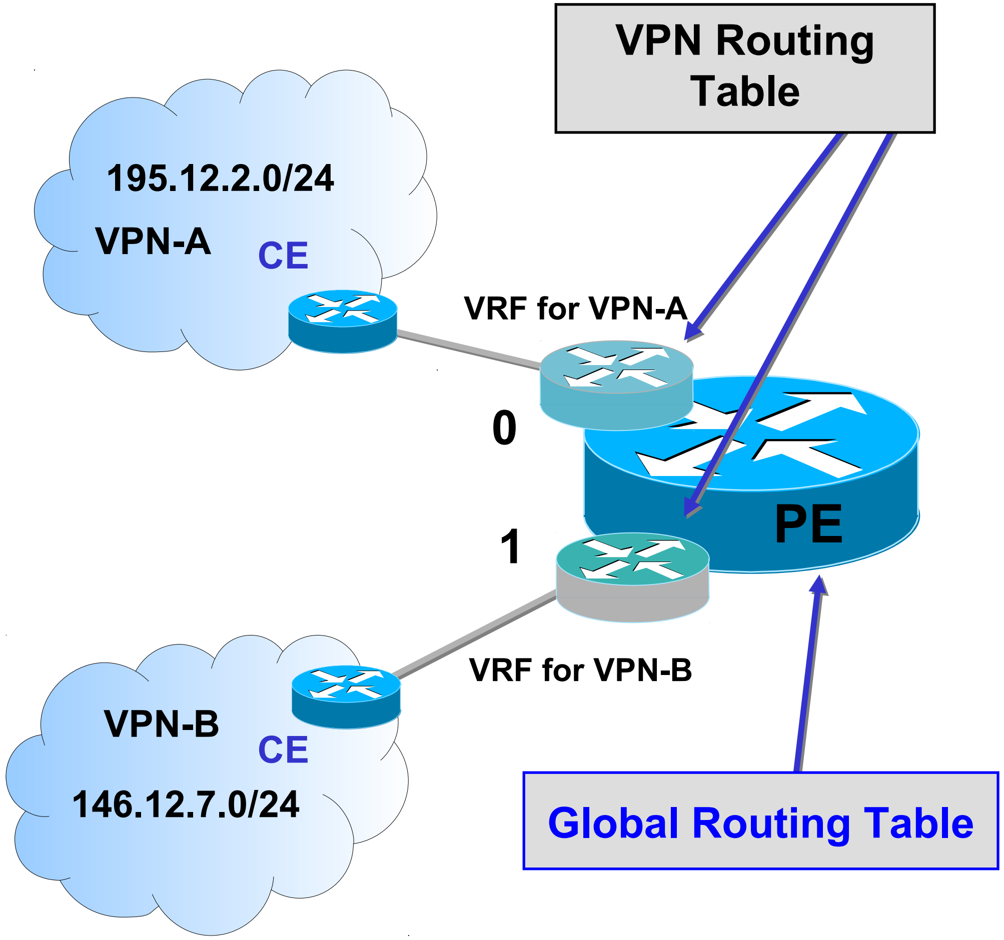
PE Router – VRF Routing Table Output

```
PE2#sh ip route vrf RED
Routing Table: RED

Gateway of last resort is 192.168.100.1 to network 0.0.0.0

     172.16.0.0/16 is variably subnetted, 8 subnets, 3 masks
C       172.16.25.0/30 is directly connected, Serial4/0
C       172.16.25.2/32 is directly connected, Serial4/0
B       172.16.20.0/24 [20/0] via 172.16.25.2, 00:07:04
     10.0.0.0/24 is subnetted, 1 subnets
B       10.0.0.0 [200/307200] via 192.168.100.1, 00:06:28
B*   0.0.0.0/0 [200/0] via 192.168.100.1, 00:07:03
```

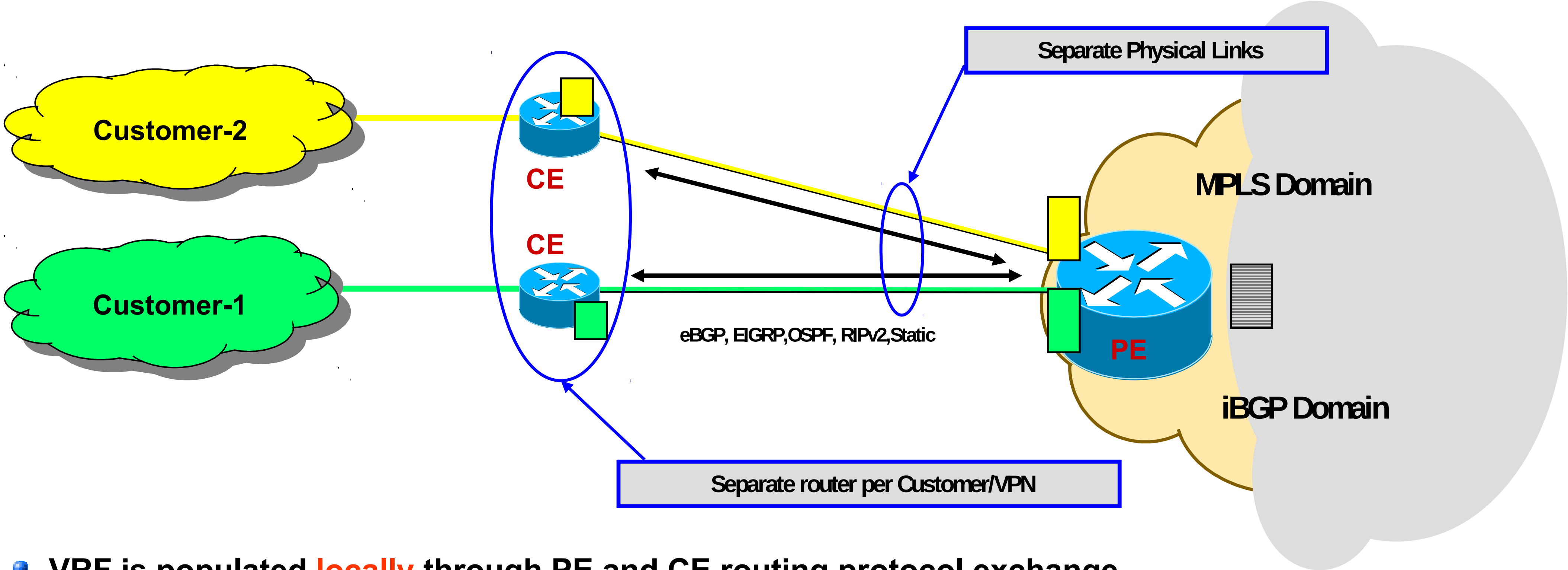


Virtual Routing and Forwarding Instances

- Define a unique VRF for interface 0
- Define a unique VRF for interface 1
  - Packets will never go between interfaces 0 and 1
- Uses VPNv4 to exchange VRF routing information between PE's



VRF Route Population



- **VRF is populated locally through PE and CE routing protocol exchange**
  - RIP Version 2, OSPF, BGP-4, EIGRP, & Static routing
  - “connected” is also supported (i.e. Default-gateway is PE)
- **Separate routing context for each VRF**
  - routing protocol context (BGP-4 & RIP V2)
  - separate process (OSPF)

Carrying VPN Routes in BGP

- VRFs by themselves aren’t all that useful
- Need some way to get the VRF routing information off the PE and to other PEs
- This is done with BGP

Additions to BGP to Carry MPLS-VPN Info

- RD: Route Distinguisher
- VPNv4 address family
- RT: Route Target
- Label

Route Distinguisher

```
!
ip vrf red
rd 1:1
route-target export 1:1
route-target import 1:1
```

- To differentiate 10.0.0.0/8 in VPN-A from 10.0.0.0/8 in VPN-B
- 64-bit quantity
- Configured as ASN:YY or IPADDR:YY
  - Almost everybody uses ASN
- Purely to make a route unique
  - Unique route is now RD:Ipaddr (96 bits) plus a mask on the IPAddr portion
  - So customers don’t see each others routes

Route Target

```
!
ip vrf red
rd 1:1
route-target export 1:1
route-target import 1:1
```

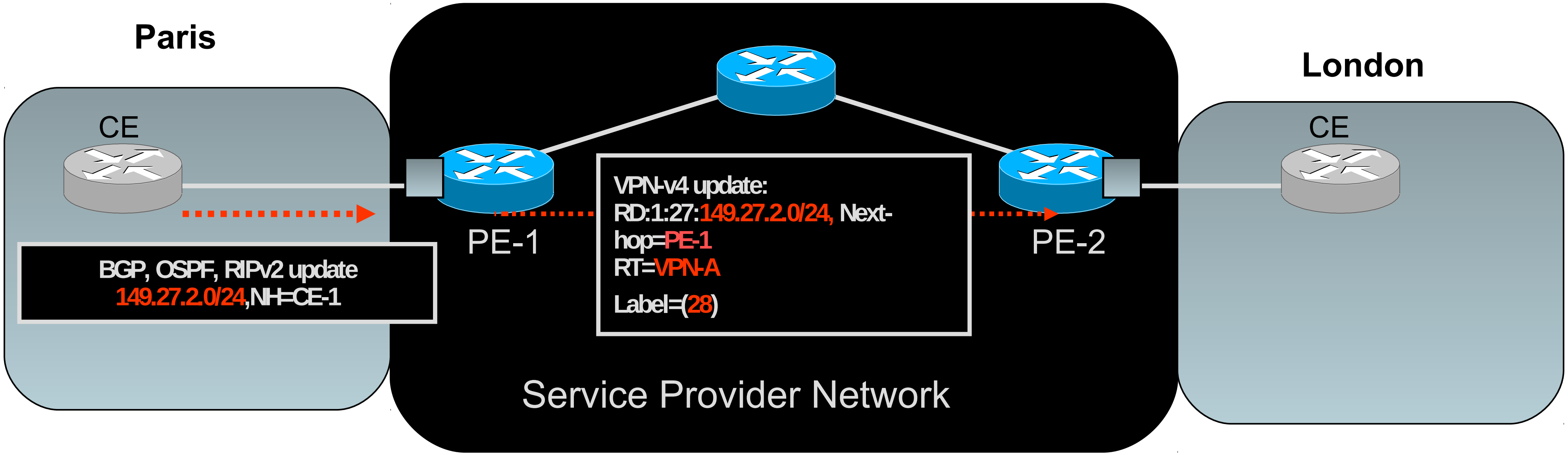
- Creates or adds to a list of VPN extended communities used to determine which routes are imported by a VRF
- To control policy about who sees what routes
- 64-bit quantity (2 bytes type, 6 bytes value)
- Carried as an extended community
- Typically written as ASN:YY
- Each VRF ‘imports’ and ‘exports’ one or more RTs
    - Exported RTs are carried in VPNv4 BGP
    - Imported RTs are local to the box
- A PE that imports an RT installs that route in its routing table
- Example: Each VRF in VPN A has the same route target in their import list and export list. Each VPN A VRF accepts only received routes that have this route target attached. Because this route target is attached to each route advertised by VPN A VRFs, every site in VPN A accepts routes only from other sites in VPN A.

VPNv4

- In BGP for IP, 32-bit address + mask makes a unique announcement
- In BGP for MPLS-VPN, (64-bit RD + 32-bit address) + 32-bit mask makes a unique announcement
- Since the route encoding is different, need a different address family in BGP
- VPNv4 = VPN routes for IPv4
  - As opposed to IPv4 or IPv6 or multicast-RPF, etc…
- VPNv4 announcement carries a label with the route
  - “If you want to reach this unique address, get me packets with this label on them”

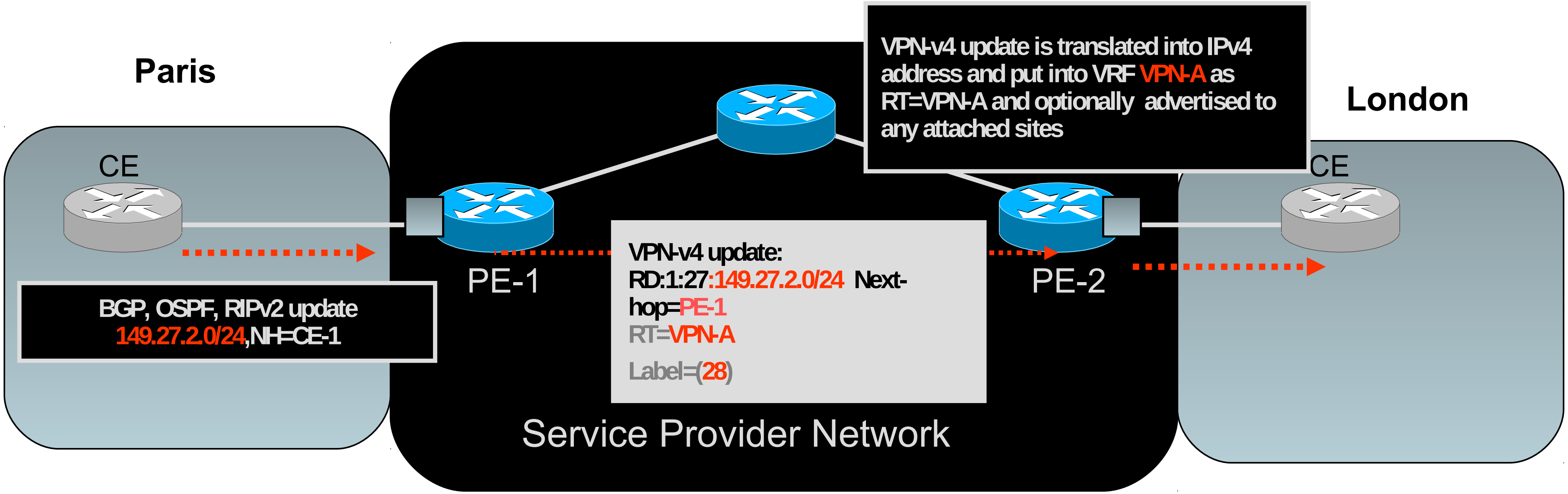
MPLS Layer-3 VPN - Operation Example

VRF Population of MP-BGP



- PE routers translate into VPN-V4 route
  - Assigns an RD, Site of Origin - SoO (if configured) and RT based on configuration
  - Re-writes Next-Hop attribute (to PE loopback)
  - Assigns a label based on VRF and/or interface
  - Sends MP-BGP update to all PE neighbors

VRF Population of MP-BGP

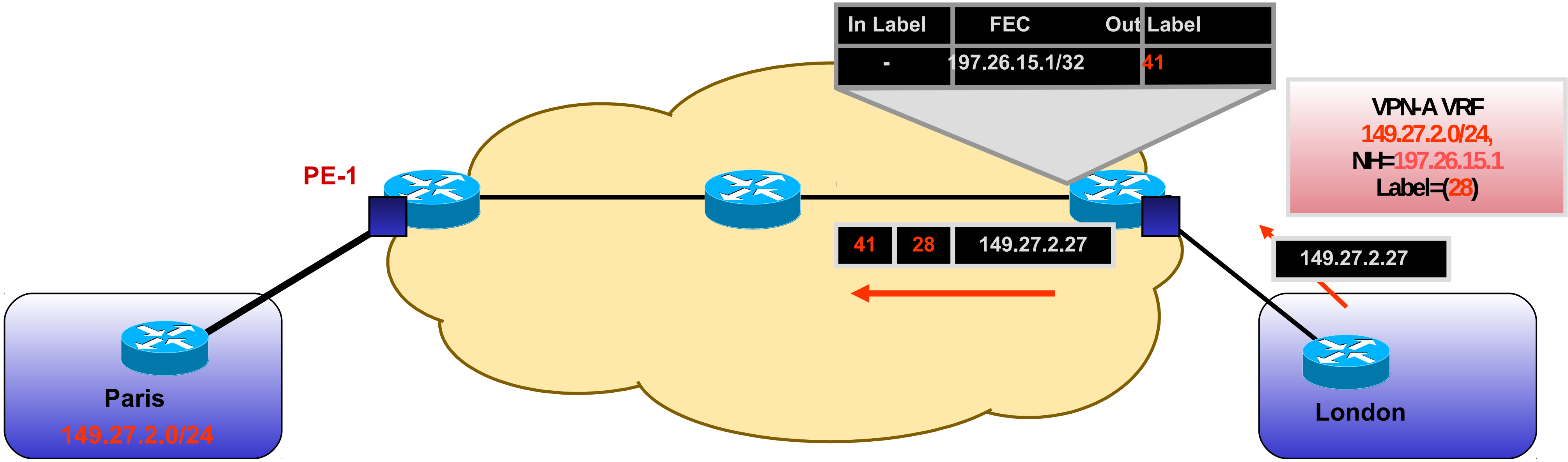

Paris
CE
BGP, OSPF, RIPv2 update
149.27.2.0/24,NH=CE-1
PE-1
VPN-v4 update:
RD:1:27:149.27.2.0/24 Next-hop=PE-1
RT=VPN-A
Label=(28)
PE-2
Service Provider Network
VPN-v4 update is translated into IPv4 address and put into VRF VPN-A as RT=VPN-A and optionally advertised to any attached sites
London
CE

- **Receiving PE routers translate to IPv4**
  - **Insert the route into the VRF identified by the RT attribute (based on PE configuration)**
- **The label associated to the VPN-V4 address will be set on packets forwarded towards the destination**

MPLS/VPN Packet Forwarding

- **Between PE and CE, regular IP packets (currently)**
- **Within the provider network—label stack**
  - **Outer label: “get this packet to the egress PE”**
  - **Inner label: “get this packet to the egress CE”**
- **MPLS nodes forward packets based on TOP label!!!**
  - any subsequent labels are ignored
- **Penultimate Hop Popping procedures used one hop prior to egress PE router (shown in example)**

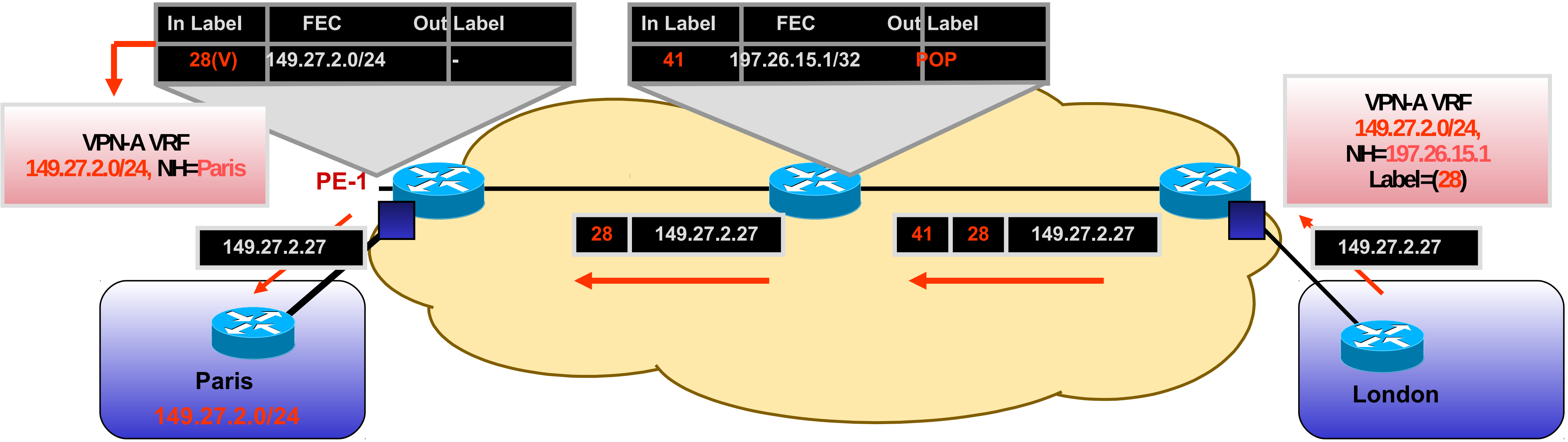
MPLS/VPN Packet Forwarding



In Label	FEC	Out Label
-	197.26.15.1/32	41

- Ingress PE receives normal IP packets
- PE router performs IP Longest Match from VPN FIB (Forwarding Table), finds iBGP next-hop and imposes a stack of labels <IGP, VPN>

MPLS/VPN Packet Forwarding

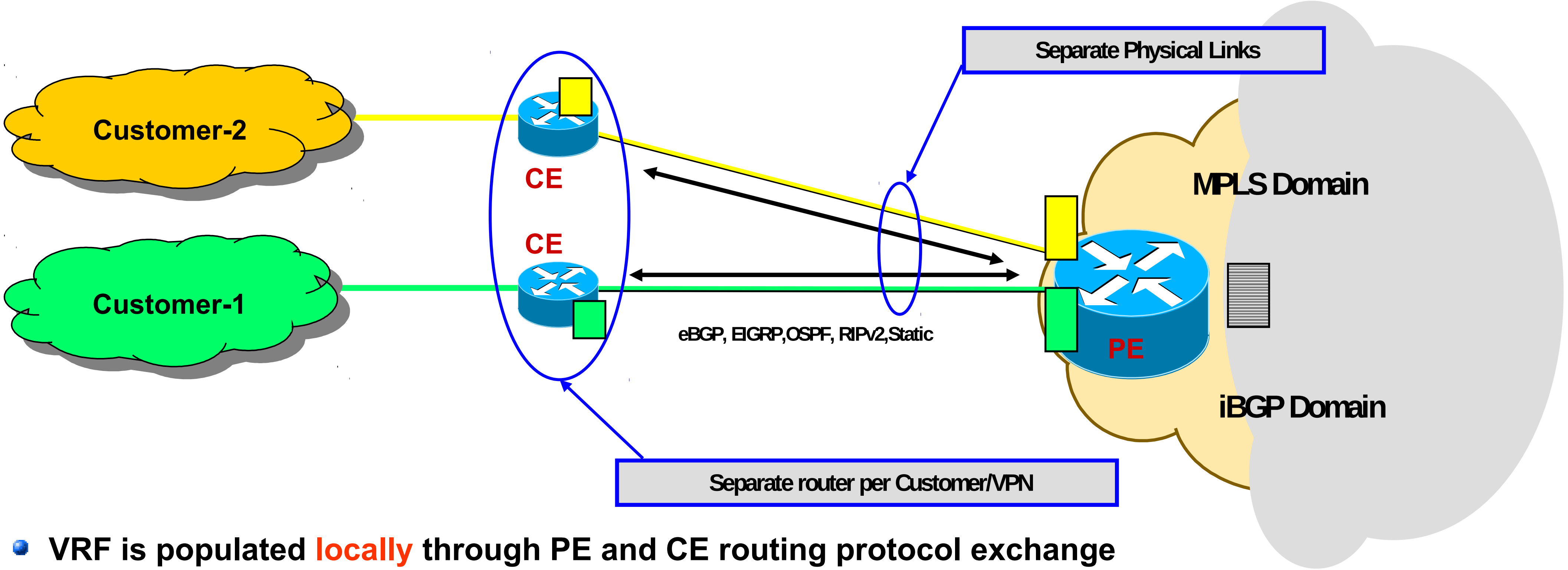


- **Penultimate PE router removes the IGP label**
  - Penultimate Hop Popping procedures (implicit-null label)
- **Egress PE router uses the VPN label to select which VPN/CE to forward the packet to**
- **VPN label is removed and the packet is routed toward the VPN site**

Things to Note

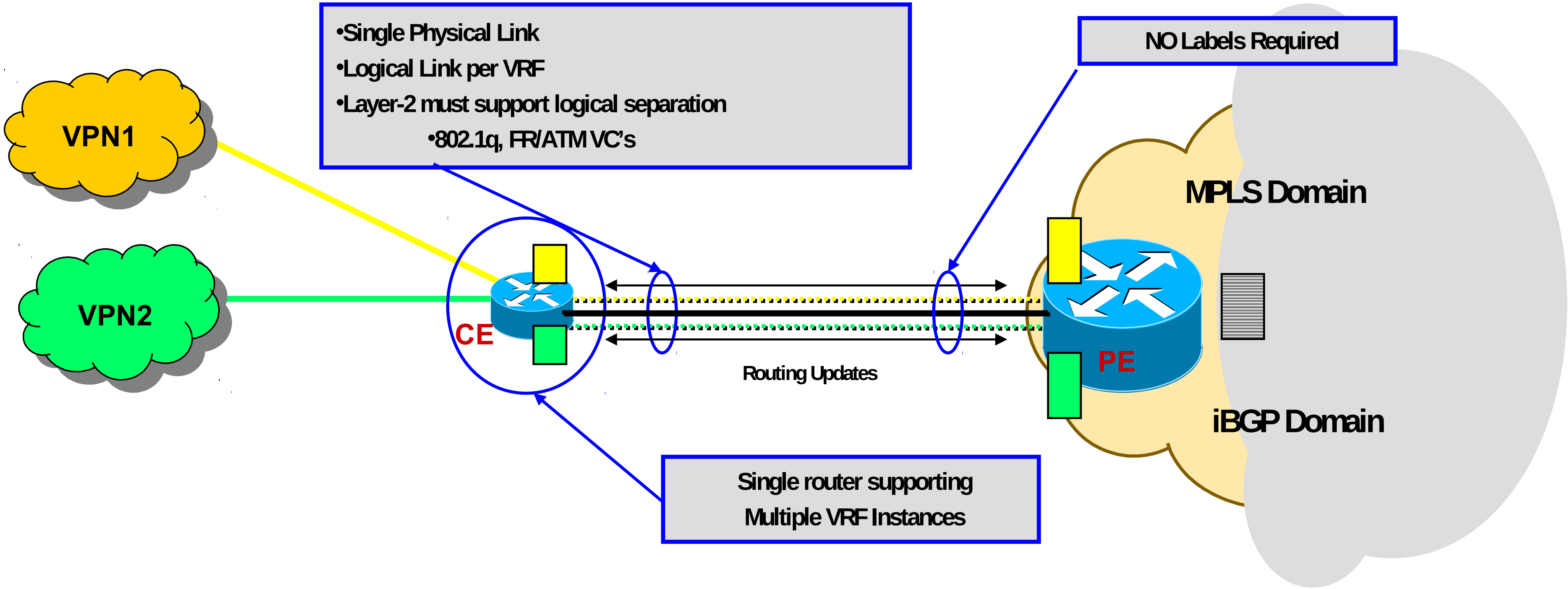
- Core does not run VPNv4 BGP!
  - Same principle can be used to run a BGP-free core for an IP network
- CE does not know it's in an MPLS-VPN
- Outer label is from LDP/RSVP
  - Getting packet to egress PE is mutually independent to MPLS-VPN
- Inner label is from BGP
  - Inner label is there so the egress PE can have the same network in multiple VRFs

VRF Route Population



- **VRF is populated locally through PE and CE routing protocol exchange**
  - RIP Version 2, OSPF, BGP-4, EIGRP, & Static routing
  - “connected” is also supported (i.e. Default-gateway is PE)
- **Separate routing context for each VRF**
  - routing protocol context (BGP-4 & RIP V2)
  - separate process (OSPF)

Multi-VRF CE (VRF-lite)



•Single Physical Link
•Logical Link per VRF
•Layer-2 must support logical separation
•802.1q, FR/ATM VC's

NO Labels Required

Single router supporting Multiple VRF Instances

- Each VRF separation on the PE is extended to the CE
- Separation is maintained via layer-2 transport that support "logical" separation (e.g. 802.1Q, FR/ATM VC's)
- CE router must be capable of supporting VRF's
- CE is not required to support MPLS labels
- Routing protocol options from CE-PE remain the same (e.g. BGP, RIPv2, OSPF, EIGRP, static)

Customers Connecting to a Layer-3 VPN Service

- What routing protocol is supported by the carrier (CE-PE)?
- What address space do they allow for CE-PE subnet?
- What layer-2 transport is required/supported from CE-PE?
- Do they provide a QoS SLA?
- Concerning QoS, do they require DSCP or ToS settings from the CE to their PE?
- Do they manipulate DSCP/ToS based on congestion in their network?
- What other services do they have on their roadmap of "Service Offerings" (Example: IPv6, IP Multicast, Tighter QoS SLA offering, other??)
- Understand the resiliency in the core
- Do they offer LEC (Local Exchange Carrier) diversification or "bypass"?

MPLS Layer-2 Transport

AToM - Any Transport Over MPLS

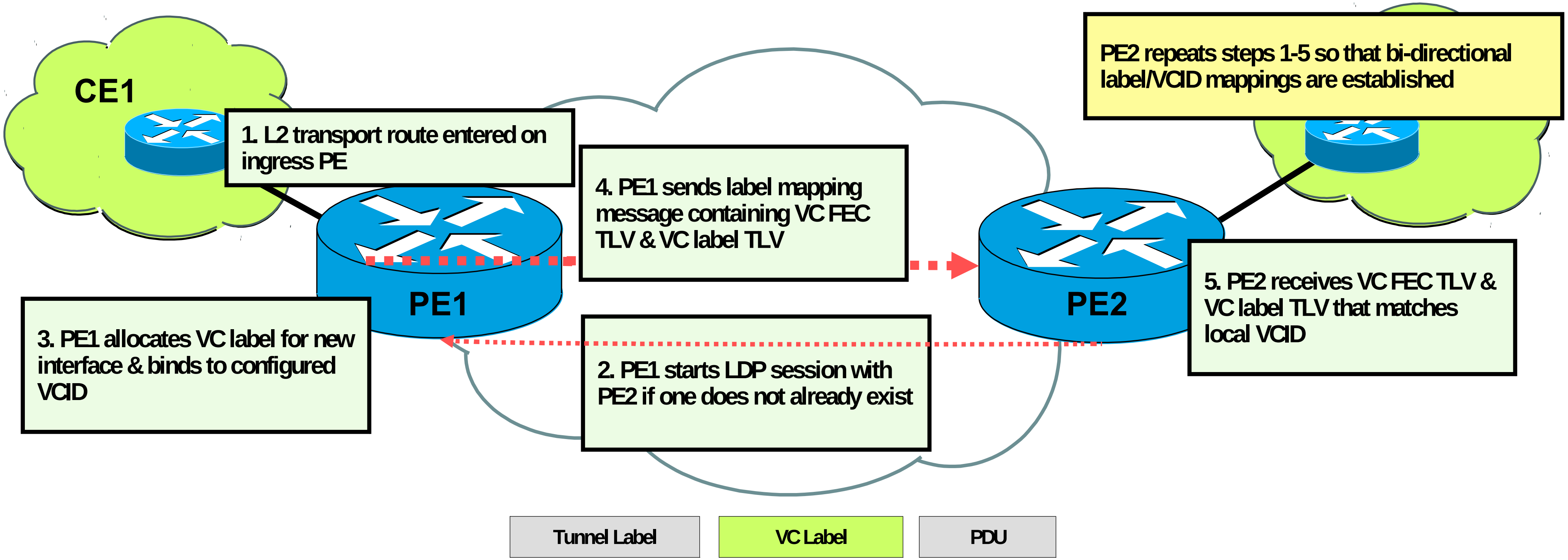
Motivation for AToM

- Protect existing investment while building packet core
  - Frame Relay, ATM and Ethernet
  - Non-IP protocols – SNA, IPX
- Trunk customer traffic
  - Trunk customer's IGP across the provider backbone
  - Especially when the customer is connecting over disparate media
- Provider devices forward customer packets based on Layer 2 information
  - Circuits (ATM/FR), MAC address
  - CPE-based Tunnels (e.g. IPSEC) analogous to circuits
  - Possibility of a new service (VPLS – emulated LAN)
- Good fit for customers that either
  - Simply want connectivity
  - Have non-IP protocols

AToM – VC Information Exchange

- VC labels are exchanged across a directed LDP session between PE routers
    - Carried in Generic Label TLV (Type, Length, Value within LDP Label Mapping Message (RFC3036 -LDP)
- New LDP FEC element defined to carry VC information
    - FEC element type ‘128 – Virtual Circuit FEC Element’;
    - Carried within LDP Label Mapping Message
- VC information exchanged using Downstream Unsolicited label distribution procedures

AToM – Label Mapping Exchange



1. L2 transport route entered on ingress PE

2. PE1 starts LDP session with PE2 if one does not already exist

3. PE1 allocates VC label for new interface & binds to configured VCID

4. PE1 sends label mapping message containing VC FEC TLV & VC label TLV

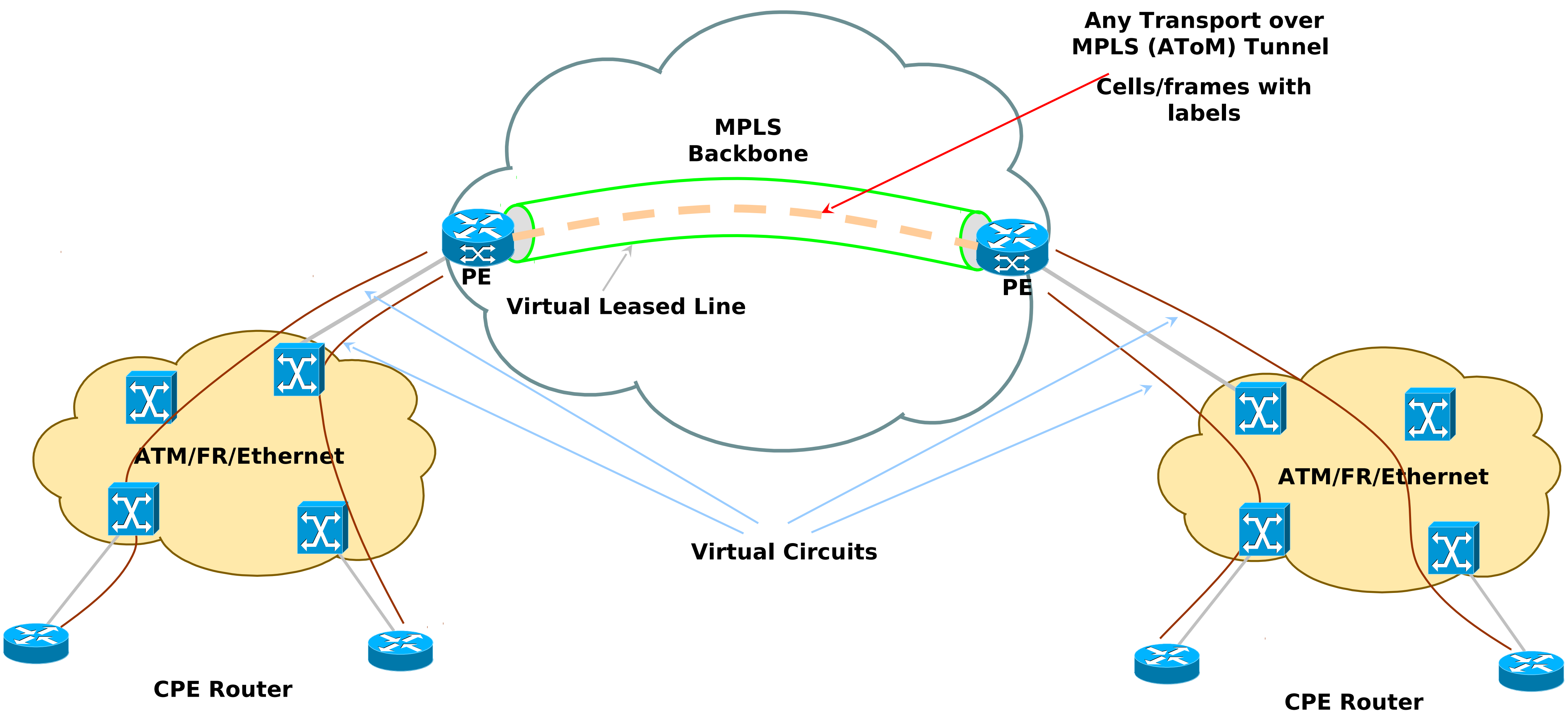
5. PE2 receives VC FEC TLV & VC label TLV that matches local VCID

PE2 repeats steps 1-5 so that bi-directional label/VCID mappings are established

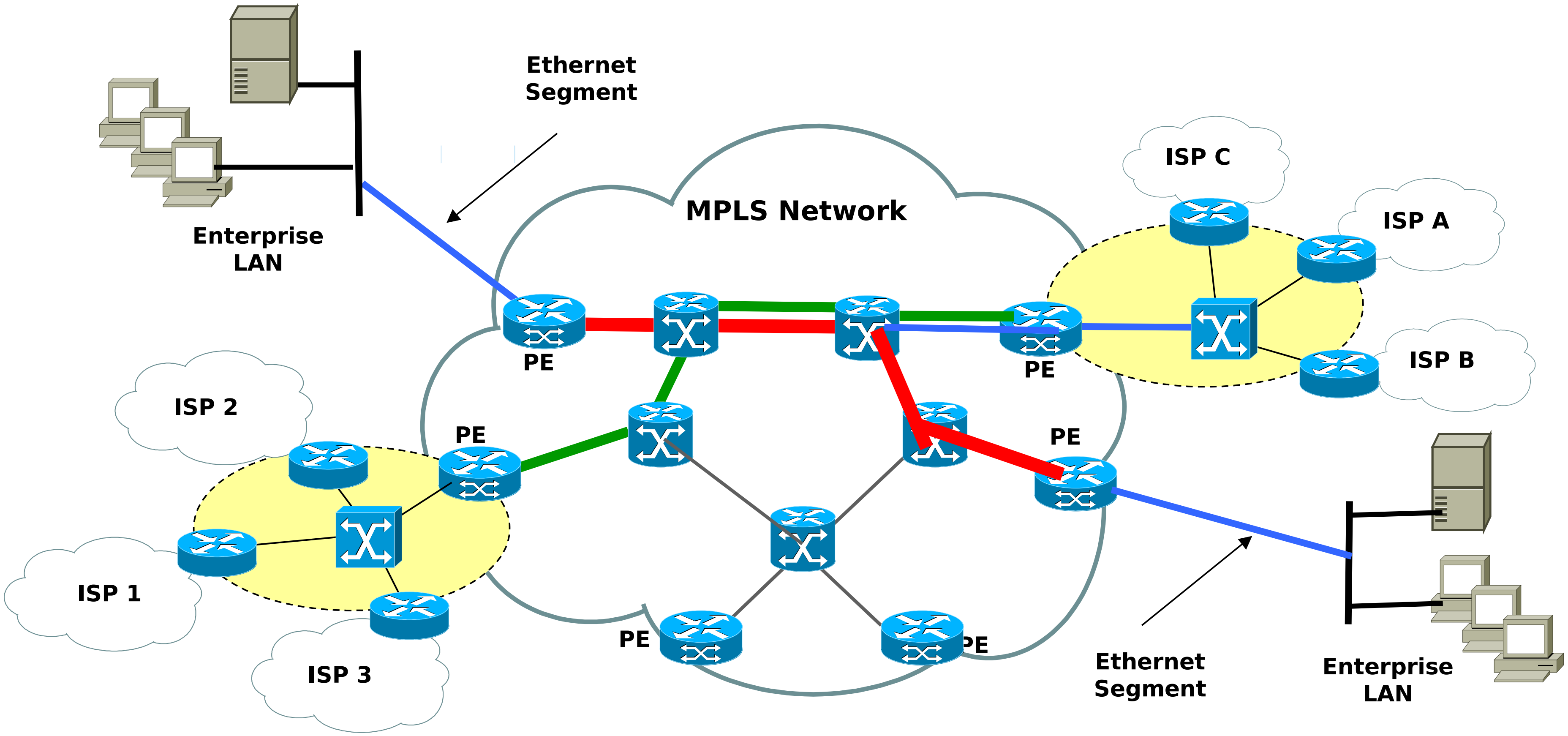
Bi-directional Label/VCID mapping exchange

Layer 2 Integration – ATM/FR over MPLS

QoS Options, Mapping: L2→IP→EXP



Layer 2 Integration - Ethernet over MPLS



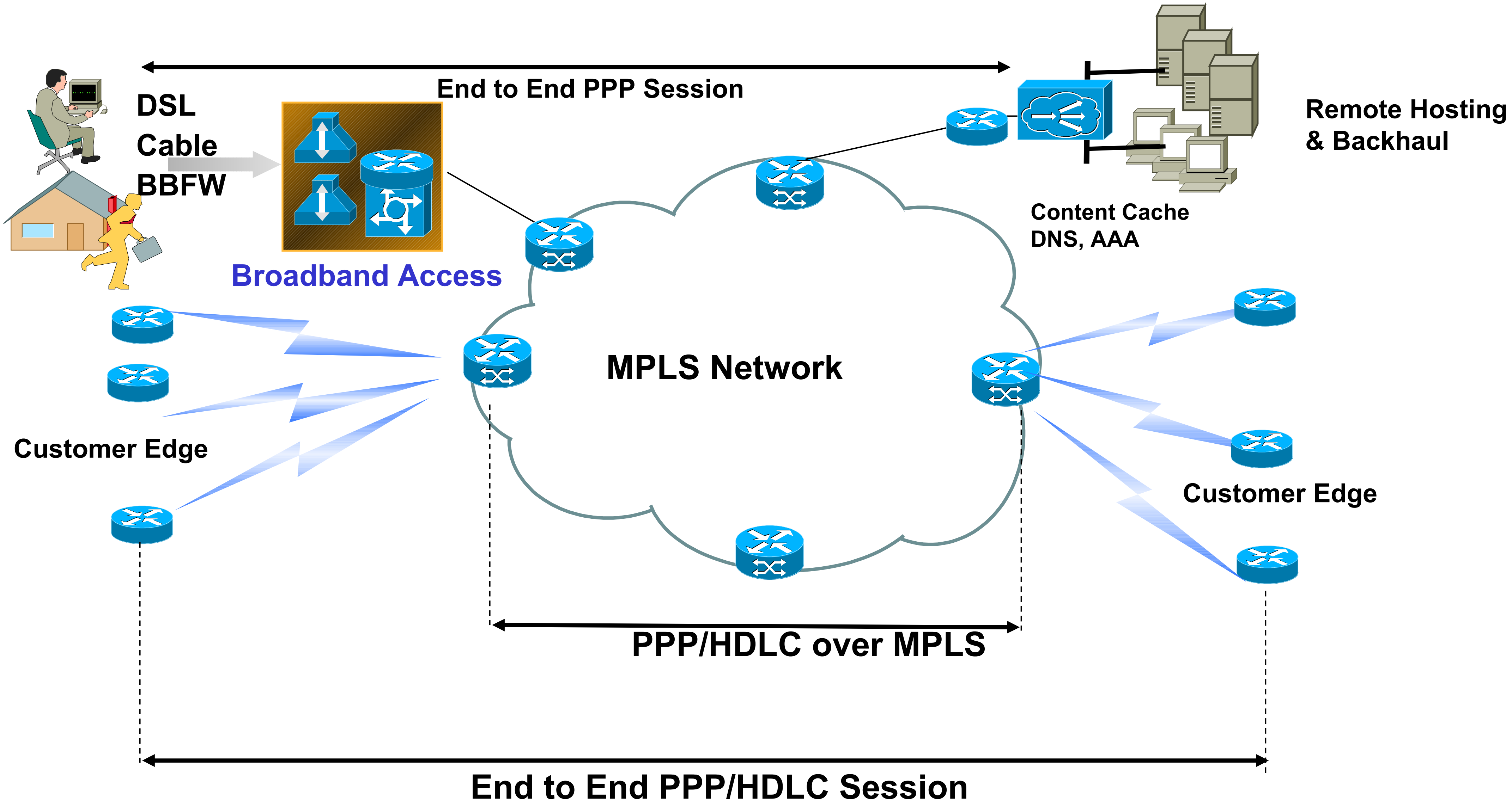
- **Port-mode**

  Allows a frame coming into an interface to be packed into an MPLS packet

- **VLAN-mode**

  Forwards frames from a SRC 802.1Q VLAN to a DST 802.1Q VLAN

PPP/HDLC over MPLS



MPLS Traffic Engineering

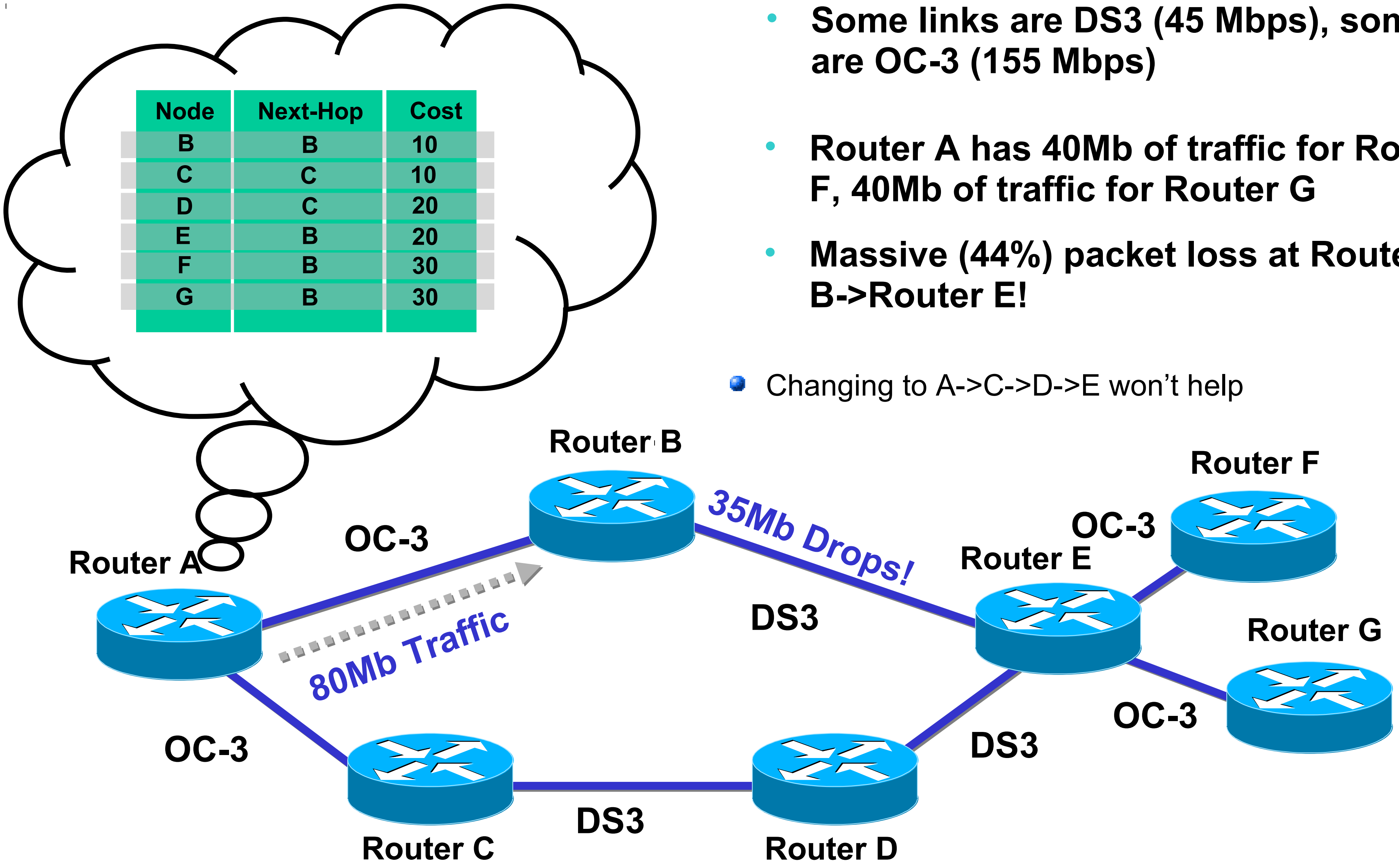
Traffic Engineering - Theory

- MPLS-TE was designed to move traffic along a path other than the IGP shortest path
  - Bring ATM/FR traffic engineering abilities to an IP network
  - Avoid full IGP mesh and n(n – 1)/2 flooding
  - Bandwidth-aware connection setup
- Fast ReRoute (FRR) is emerging as another application of MPLS-TE
  - Bandwidth Protection: Allows for tighter control on bandwidth – packet loss, delay & jitter
  - Minimal packet loss (msec) when a link goes down
  - Can be used in conjunction with MPLS-TE for primary paths, can also be used in standalone
- Provide Virtual Leased Lines – DS-TE + QoS
  - Intelligent network infrastructure for better bandwidth guarantees (DS-TE, Online Bandwidth Protection, Voice VPNs etc)

The Problem with Shortest-Path

- **Some links are DS3 (45 Mbps), some are OC-3 (155 Mbps)**
- **Router A has 40Mb of traffic for Route F, 40Mb of traffic for Router G**
- **Massive (44%) packet loss at Router B->Router E!**

- Changing to A->C->D->E won't help

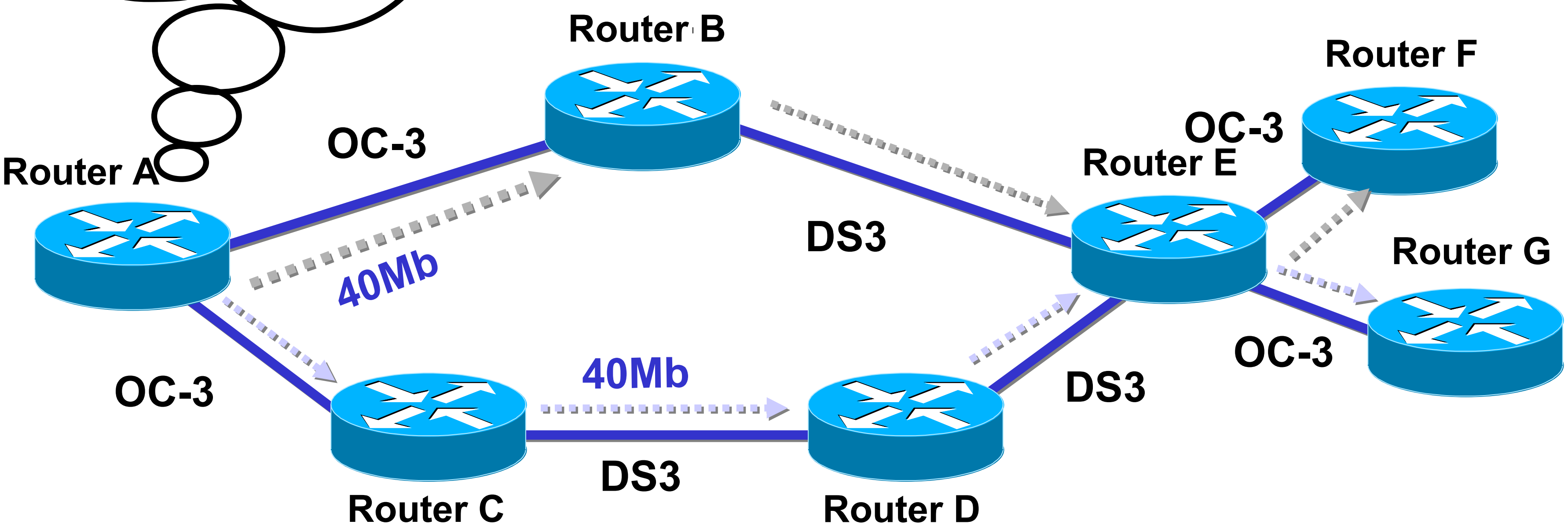


Node	Next-Hop	Cost
B	B	10
C	C	10
D	C	20
E	B	20
F	B	30
G	B	30

Path Calculation

Node	Next-Hop	Cost
B	B	10
C	C	10
D	C	20
E	B	20
F	Tunnel 0	30
G	Tunnel 1	30

- PCALC takes bandwidth, other constraints into account
- Link state protocol advertises “unreserved capacity”
- Constraints (required bandwidth and policy) are specified for a TE “trunk”
- End result: Bandwidth used more efficiently!



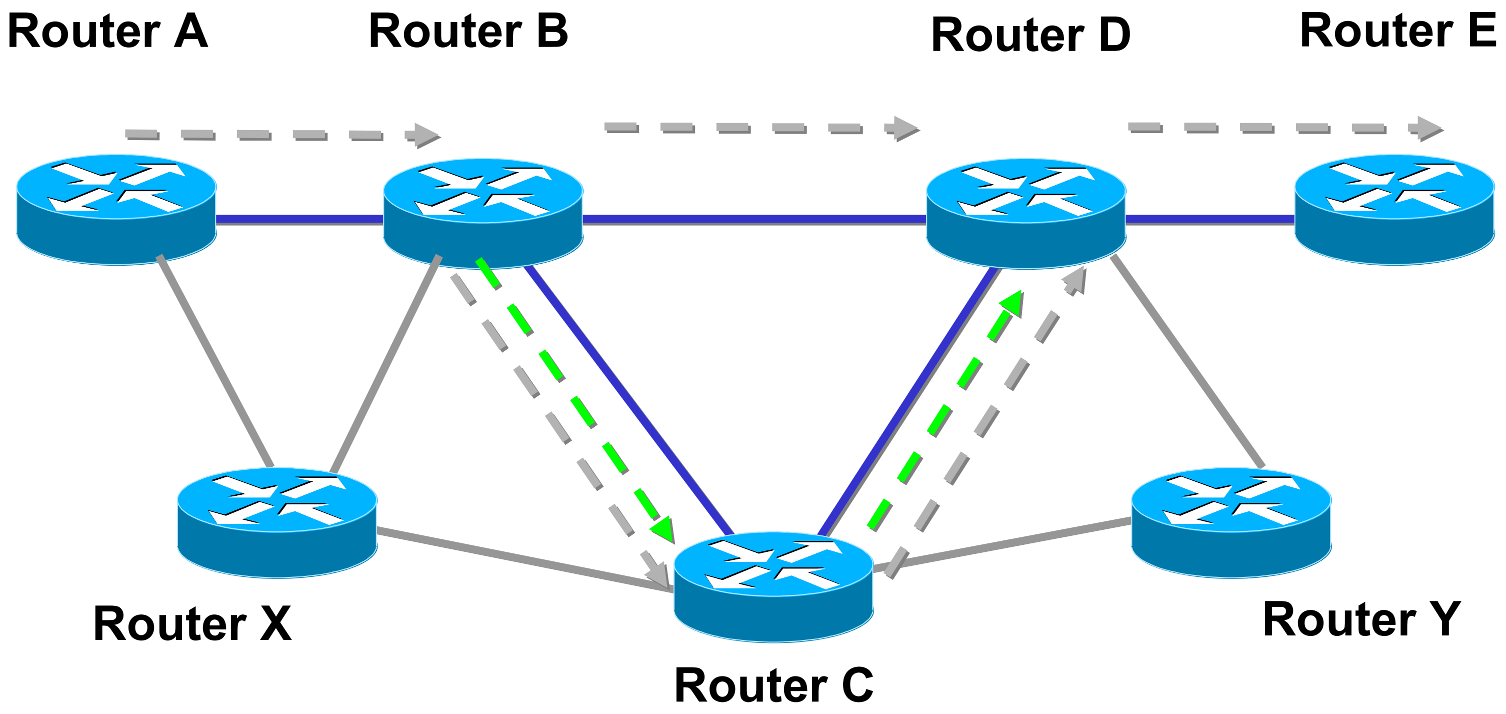
Forwarding Traffic Down a Tunnel

- There are three ways traffic can be forwarded down a TE tunnel
  - Auto-route
  - Static routes
  - Policy routing
- With the first two, MPLS-TE gets you unequal cost load balancing

Fast ReRoute

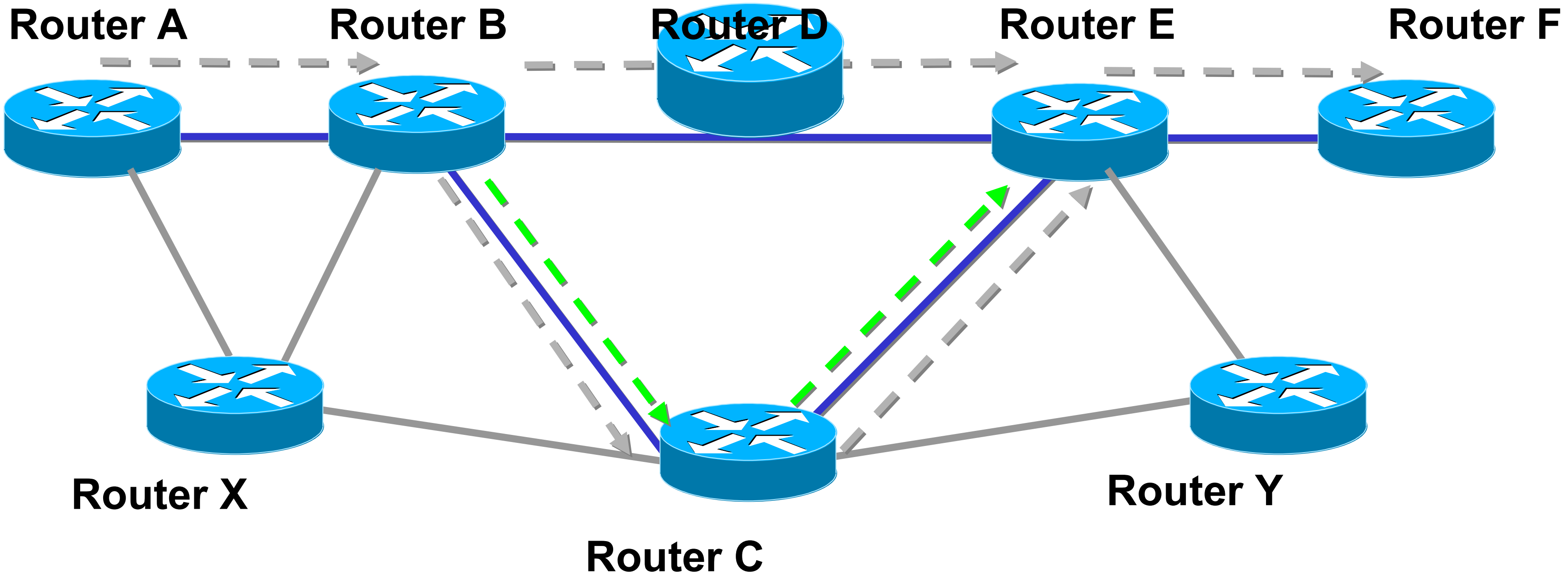
- FRR: A mechanism to minimize packet loss during a failure
- Pre-provision protection tunnels that carry traffic when a protected resource (link/node) goes down
- Use MPLS-TE to signal the FRR protection tunnels, taking advantage of the fact that MPLS-TE traffic doesn’t have to follow the IGP shortest path
- Used as a mechanism (along with DS-TE) for tight SLA offerings for “Guaranteed Bandwidth Services”

Link Protection*



- Primary Tunnel: A -> B -> D -> E
- BackUp Tunnel: B -> C -> D (Pre-provisioned)
- Recovery = ~50ms

Node Protection



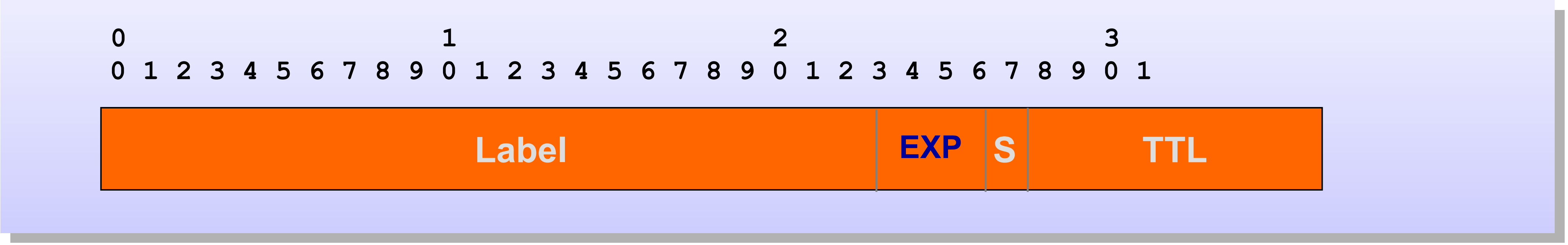
- Primary Tunnel: A -> B -> D -> E -> F
- BackUp Tunnel: B -> C -> E (Pre-provisioned)
- Recovery = ~100ms

MPLS QoS

DiffServ over MPLS

- MPLS doesn’t define a new QoS architecture
- Most of the work on MPLS QoS has focused on supporting current IP QoS architectures
- Same traffic conditioning and Per-Hop behaviors as defined by DiffServ

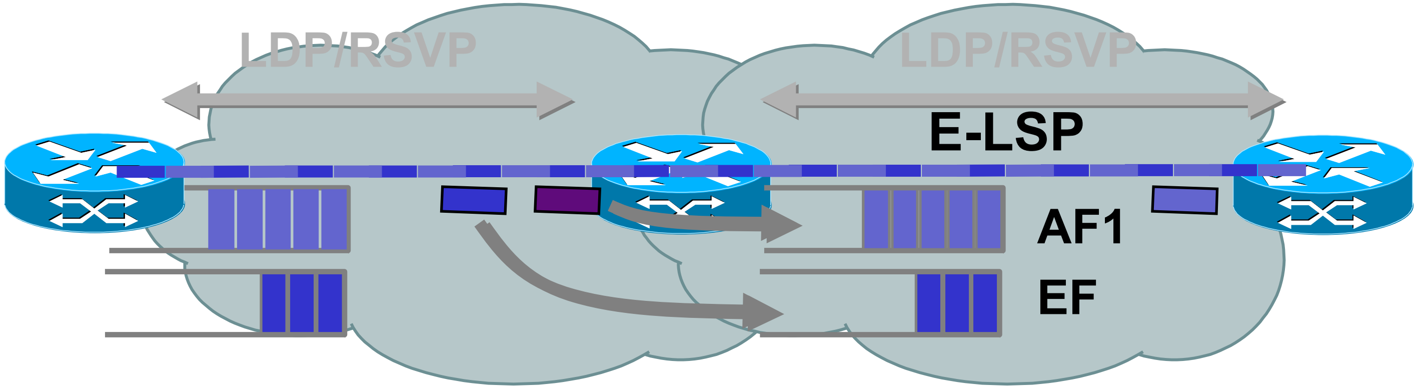
Label Header for Packet Media



Label	**20 bits**
EXP	**Experimental Field, 3 bits (Class of service information)**
S	**Bottom of Stack, 1 Bit**
TTL	**Time to Live, 8 Bits**

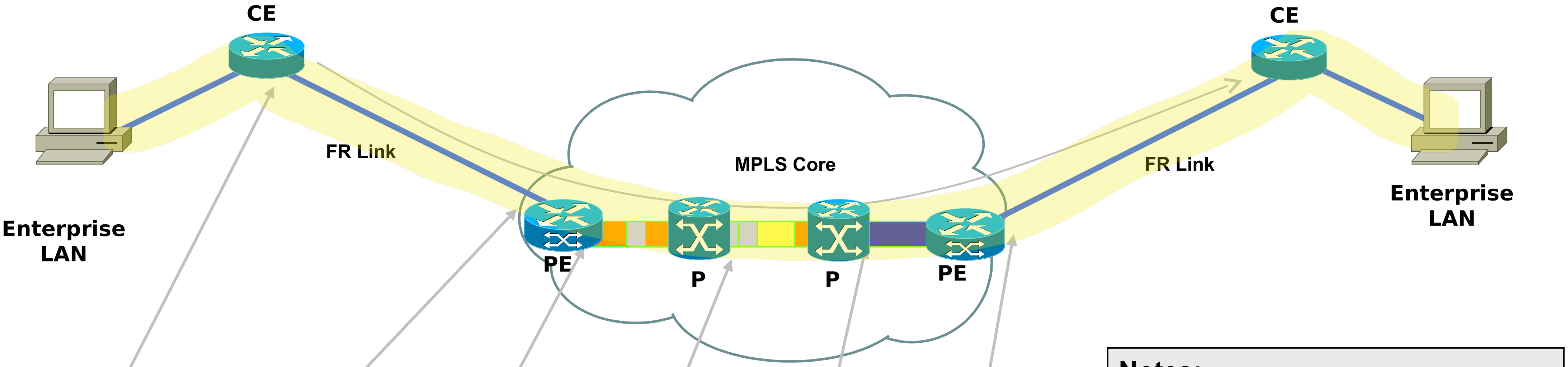
- Can be used over other layer-2 technologies
- Contains all information needed at forwarding time
- One 32-bit word per label
- EXP field size limitation by standards

Diff-Serv Support Over MPLS



- Diff-Serv is supported today over MPLS
  - RFC3270
  - Neither more nor less than “plain old” Diff-Serv
- Example above illustrates support of EF and AF1 on single E-LSP
  - EF (Expedited Forwarding) and AF1 (Assured Forwarding) packets travel on single LSP (single label) but are enqueued in different queues (different EXP values)

DiffServ MPLS QoS Implementation



CE Out
FR TS
LLQ
WRED
FRF.12
cRTP

PE In
Police
Mark

PE - P
LLQ
WRED

P - P
LLQ
WRED

P - PE
LLQ
WRED

PE Out
LLQ
WRED

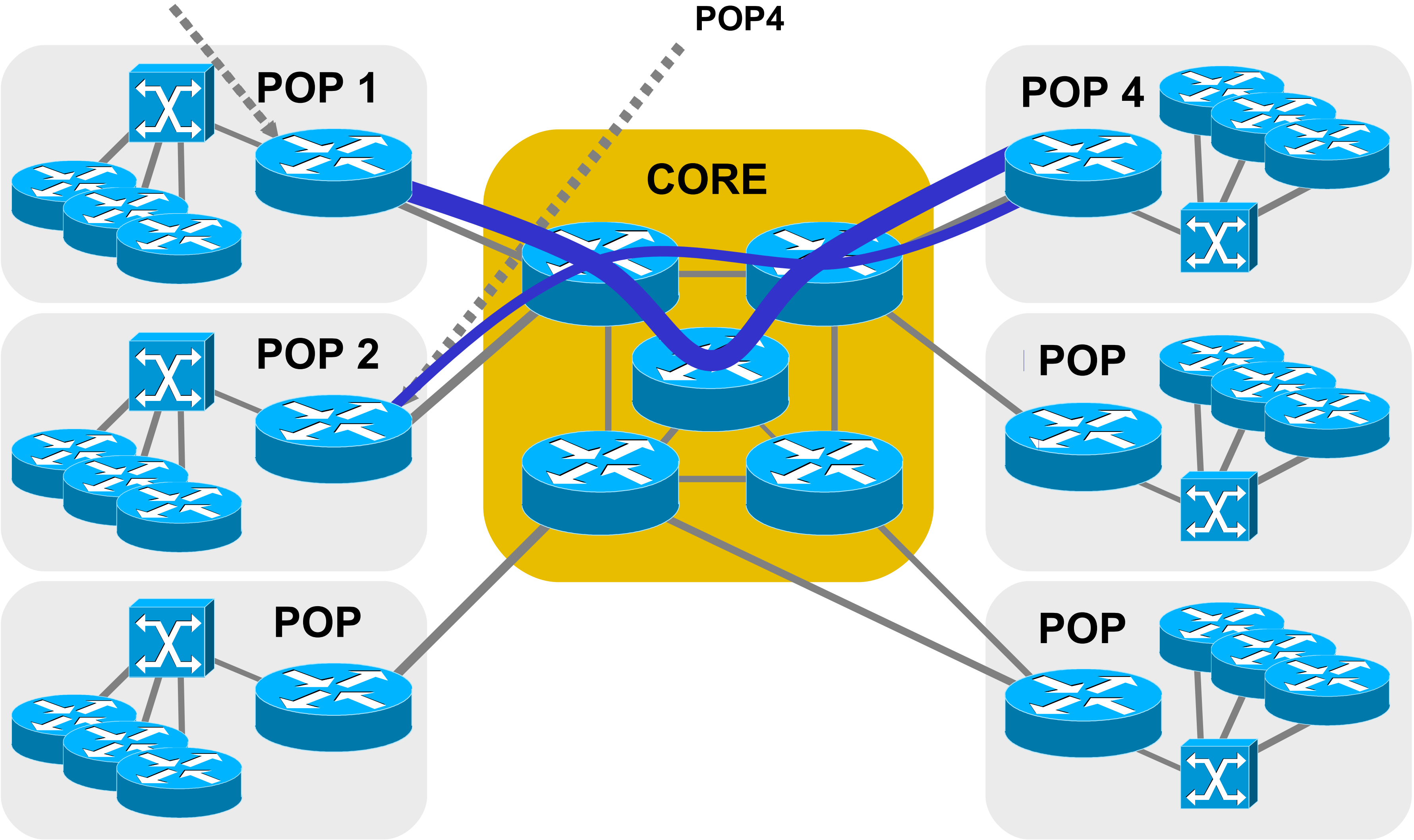
Notes:
-Traffic Classified by EXP
- Core is MPLS Frame-mode
- LLQ on MPLS packets
- WRED based on EXP
- No need for inbound policy in Core
-LLQ for Min B/W guarantee
-Unmanaged CE example shown

LLQ – Low Latency Queing

Relationship between MPLS TE and MPLS Diff-Serv

- Diff-Serv specified independently of Routing/Path Computation
- MPLS Diff-Serv (RFC3270) specified independently of Routing/Path Computation
- MPLS TE designed as tool to improve backbone efficiency independently of QoS:
  - MPLS TE compute routes for aggregates across all Classes
  - MPLS TE performs admission control over "global" bandwidth pool for all Classes (i.e., unaware of bandwidth allocated to each queue)
- MPLS TE and MPLS Diff-Serv:
  - can run simultaneously
  - can provide their own benefit (i.e. TE distributes aggregate load, Diff-Serv provides differentiation)
  - are unaware of each other (TE cannot provide its benefit on a per class basis such as CAC and constraint based routing)

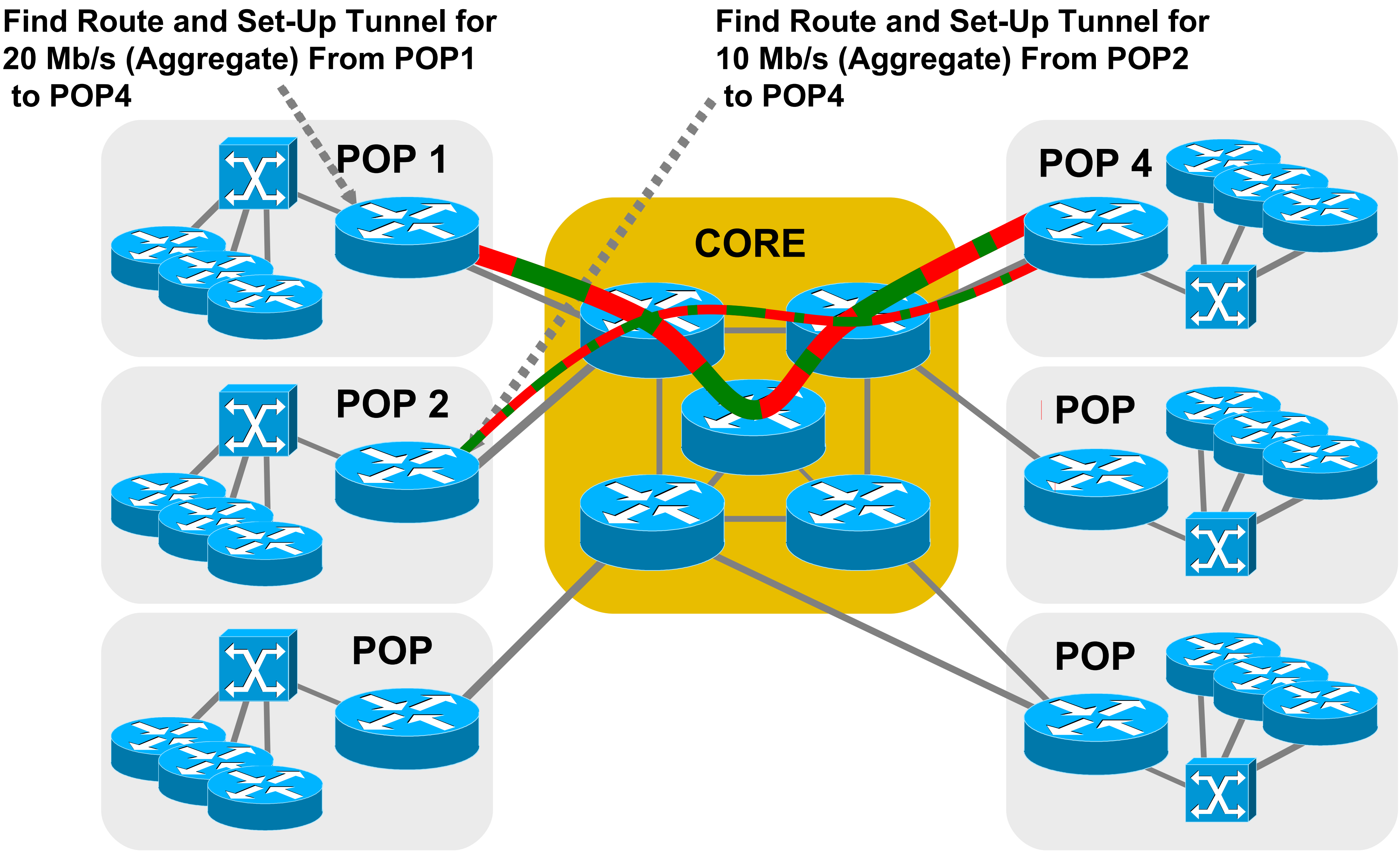
MPLS TE with Best Effort Network



Find Route and Set-Up Tunnel for 20 Mb/s (Aggregate) From POP1 to POP4

Find Route and Set-Up Tunnel for 10 Mb/s (Aggregate) From POP2 to POP4

MPLS TE with DiffServ Network



Find Route and Set-Up Tunnel for 20 Mb/s (Aggregate) From POP1 to POP4

Find Route and Set-Up Tunnel for 10 Mb/s (Aggregate) From POP2 to POP4

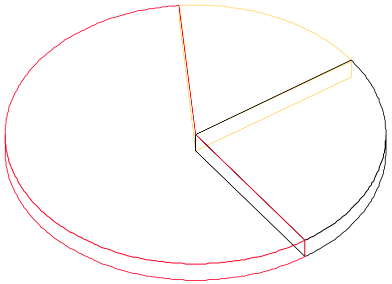
DiffServ aware Traffic Engineering (DS-TE)

- DS-TE is more than MPLS TE + MPLS DiffServ
- DS-TE makes MPLS TE aware of DiffServ:
  - DS-TE establishes separate tunnels for different classes
  - DS-TE takes into account the “bandwidth” available to each class (e.g. to queue)
  - DS-TE takes into account separate engineering constraints for each class
    - e.g. I want to limit Voice traffic to 70% of link max, but I don’t mind having up to 100% of BE traffic.
    - e.g I want overbook ratio of 1 for voice but 3 for BE
- DS-TE ensures specific QoS level of each DiffServ class is achieved

DS-TE Configuration Example
Tunnel Midpoint

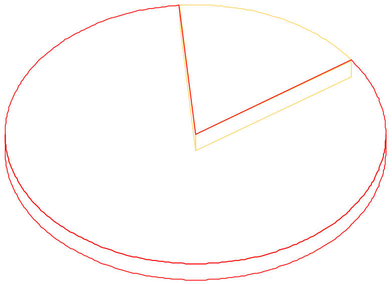
Data Plane

Bandwidth Allocation



Control Plane

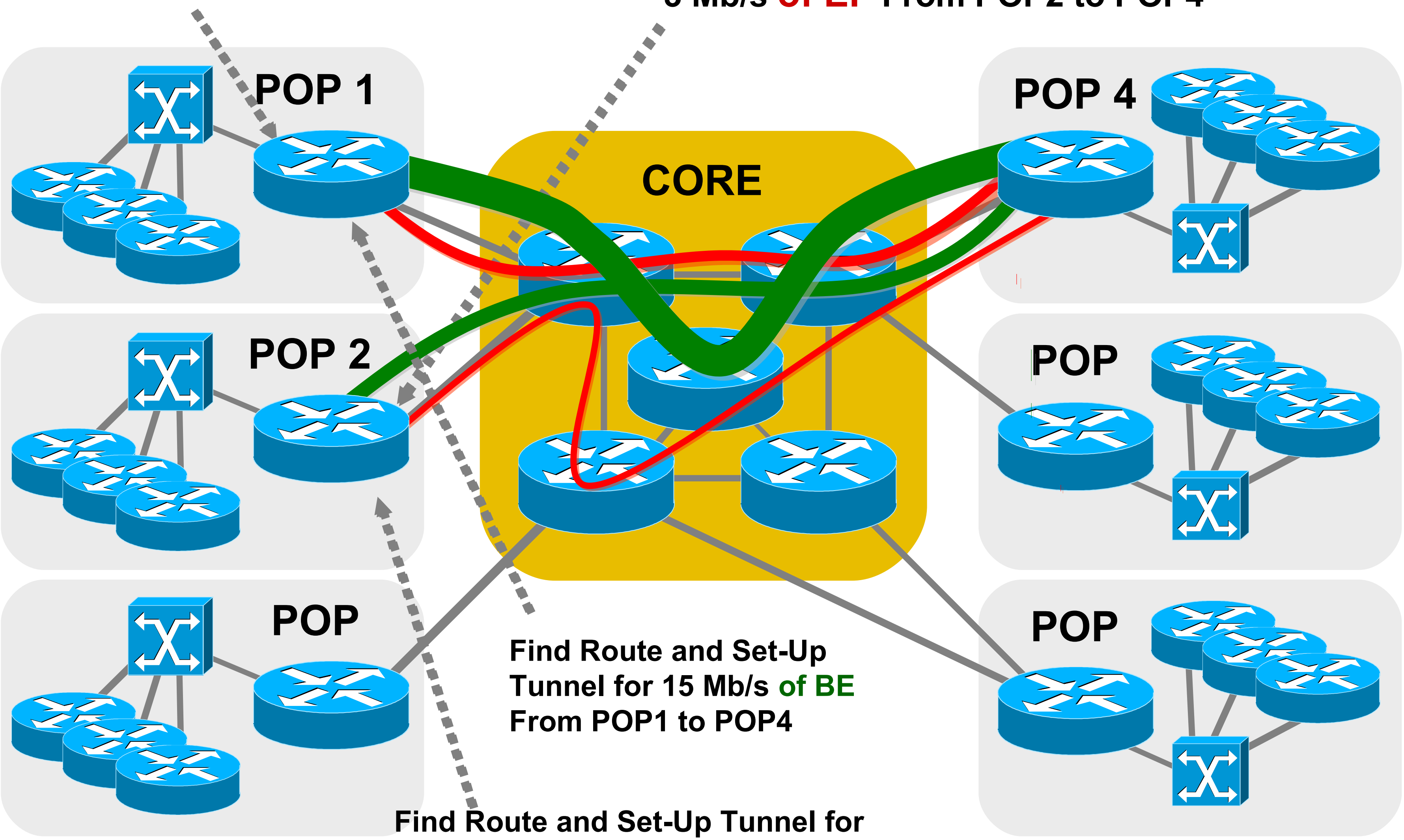
Bandwidth Allocation



```
!
class-map match-all PREMIUM
 match mpls experimental 5
!
class-map match-all BUSINESS
  match mpls experimental 3 4
!
policy-map OUT-POLICY
  class GOLD
    priority 16384
  class SILVER
    bandwidth 65536
    random-detect
  class class-default
    random-detect
!
interface POS1/0
 ip address 10.150.1.1 255.255.255.0
 ip rsvp bandwidth 155000 155000 sub-pool 16384
 service-policy output OUT-POLICY
 mpls traffic-eng tunnels
 mpls ip
!
```

Bandwidth Allocation

MPLS DS-TE with DiffServ Network



Find Route and Set-Up Tunnel for 5 Mb/s of EF From POP1 to POP4

Find Route and Set-Up Tunnel for 3 Mb/s of EF From POP2 to POP4

Find Route and Set-Up Tunnel for 15 Mb/s of BE From POP1 to POP4

Find Route and Set-Up Tunnel for 7 Mb/s of BE From POP2 to POP4

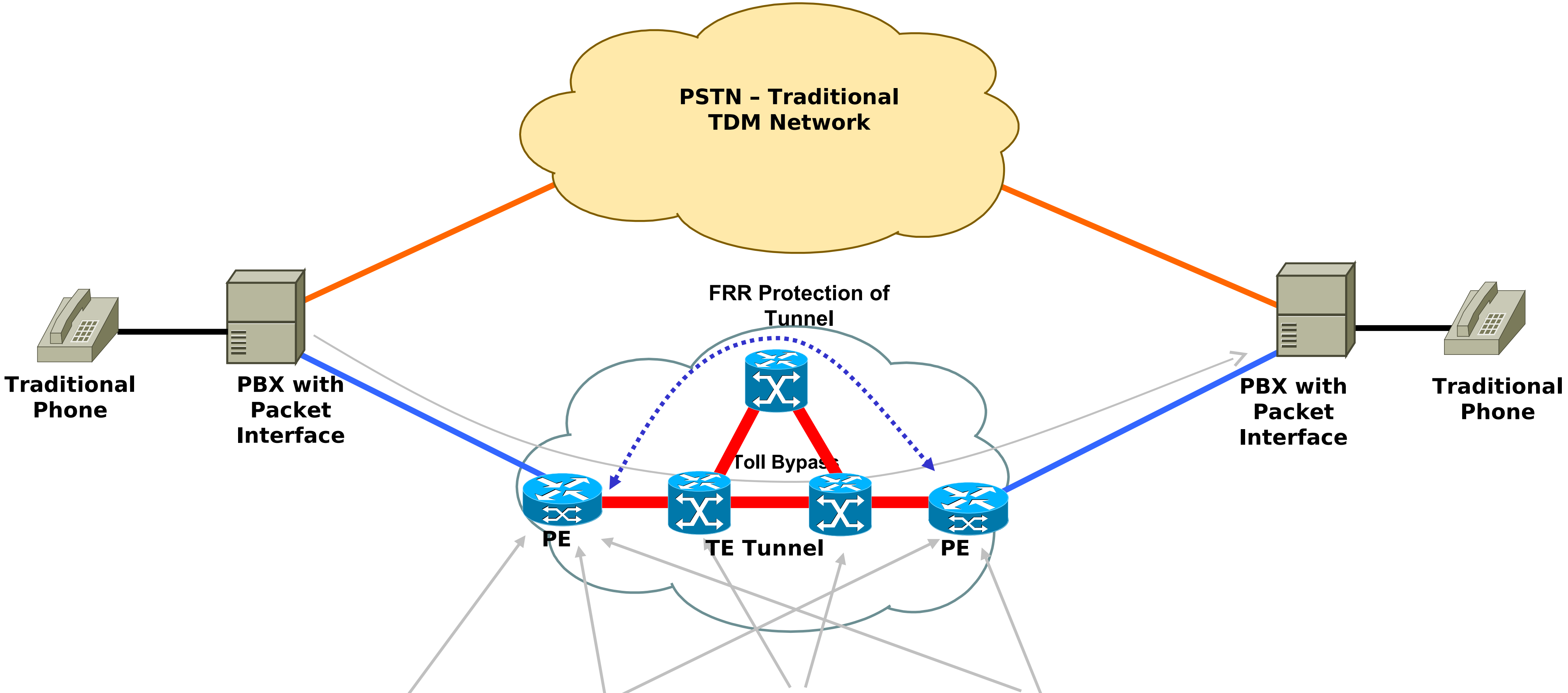
MPLS QoS Applications for Multi-Service

MPLS QoS Applications for Multi-Service

- MPLS QoS General
  - MPLS Diffserv
  - MPLS TE
  - MPLS FRR (applies to strict QoS)
  - Diffserv-TE (DS-TE)

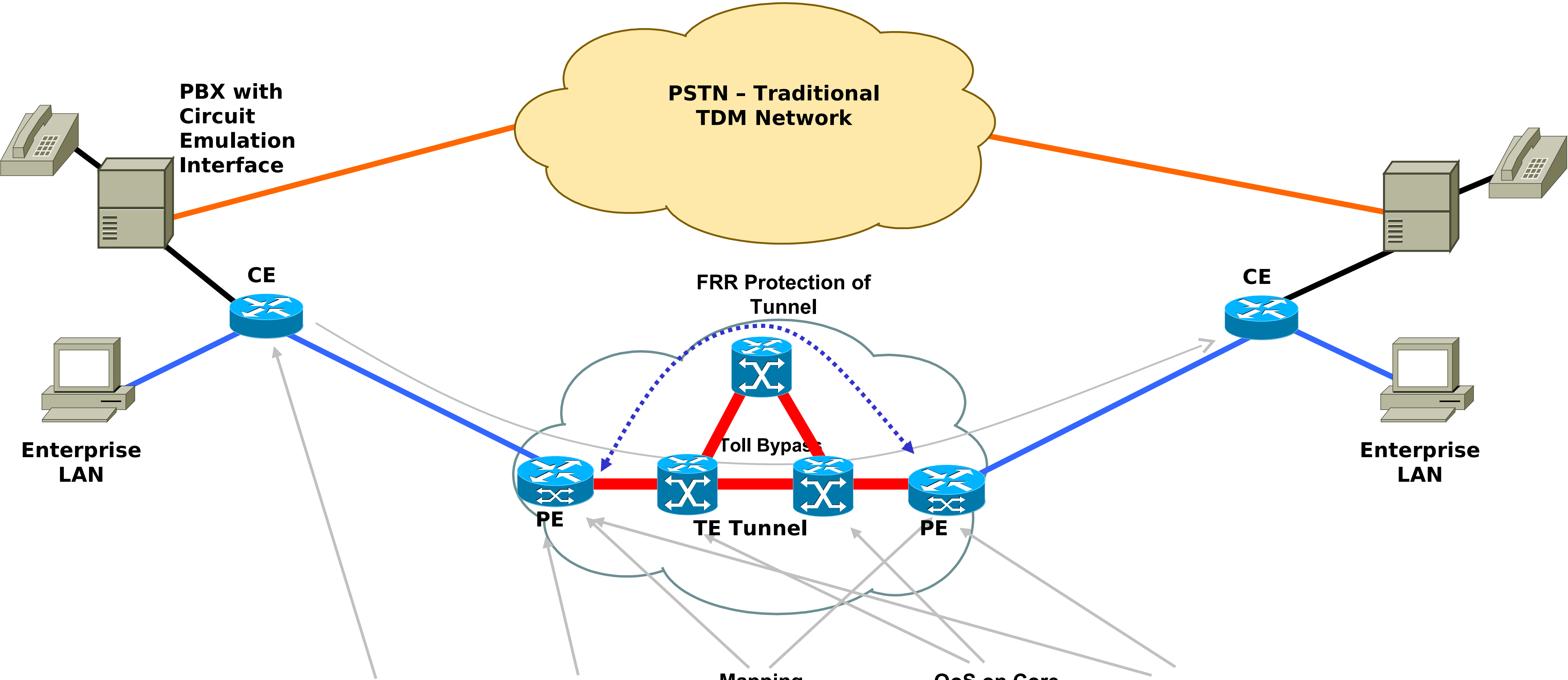
- Combination = Guaranteed Bandwidth Services
  - Applications
  - Voice Trunking over TE
  - Virtual Leased Line Services

Solution 1: Toll Bypass with Voice Network



Solution Requirements ⇒ QoS on PE Router **+** Mapping Traffic to Tunnels **+** QoS on Core Routers **+** TE or DS-TE

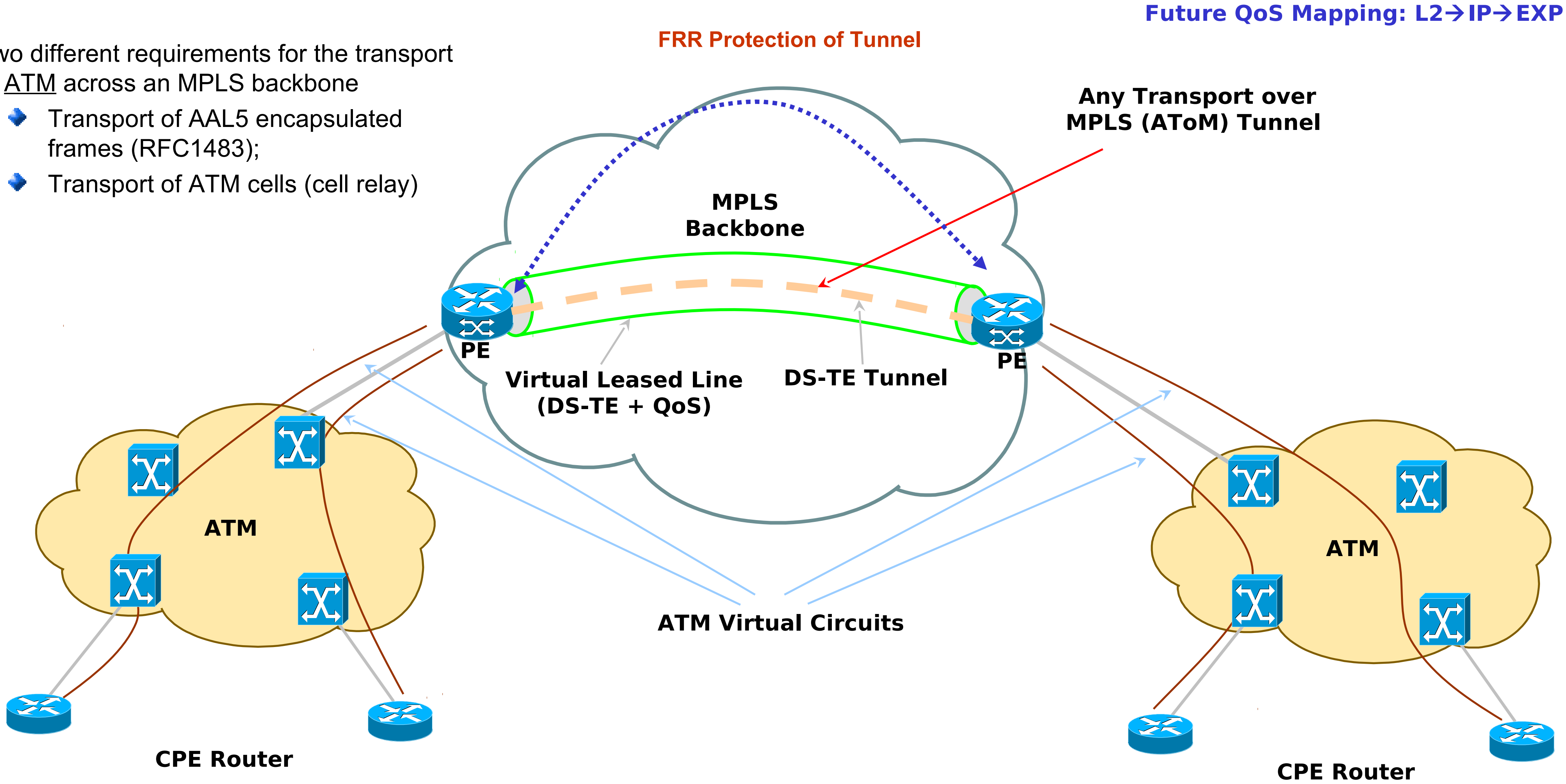
Solution 2: Toll Bypass with Voice/Data Converged Network



Solution Requirements ⇒ QoS on CE Router + QoS on PE Router + Mapping Traffic to Tunnels + QoS on Core Routers + TE or DS-TE

Solution 3: Virtual Leased Lines – ATM Networks Using AToM

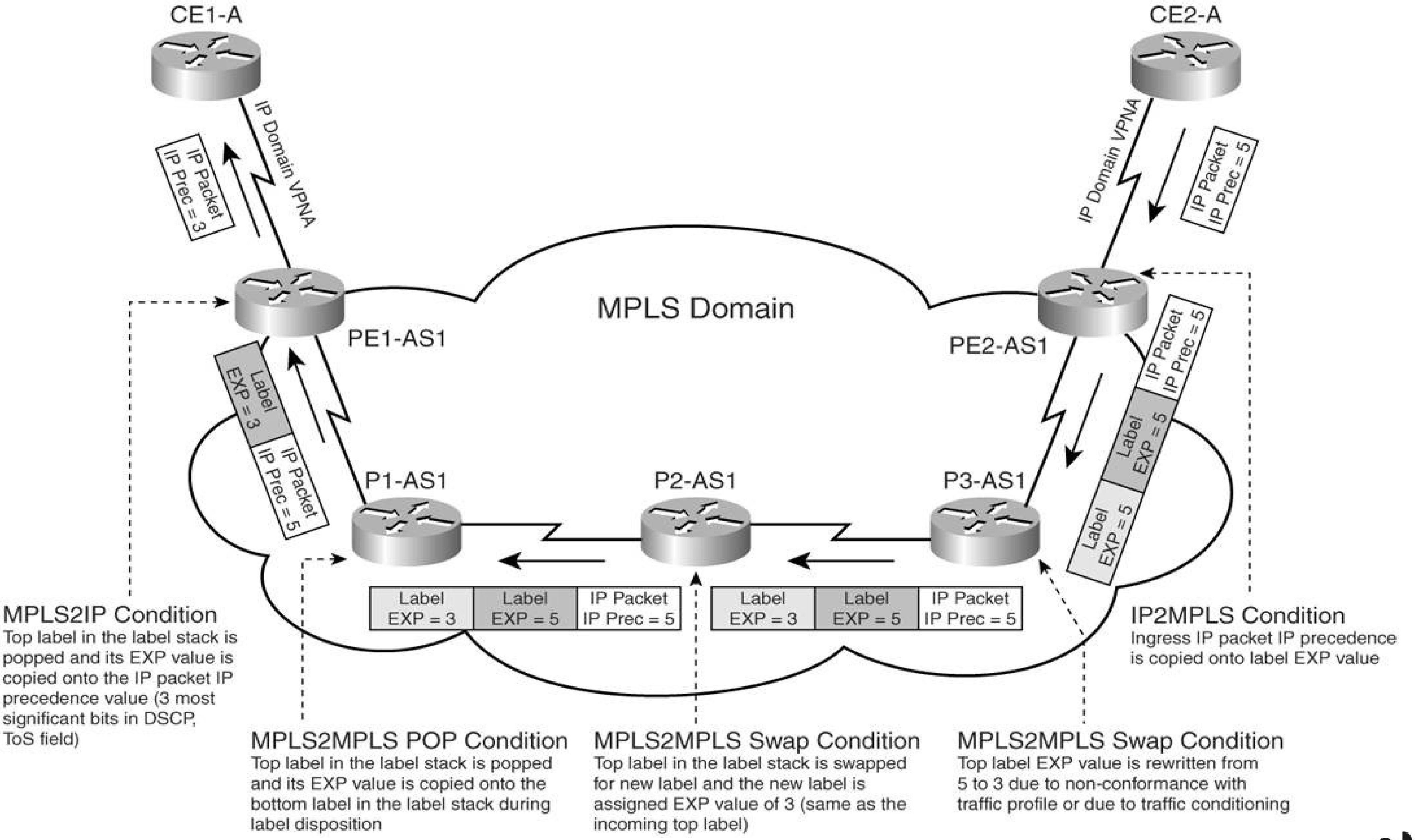
- Two different requirements for the transport of ATM across an MPLS backbone
  - Transport of AAL5 encapsulated frames (RFC1483);
  - Transport of ATM cells (cell relay)


Future QoS Mapping: L2→IP→EXP
FRR Protection of Tunnel
Any Transport over MPLS (AToM) Tunnel
MPLS Backbone
PE
PE
Virtual Leased Line (DS-TE + QoS)
DS-TE Tunnel
ATM
ATM
ATM Virtual Circuits
CPE Router
CPE Router

TE Tunnel Selection for AToM Attachment VCs

MPLS Tunnel Modes

Uniform Tunnel Mode



MPLS2IP Condition
Top label in the label stack is popped and its EXP value is copied onto the IP packet IP precedence value (3 most significant bits in DSCP, ToS field)

MPLS2MPLS POP Condition
Top label in the label stack is popped and its EXP value is copied onto the bottom label in the label stack during label disposition

MPLS2MPLS Swap Condition
Top label in the label stack is swapped for new label and the new label is assigned EXP value of 3 (same as the incoming top label)

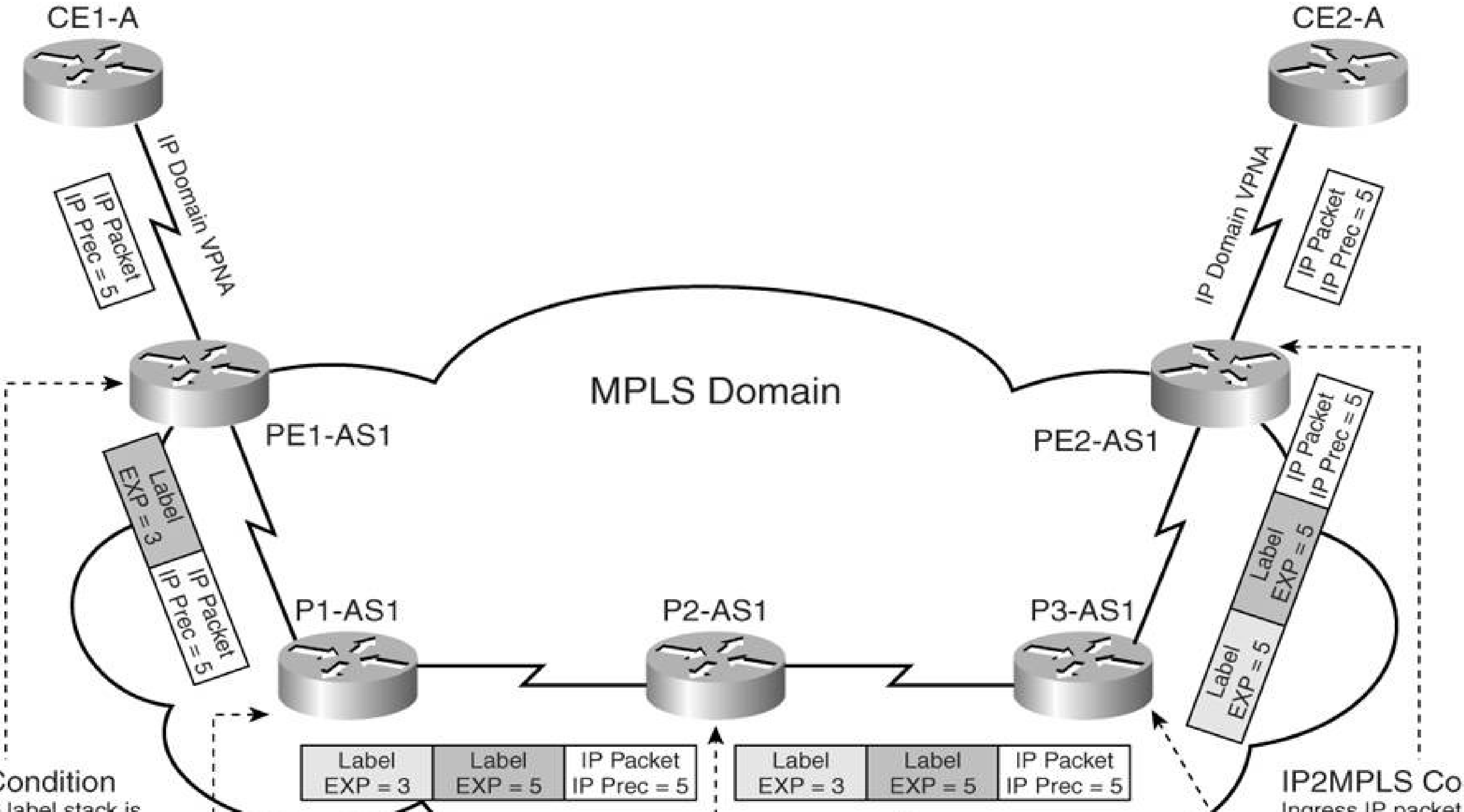
MPLS2MPLS Swap Condition
Top label EXP value is rewritten from 5 to 3 due to non-conformance with traffic profile or due to traffic conditioning

IP2MPLS Condition
Ingress IP packet IP precedence is copied onto label EXP value

Uniform Tunnel Mode implementation

- MPLS VPN service
- Uniform mode is used in a managed scenario where the SP controls QoS from CE to CE via the MPLS domain.
- An IP packet destined for CE1-A from CE2-A is given a label stack, the labels are marked with an EXP value of 5 mapping to the ingress IP packet's IP Precedence on PE2-AS1.
- P3-AS1 reassigns the top label's EXP value from 5 to 3 during the label swapping process.
- P2-AS1 performs a simple MPLS2MPLS swap function and forwards the labeled packet to P1-AS1 while preserving the EXP value at 3.
- P1-AS1 removes the top label in the label stack (penultimate hop popping). During this process, the top label's EXP value is copied onto the bottom label.
- PE1-AS1 receives the labeled packet and rewrites the outgoing IP packets IP precedence to 3 to map to the ingress labeled packet's EXP value.
- In this mode of operation, the PEs and CEs function as a single differential services domain as the QoS associated with a packet is carried across the MPLS domain as well as the remote CE's IP domain.

Pipe Tunnel Mode



MPLS2IP Condition
Top label in the label stack is popped and its EXP value is copied onto the IP packet IP precedence value (3 most significant bits in DSCP, ToS field)

IP2MPLS Condition
Ingress IP packet IP precedence is **NOT** copied onto label EXP value

MPLS2MPLS POP Condition
Top label in the label stack is popped and its EXP value is copied onto the bottom label in the label stack during label disposition

MPLS2MPLS Swap Condition
Top label in the label stack is swapped for new label and the new label is assigned EXP value of 3 (same as the incoming top label)

MPLS2MPLS Swap Condition
Top label EXP value is rewritten from 5 to 3 due to non-conformance with traffic profile or due to traffic conditioning

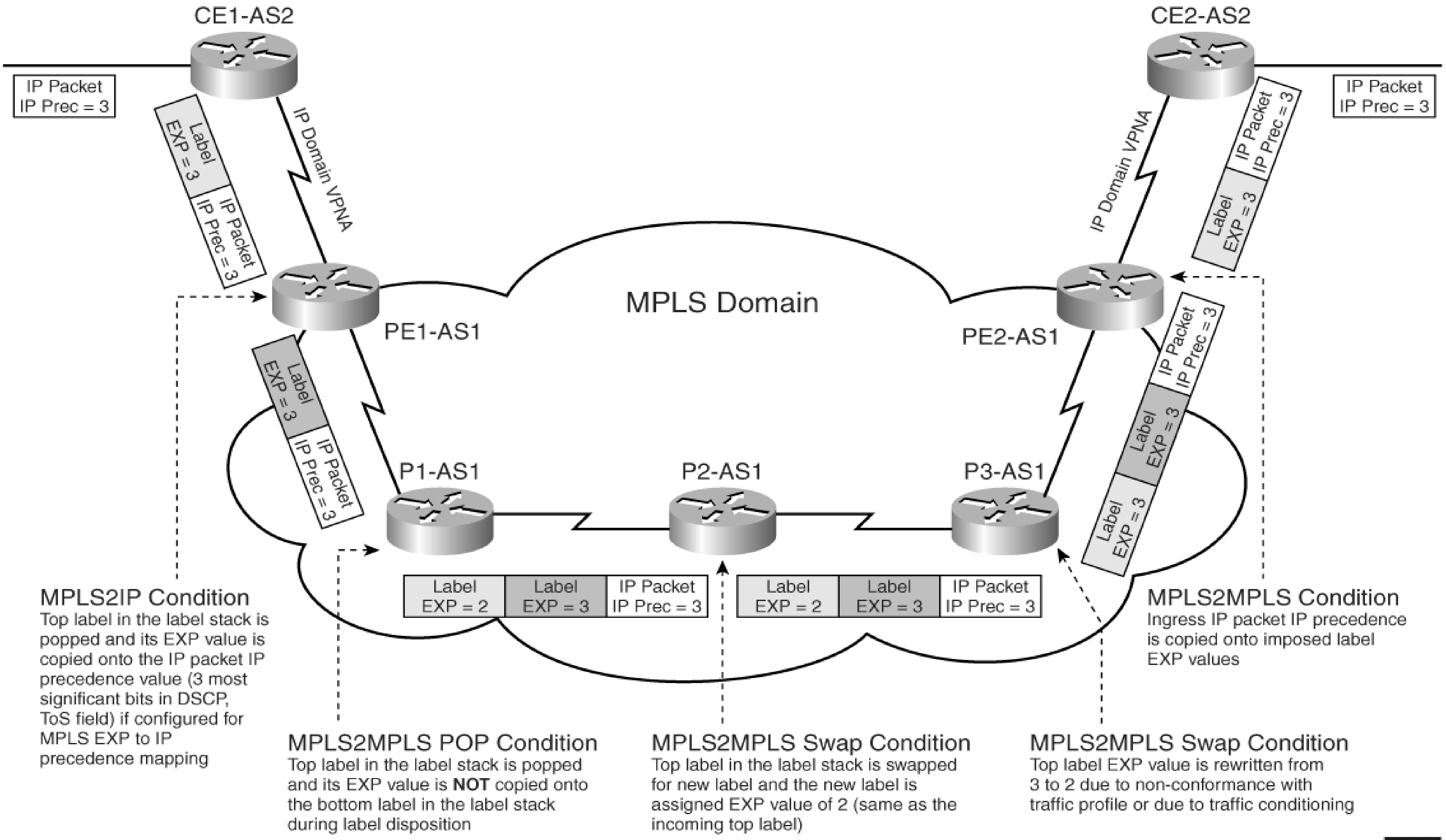
Pipe Tunnel Mode

- PE1-AS1 does not copy the ingress label EXP value onto the egress IP packet's IP Precedence value.
- However, the queuing characteristics of the labeled packet on PE1-AS1 still depend on the ingress label EXP value that is copied onto the qos-group value.
- This implementation is used when the SP would like to implement the PHB based on the QoS policy implementation in the SP core versus the customer's QoS policy when forwarding data to the attached CE routers.
- Hence, the QoS PHB of the same packet in the IP and the MPLS domain are independent of one another.

Short Pipe Tunnel Mode

- In Short Pipe mode, the difference occurs on egress from the MPLS to the IP domain (MPLS2IP condition).
- The packet's PHB is not associated to the ingress labeled packet's EXP value but only on the underlying IP packet's IP Precedence/DSCP value.
- The egress LSR does not maintain a copy of the ingress labeled packet's EXP value in the qos-group variable, which can be used to identify the egress PHB of the IP packet.
- This procedure is implemented when the QoS associated with the packet needs to conform to the customer's QoS policy.

Long Pipe Tunnel Mode



MPLS2IP Condition
Top label in the label stack is popped and its EXP value is copied onto the IP packet IP precedence value (3 most significant bits in DSCP, ToS field) if configured for MPLS EXP to IP precedence mapping

MPLS2MPLS POP Condition
Top label in the label stack is popped and its EXP value is **NOT** copied onto the bottom label in the label stack during label disposition

MPLS2MPLS Swap Condition
Top label in the label stack is swapped for new label and the new label is assigned EXP value of 2 (same as the incoming top label)

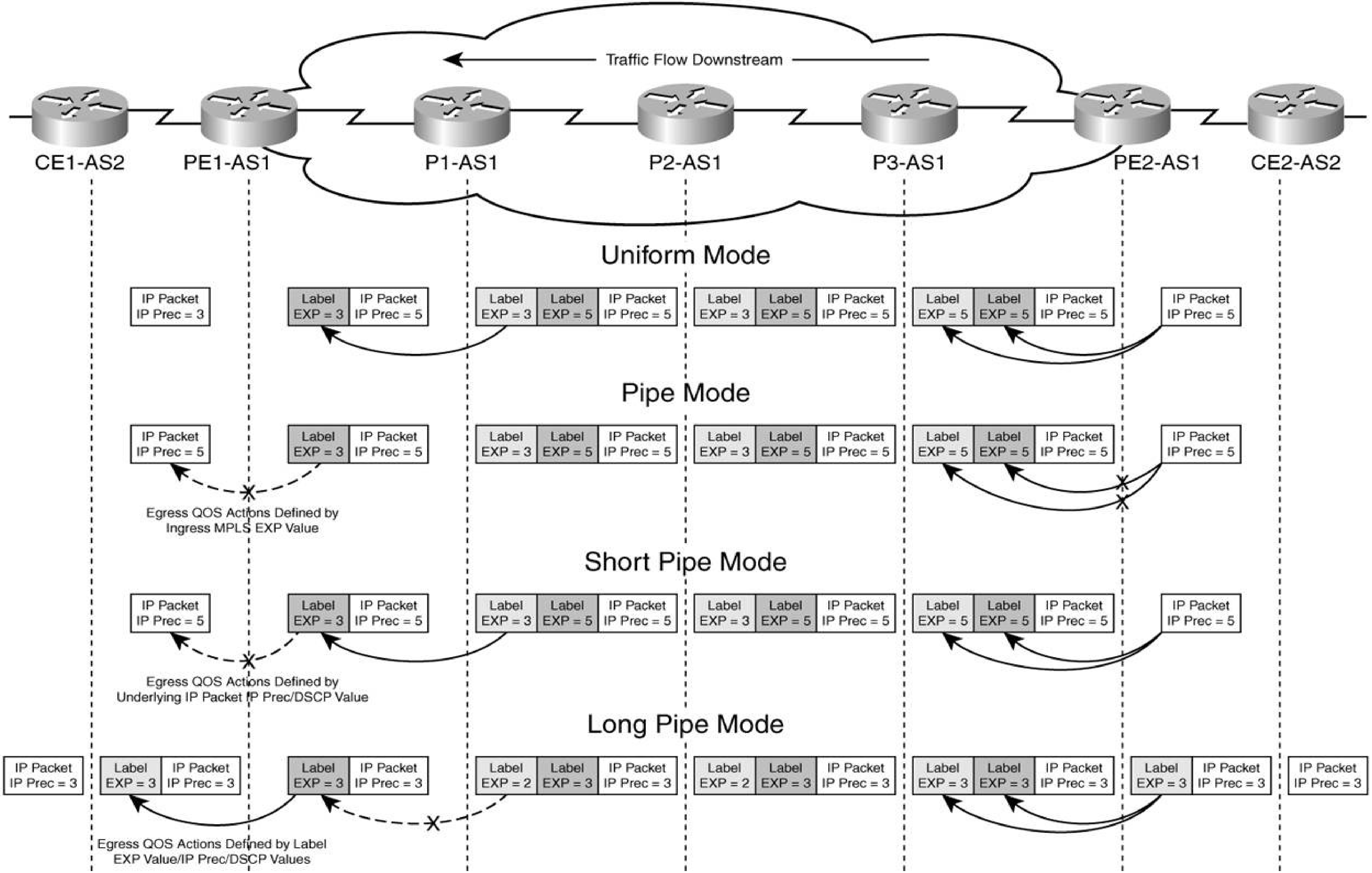
MPLS2MPLS Swap Condition
Top label EXP value is rewritten from 3 to 2 due to non-conformance with traffic profile or due to traffic conditioning

MPLS2MPLS Condition
Ingress IP packet IP precedence is copied onto imposed label EXP values

Long Pipe Tunnel Mode

- When a labeled packet is received by CE2-AS2 destined for CE1-AS2, the label is associated with the destination, and the label EXP value is copied as the ingress IP packet's IP Precedence value.
- When PE2-AS1 receives the ingress labeled packet, the label stack is applied with EXP value equal to the ingress label's EXP value.
- Although P3-AS1 rewrites the top label's EXP value to 2 (from 3) upon label disposition at P1-AS1, this value is not copied back down the label stack.
- PE1-AS1 performs the MPLS2MPLS label swapping function with direct mapping of EXP bits.
- On receiving the labeled packet on CE1-AS2, the router can perform PHB based on the ingress labeled packets EXP value or underlying IP packet's IP Precedence value.

Summary of MPLS QoS Modes



Traffic Flow Downstream

CE1-AS2 PE1-AS1 P1-AS1 P2-AS1 P3-AS1 PE2-AS1 CE2-AS2

Uniform Mode

IP Packet IP Prec = 3 | Label EXP = 3, IP Packet IP Prec = 5 | Label EXP = 3, Label EXP = 5, IP Packet IP Prec = 5 | Label EXP = 3, Label EXP = 5, IP Packet IP Prec = 5 | Label EXP = 5, Label EXP = 5, IP Packet IP Prec = 5 | IP Packet IP Prec = 5

Pipe Mode

IP Packet IP Prec = 5 | Label EXP = 3, IP Packet IP Prec = 5 | Label EXP = 3, Label EXP = 5, IP Packet IP Prec = 5 | Label EXP = 3, Label EXP = 5, IP Packet IP Prec = 5 | Label EXP = 5, Label EXP = 5, IP Packet IP Prec = 5 | IP Packet IP Prec = 5

Egress QOS Actions Defined by
Ingress MPLS EXP Value

Short Pipe Mode

IP Packet IP Prec = 5 | Label EXP = 3, IP Packet IP Prec = 5 | Label EXP = 3, Label EXP = 5, IP Packet IP Prec = 5 | Label EXP = 3, Label EXP = 5, IP Packet IP Prec = 5 | Label EXP = 5, Label EXP = 5, IP Packet IP Prec = 5 | IP Packet IP Prec = 5

Egress QOS Actions Defined by
Underlying IP Packet IP Prec/DSCP Value

Long Pipe Mode

IP Packet IP Prec = 3 | Label EXP = 3, IP Packet IP Prec = 3 | Label EXP = 3, IP Packet IP Prec = 3 | Label EXP = 2, Label EXP = 3, IP Packet IP Prec = 3 | Label EXP = 2, Label EXP = 3, IP Packet IP Prec = 3 | Label EXP = 3, Label EXP = 3, IP Packet IP Prec = 3 | Label EXP = 3, IP Packet IP Prec = 3 | IP Packet IP Prec = 3

Egress QOS Actions Defined by Label
EXP Value/IP Prec/DSCP Values

Terminology

Terminology, 1/2

- RR—Route Reflector
    - A router (usually not involved in packet forwarding) that distributes BGP routes within a provider's network
- PE—Provider Edge router
    - The interface between the customer and the MPLS-VPN network; only PEs (and maybe RRs) know anything about MPLS-VPN routes
- P—Provider router
    - A router in the core of the MPLS-VPN network, speaks LDP/RSVP but not VPNv4
- CE—Customer Edge router
    - The customer router which connects to the PE; does not know anything about labels, only IP (most of the time)
- LDP—Label Distribution Protocol
    - Distributes labels with a provider's network that mirror the IGP, one way to get from one PE to another
- LSP—Label Switched Path
    - The chain of labels that are swapped at each hop to get from one PE to another

Terminology, 2/2

- VPN—Virtual Private Network
  - A network deployed on top of another network, where the two networks are separate and never communicate
- VRF—Virtual Routing and Forwarding instance
  - Mechanism in IOS used to build per-interface RIB and FIB
- VPNv4
  - Address family used in BGP to carry MPLS-VPN routes
- RD
  - Route Distinguisher, used to uniquely identify the same network/mask from different VRFs (i.e., 10.0.0.0/8 from VPN A and 10.0.0.0/8 from VPN B)
- RT
  - Route Target, used to control import and export policies, to build arbitrary VPN topologies for customers