Practical Deployment Guidelines for MPLS-VPN Networks

Azhar Sayeed and Monique Morrow

asayeed@cisco.com, mmorrow@cisco.com

Prerequisites

Cisco.com

- **Must understand fundamental MPLS principles**
- **Must understand basic routing especially BGP**

Introduction to MPLS

Azhar Sayeed

Agenda

- Background
- Technology Basics
    - What is MPLS? Where Is it Used?
- Label Distribution in MPLS Networks
    - LDP, RSVP, BGP
- Building MPLS Based Services
    - VPNs
    - AToM
    - Traffic Engineering
- Configurations
    - Configuring MPLS, LDP, TE
- Summary

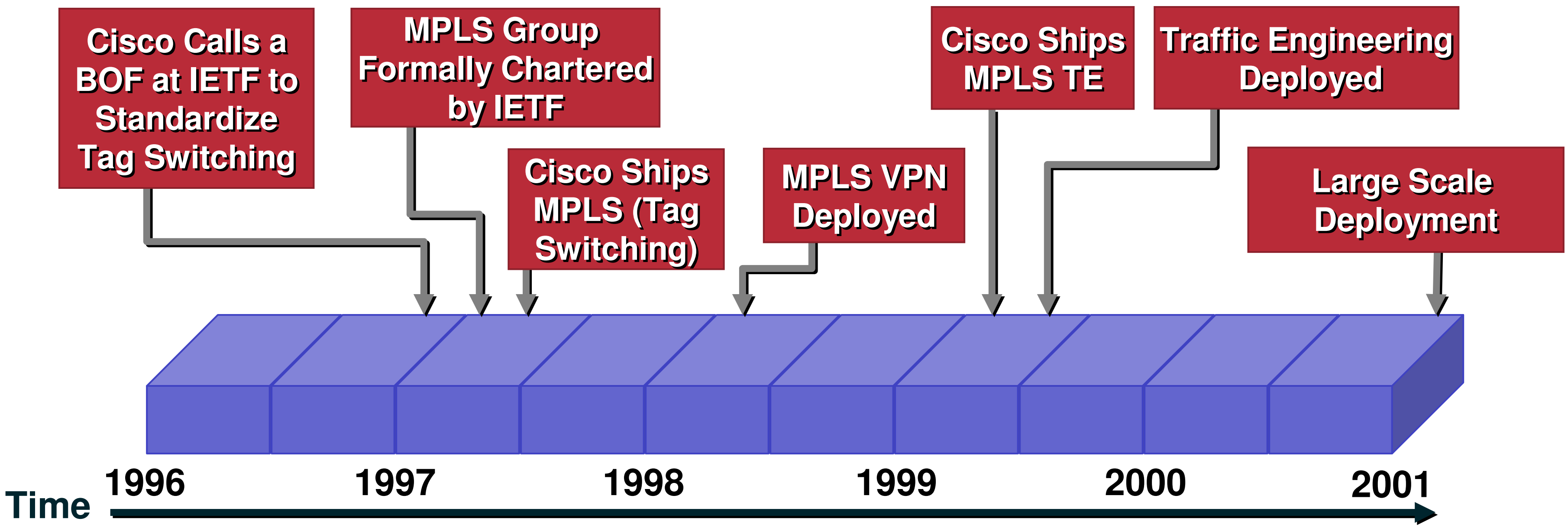
Background

Terminology

- ## Acronyms
  - PE—provider edge router
  - P—Provider core router
  - CE—Customer Edge router (also referred to as CPE)
  - ASBR—Autonomous System Boundary Router
  - RR—Route Reflector
  - 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
- ## TE—Traffic Engineering
  - TE Head end—Router that initiates a TE tunnel
  - TE Midpoint—Router where the TE Tunnel transits
- ## VPN—Collection of sites that share common policies
  - 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
- ## AToM—Any Transport over MPLS
  - Commonly known scheme for building layer 2 circuits over MPLS
  - Attachment Circuit—Layer 2 circuit between PE and CE
  - Emulated circuit—Pseudowire between PEs

Evolution of MPLS

- From Tag Switching
- Proposed in IETF—Later combined with other proposals from IBM (ARIS), Toshiba (CSR)



What Is MPLS?

- Multi Protocol Label Switching
- MPLS is an efficient encapsulation mechanism
- Uses "Labels" appended to packets (IP packets, AAL5 frames) for transport of data
- MPLS packets can run on other layer 2 technologies such as ATM, FR, PPP, POS, Ethernet
- Other layer 2 technologies can be run over an MPLS network
- Labels can be used as designators

  For example—IP prefixes, ATM VC, or a bandwidth guaranteed path
- MPLS is a technology for delivery of IP Services

Original Motivation of MPLS

Cisco.com

- Allow Core routers/networking devices to switch packets based some simplified header
- Provide a highly scalable mechanism that was topology driven rather than flow driven
- Leverage hardware so that simple forwarding paradigm can be used
- It has evolved a long way from the original goal

  Hardware became better and looking up longest best match was no longer an issue

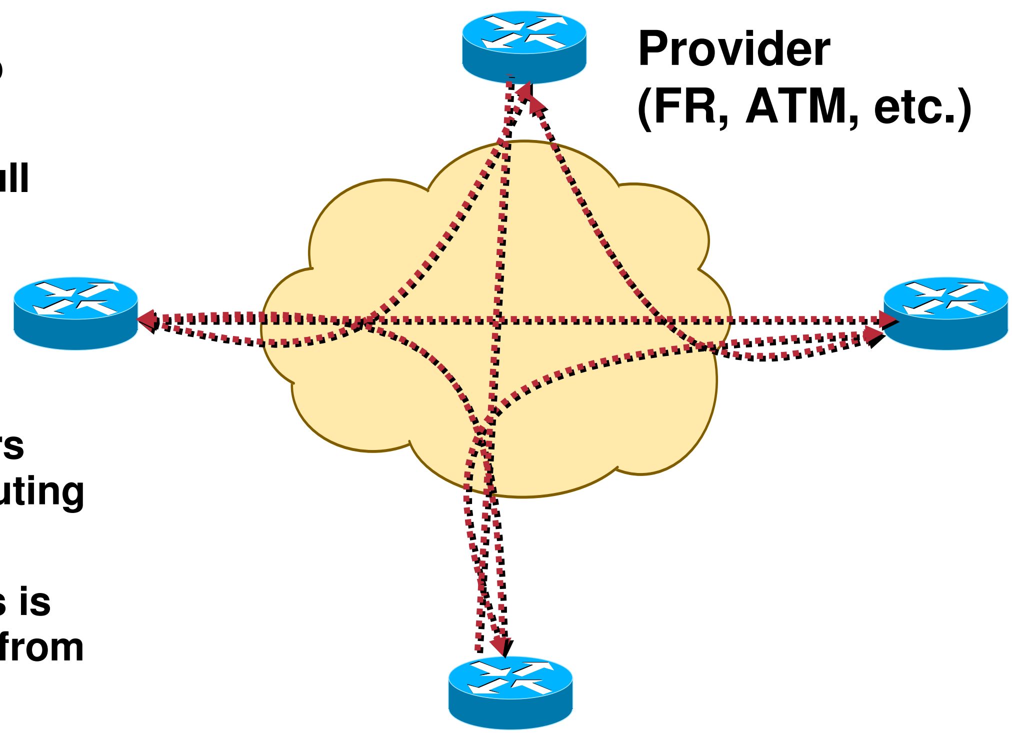
  By associating Labels with prefixes, groups of sites or bandwidth paths or light paths new services such as MPLS VPNs and Traffic engineering, GMPLS were now possible

Overlay vs. Peer Networks

- **Overlay network: customer's IP network is overlaid on top of the provider's network**
    - Provider's IP transport (FR, ATM, etc.) creates private IP network for customer
    - Most technologies that carry IP are p2p
    - Large p2p networks are hard to maintain
        - N^2 provisioning vs. inefficient routing
        - Even with hub and spoke, need lots of stuff at the hub

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



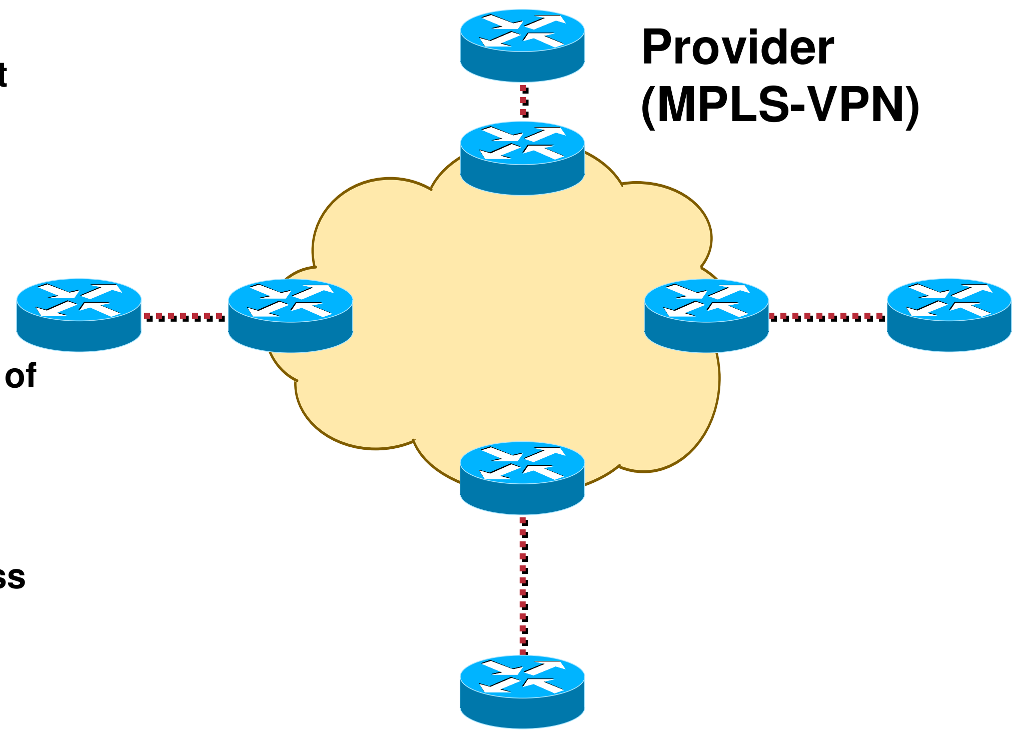
Peer Network

Cisco.com

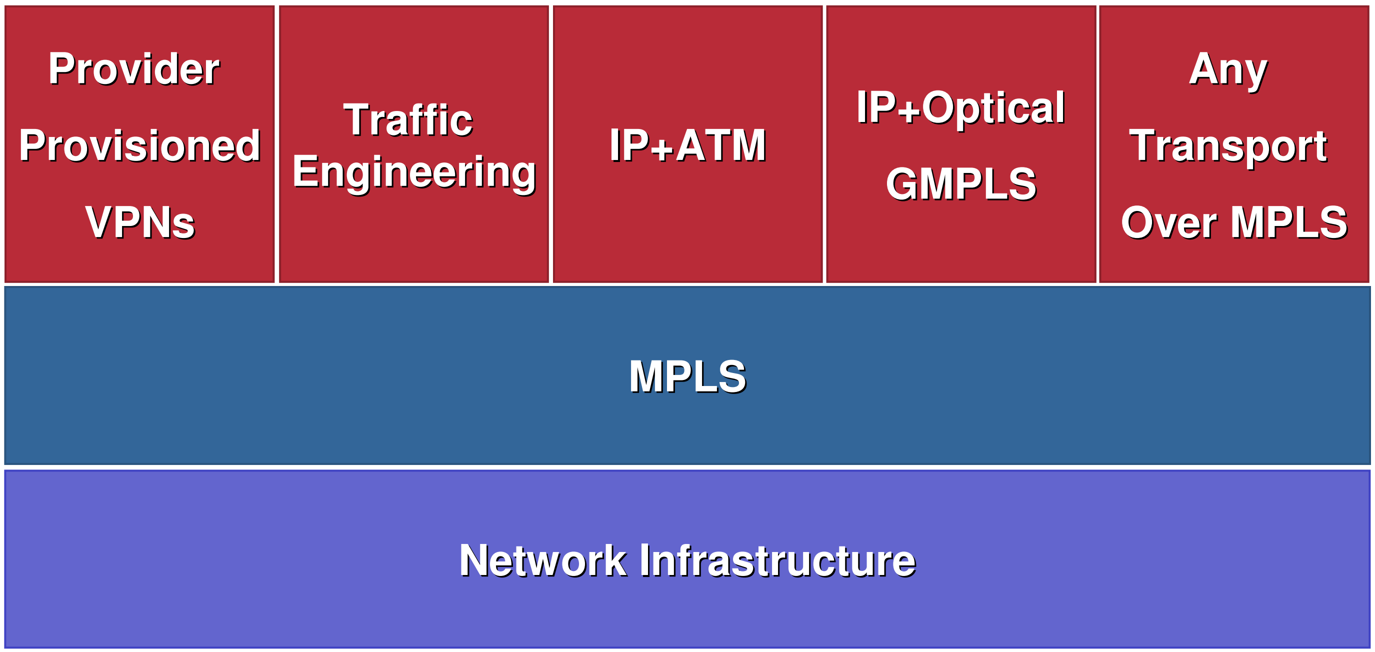
- **Provider and customer exchange IP routing information directly**
  - **Customer only has one routing peer per site**
- **Need to separate customer's IP network from provider's network**
  - **Customer A and Customer B need to not talk to each other**
  - **Customer A and Customer B may have the same address space (10.0.0.0/8, 161.44.0.0/16, etc.)**
- **VPN is provisioned and run by the provider**
- **MPLS-VPN does this without p2p connections**

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



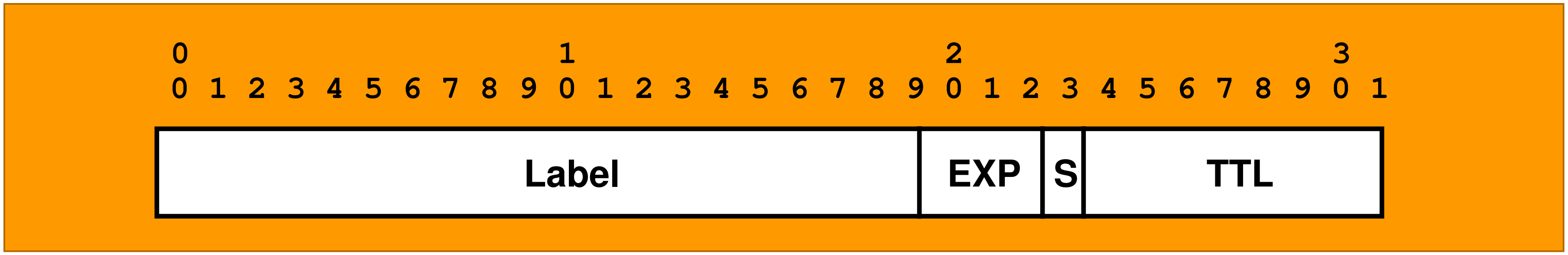
MPLS as a Foundation for Value Added Services



Technology Basics

Azhar Sayeed

Label Header for Packet Media



Label = 20 bits
COS/EXP = Class of Service, 3 bits
S = Bottom of Stack, 1 bit
TTL = Time to Live, 8 bits

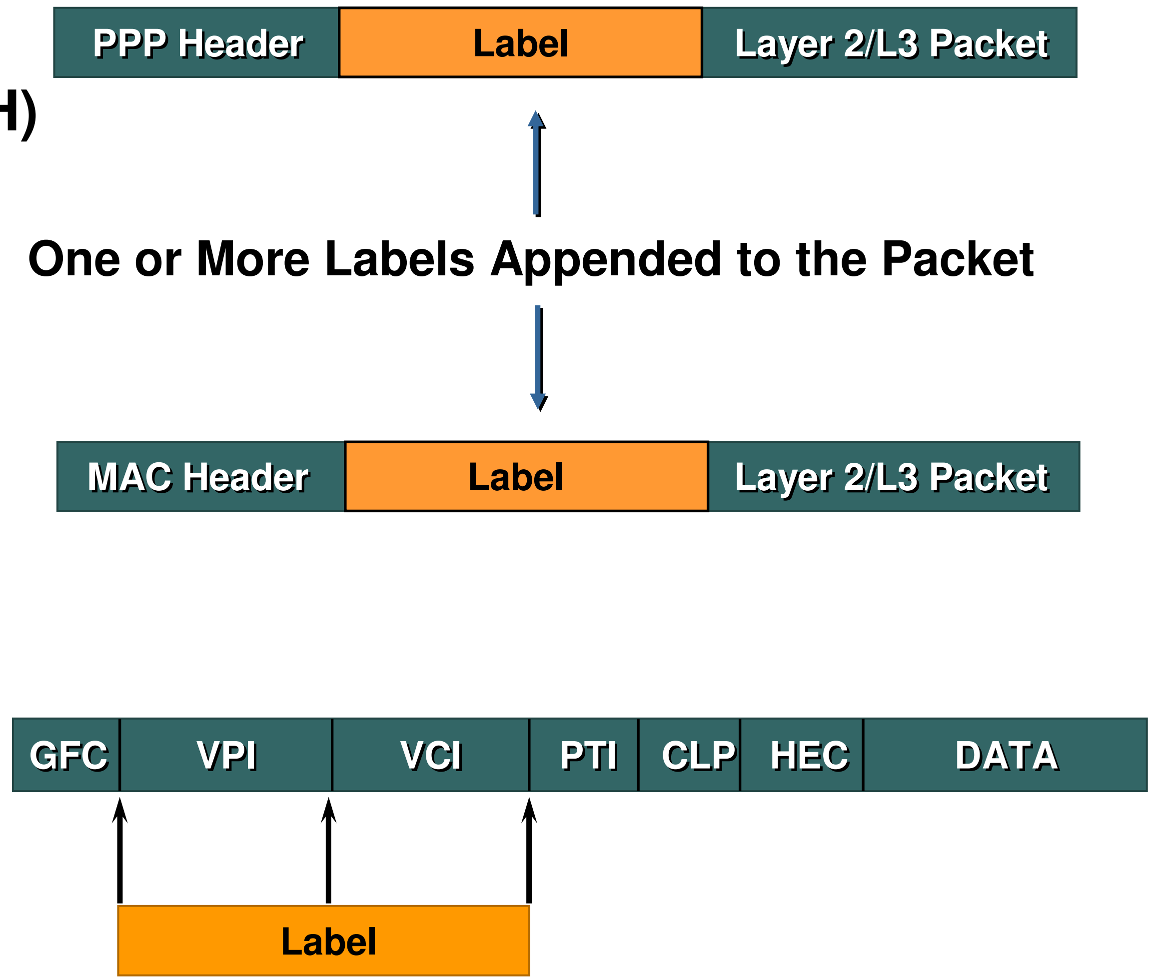
- **Can be used over Ethernet, 802.3, or PPP links**
- **Uses two new Ethertypes/PPP PIDs**
- **Contains everything needed at forwarding time**
- **One word per label**

Encapsulations

PPP Header (Packet over SONET/SDH)

LAN MAC Label Header

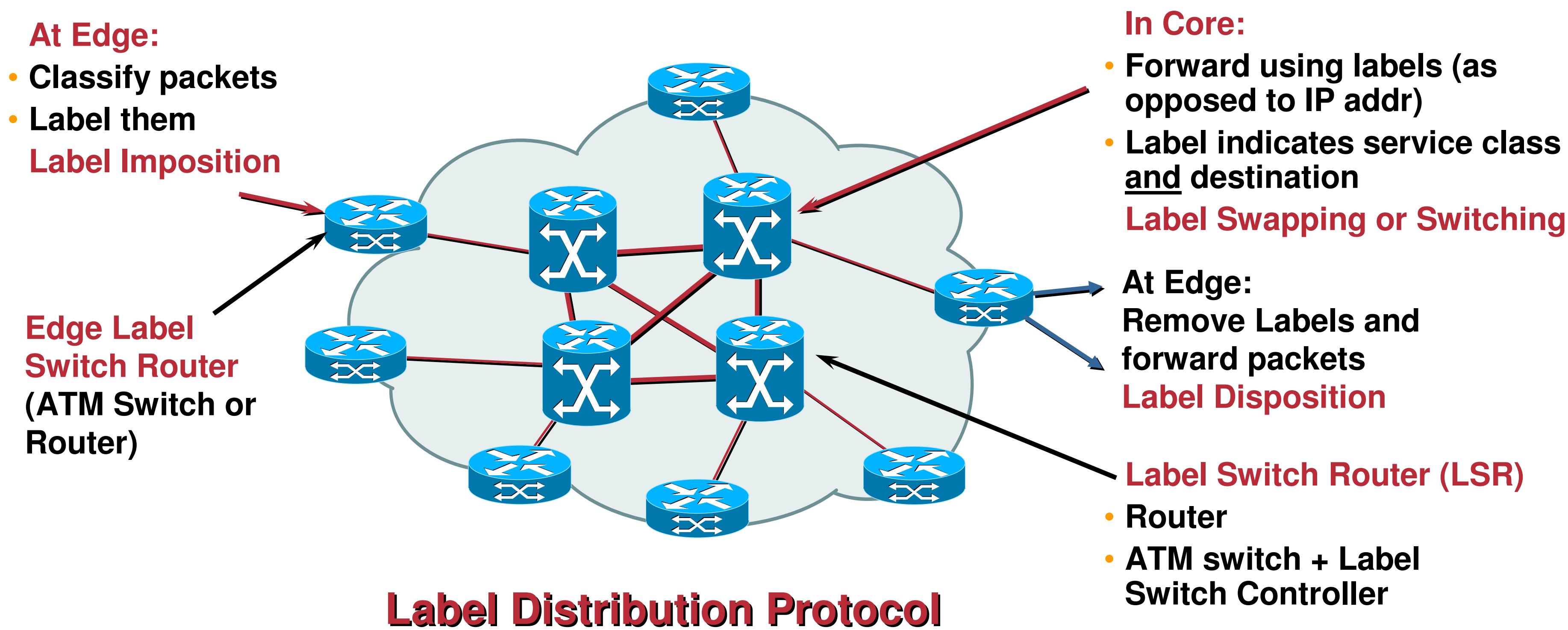
ATM MPLS Cell Header



Forwarding Equivalence Class

- **Determines how packets are mapped to LSP**
    - IP Prefix/host address
    - Layer 2 Circuits (ATM, FR, PPP, HDLC, Ethernet)
    - Groups of addresses/sites—VPN x
    - A Bridge/switch instance—VSI
    - Tunnel interface—Traffic Engineering

MPLS Concepts



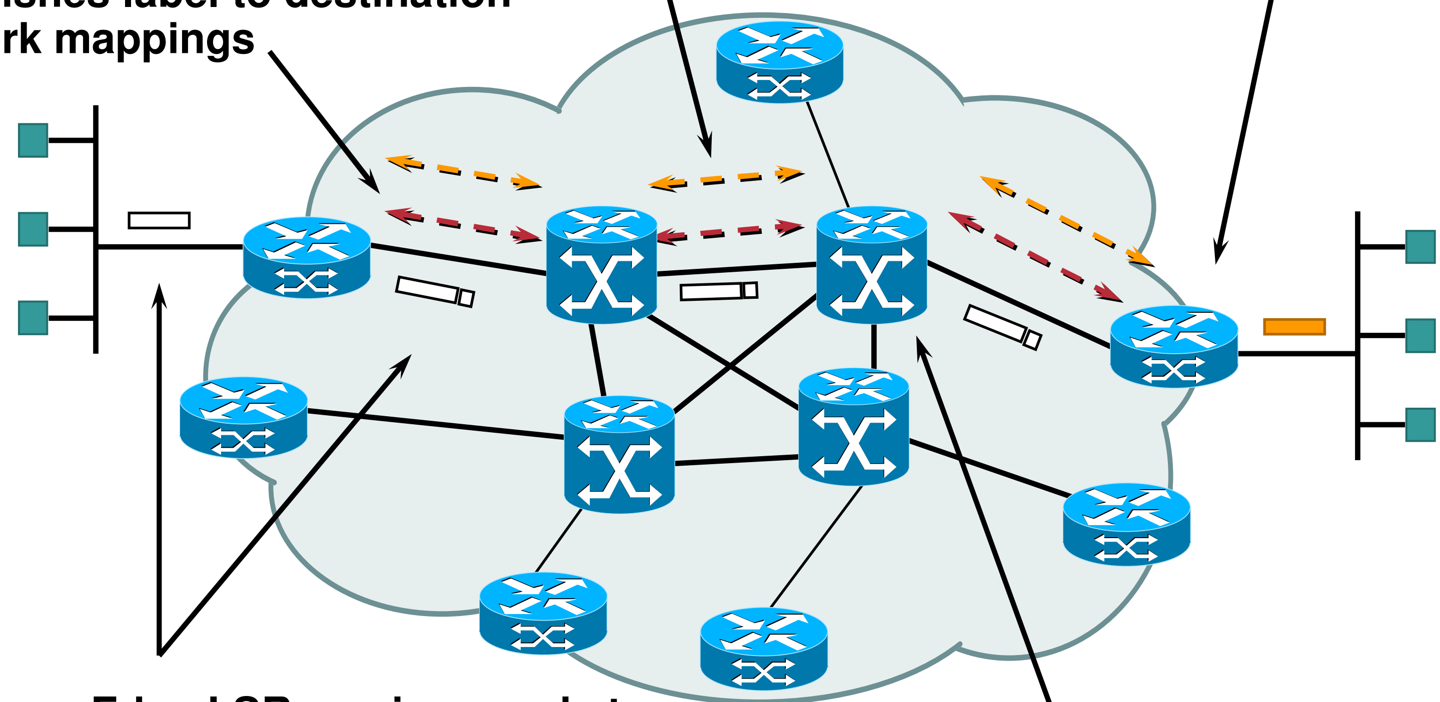
- Create new services via flexible classification
- Provides the ability to setup bandwidth guaranteed paths
- Enable ATM switches to act as routers

MPLS Operation

1a. Existing routing protocols (e.g. OSPF, IS-IS) establish reachability to destination networks

1b. Label Distribution Protocol (LDP) establishes label to destination network mappings

4. Edge LSR at egress removes label and delivers packet



2. Ingress Edge LSR receives packet, performs Layer 3 value-added services, and “labels” packets

3. LSR switches packets using label swapping

Label Distribution in MPLS Networks

Azhar Sayeed

Unicast Routing Protocols

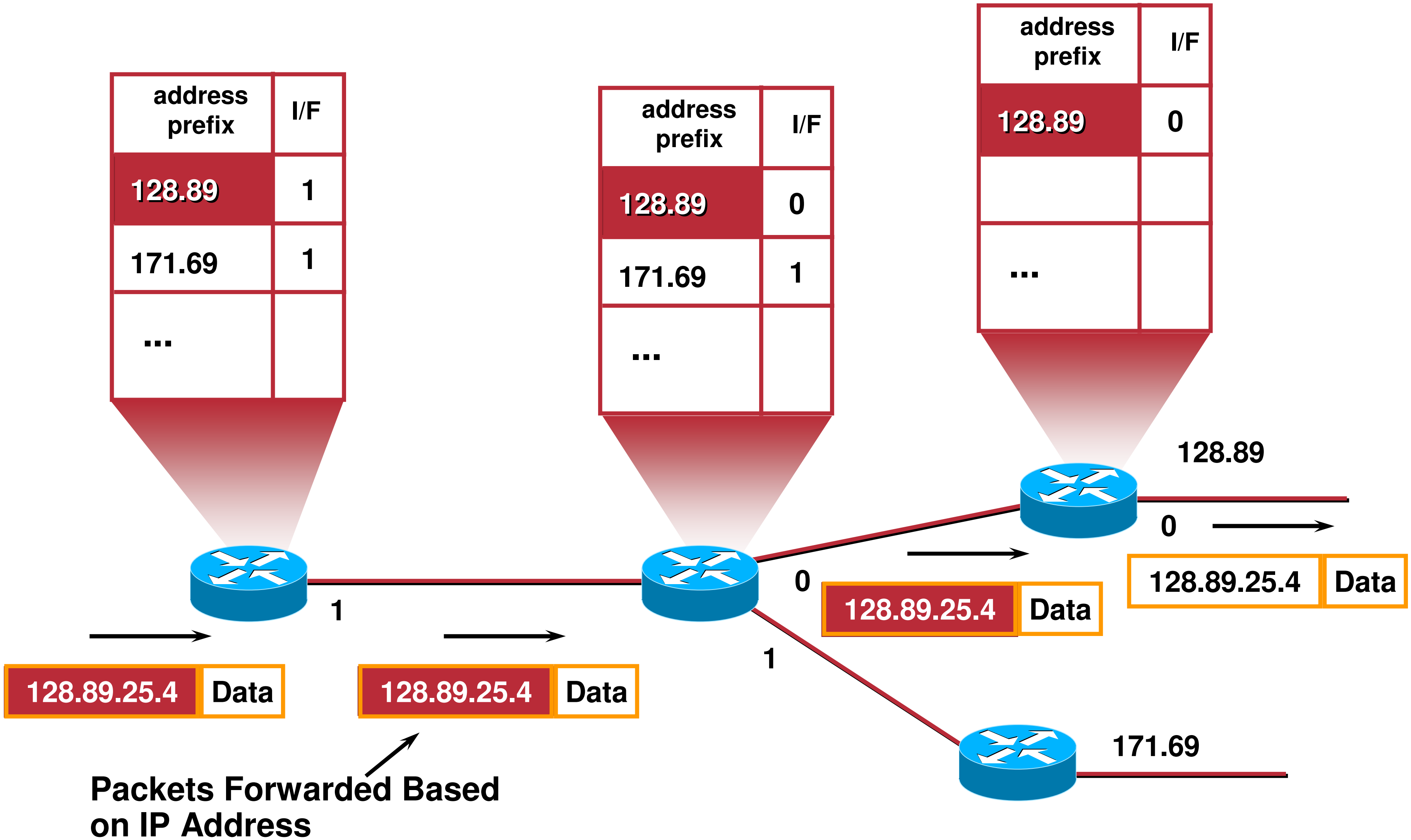
- OSPF, IS-IS, BGP are needed in the network
- They provide reachability
- Label distribution protocols distribute labels for prefixes advertised by unicast routing protocols using
  - Either a dedicated Label Distribution Protocol (LDP)
  - Extending existing protocols like BGP to distribute Labels

Label Distribution Protocol

- Defined in RFC 3035 and 3036
- Used to distribute Labels in a MPLS network
- Forwarding Equivalence Class
  - How packets are mapped to LSPs (Label Switched Paths)
- Advertise Labels per FEC
  - Reach destination a.b.c.d with label x
- Discovery

Router Example: Forwarding Packets

Cisco.com



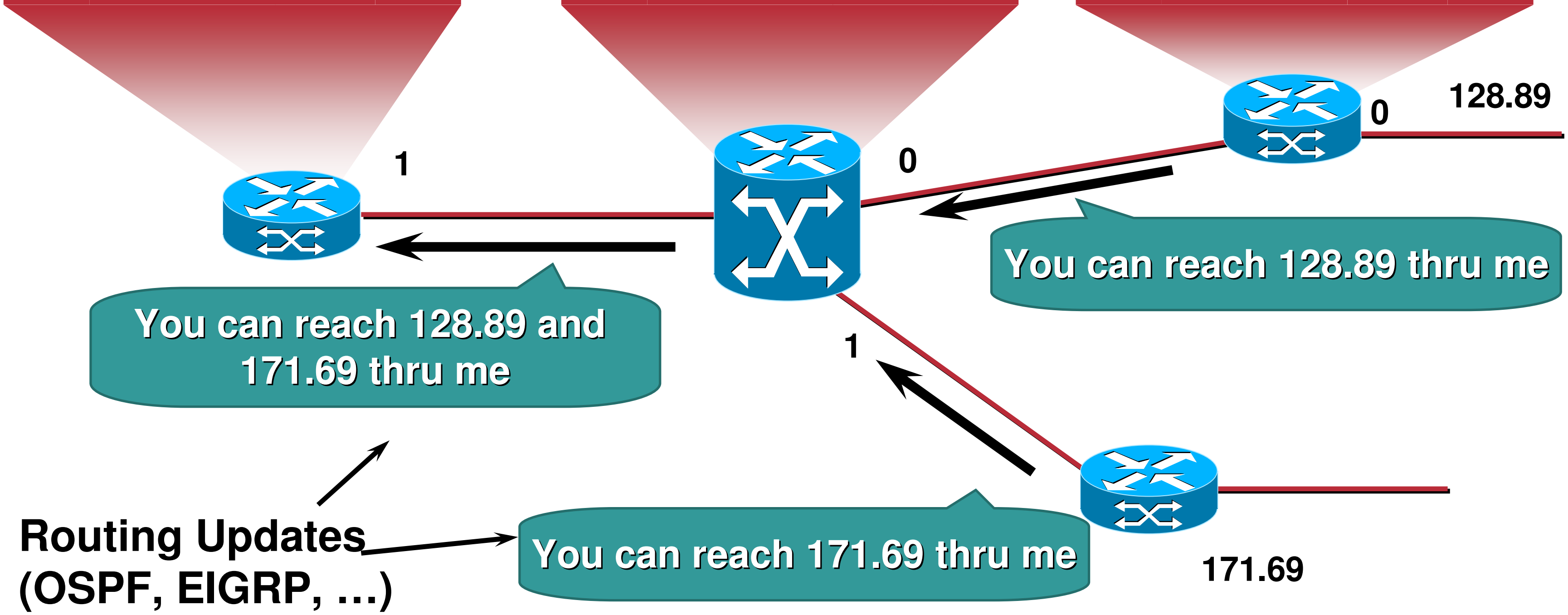
MPLS Example: Routing Information

Cisco.com

In label	Address Prefix	Out I'face	Out label
	128.89	1	
	171.69	1	
	...	...	

In label	Address Prefix	Out I'face	Out label
	128.89	0	
	171.69	1	
	...	...	

In label	Address Prefix	Out I'face	Out label
	128.89	0	
	...	...	

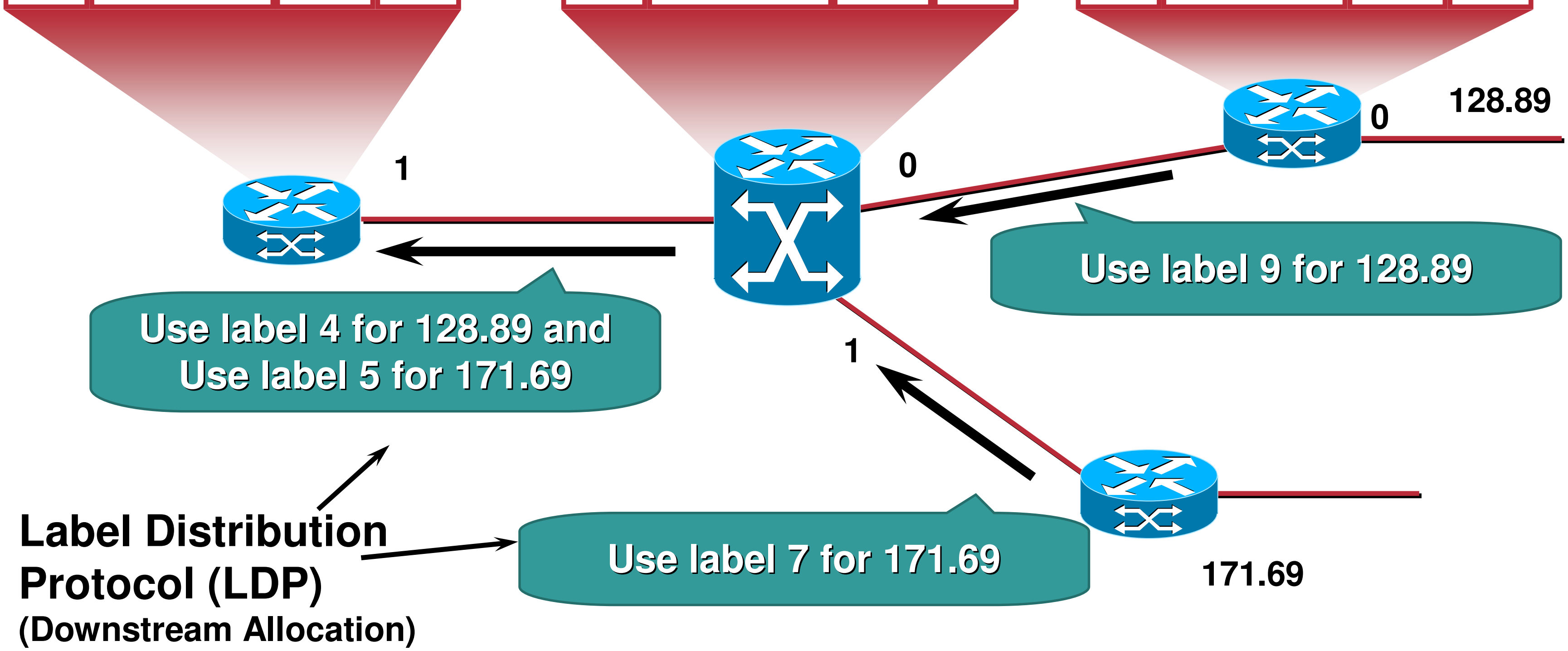


MPLS Example: Assigning Labels

In label	Address Prefix	Out I'face	Out label
-	128.89	1	4
-	171.69	1	5
...	...	...	...

In label	Address Prefix	Out I'face	Out label
4	128.89	0	9
5	171.69	1	7
...	...	...	...

In label	Address Prefix	Out I'face	Out label
9	128.89	0	-
...	...	...	...

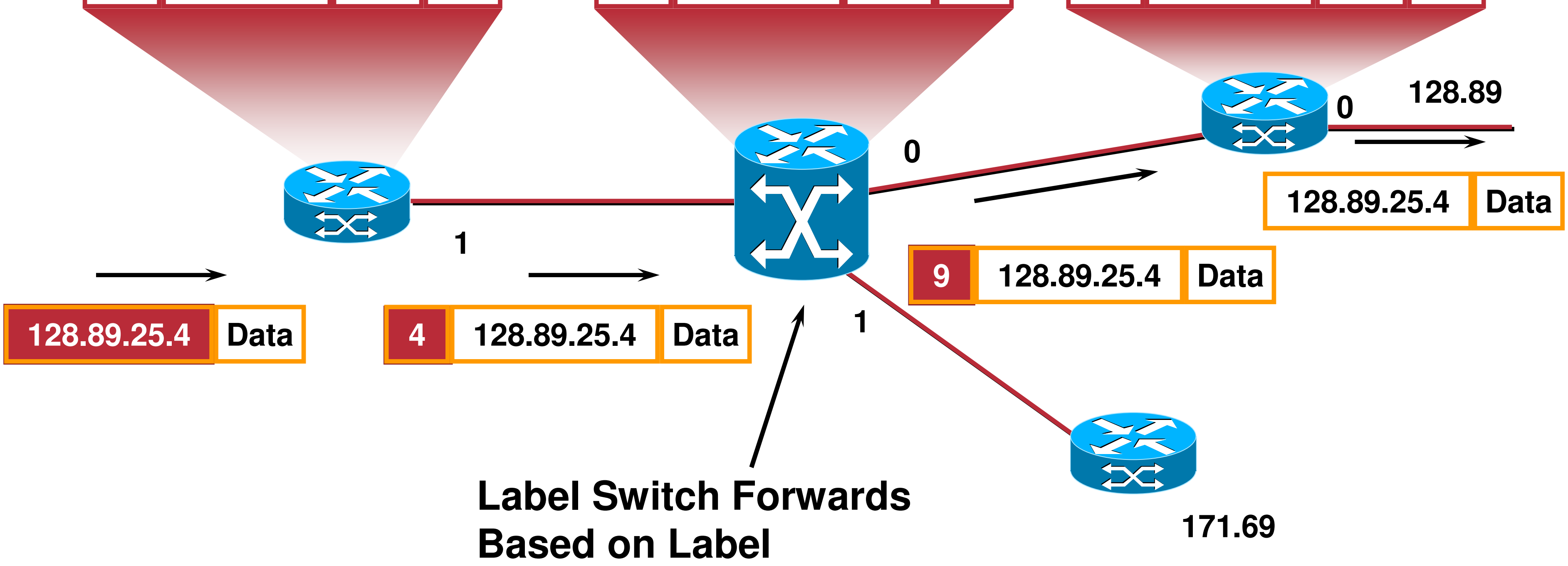


MPLS Example: Forwarding Packets

In label	Address Prefix	Out I'face	Out label
-	128.89	1	4
-	171.69	1	5
...	...	...	...

In label	Address Prefix	Out I'face	Out label
4	128.89	0	9
5	171.69	1	7
...	...	...	...

In label	Address Prefix	Out I'face	Out label
9	128.89	0	-
...	...	...	...



Label Distribution Modes

Cisco.com

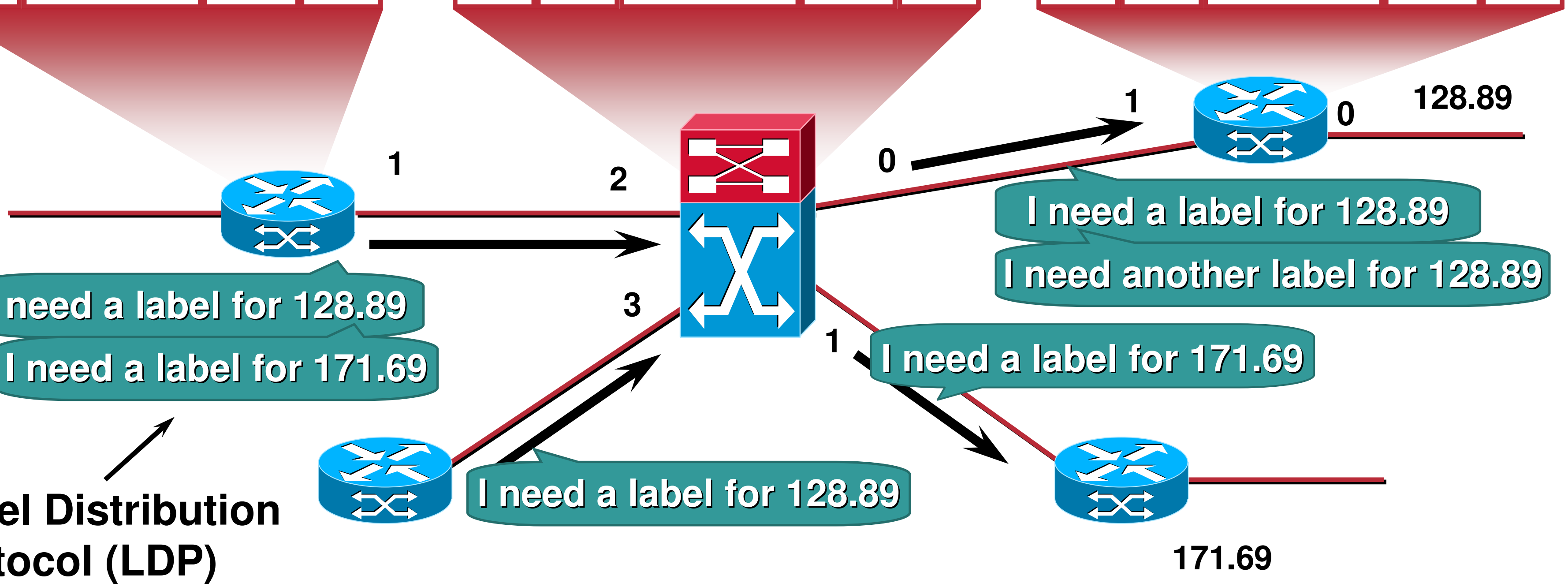
- **Downstream unsolicited**
  - Downstream node just advertises labels for prefixes/FEC reachable via that device
  - Previous example
- **Downstream on-demand**
  - Upstream node requests a label for a learnt prefix via the downstream node
  - Next example—ATM MPLS

ATM MPLS Example: Requesting Labels

In label	Address Prefix	Out I'face	Out label
	128.89	1	
	171.69	1	
	...	...	

In label	In I/F	Address Prefix	Out I'face	Out label
		128.89	0	
		171.69	1	
		...	...	

In label	In I/F	Address Prefix	Out I'face	Out label
		128.89	0	
		...	...	



Label Distribution Protocol (LDP)
(Downstream Allocation on Demand)

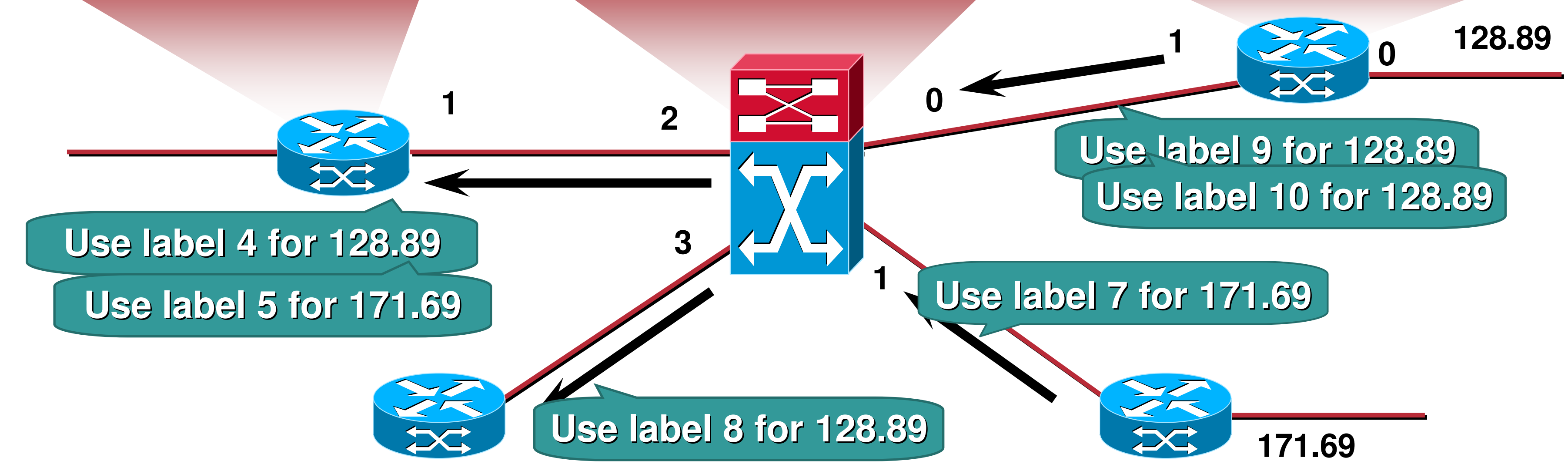
ATM MPLS Example: Assigning Labels

Cisco.com

In label	Address Prefix	Out I'face	Out label
-	128.89	1	4
-	171.69	1	5
	...	...	

In label	In I/F	Address Prefix	Out I'face	Out label
4	2	128.89	0	9
8	3	128.89	0	10
5	2	171.69	1	7

In label	In I/F	Address Prefix	Out I'face	Out label
9	1	128.89	0	-
10	1	128.89	0	-
		...	...	

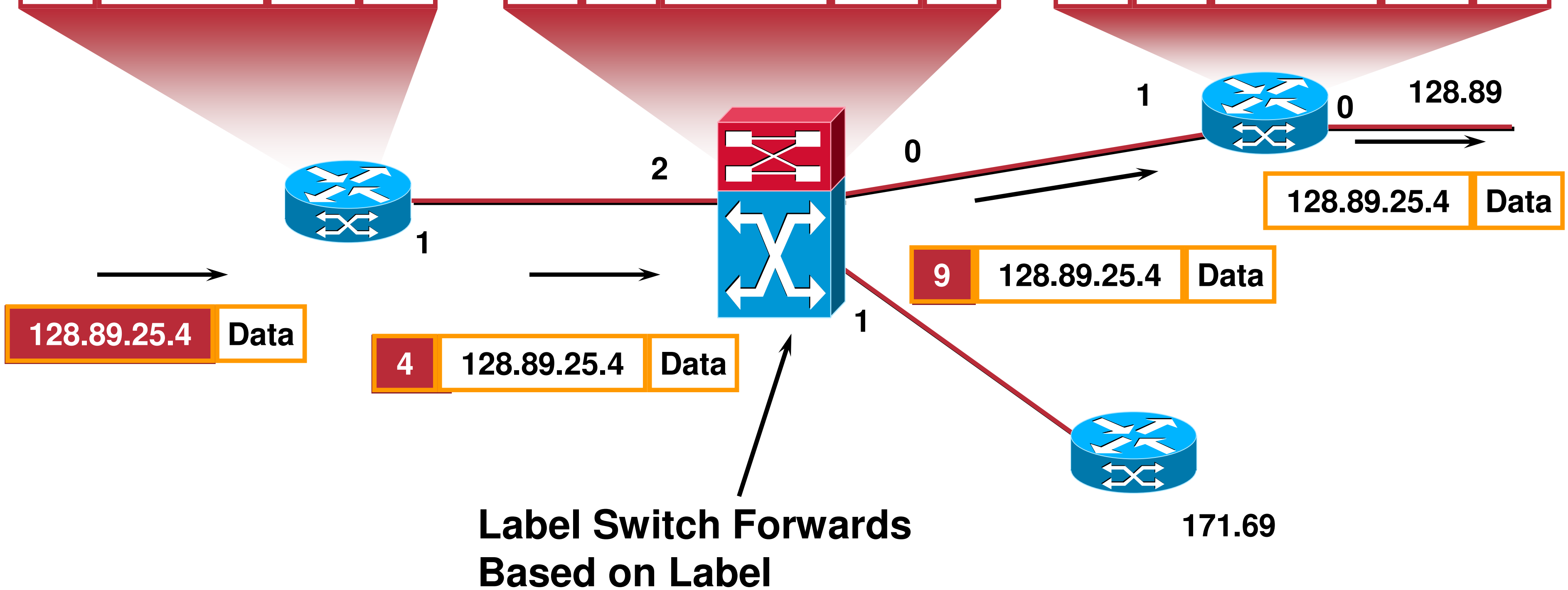


ATM MPLS Example: Packet Forwarding

In label	Address Prefix	Out I'face	Out label
-	128.89	1	4
-	171.69	1	5
	...	...	

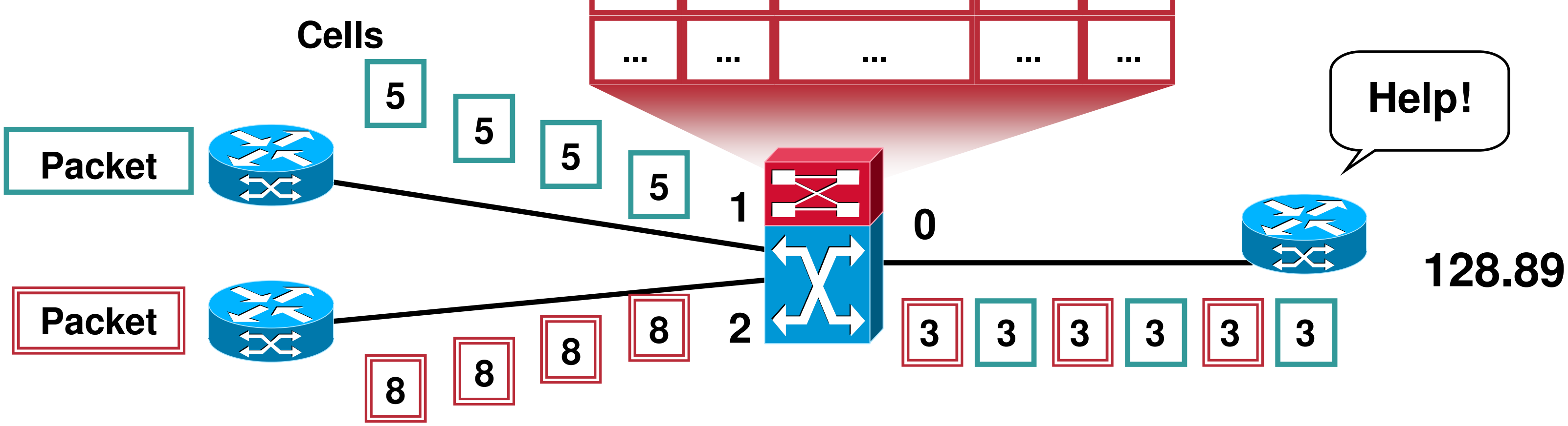
In label	In I/F	Address Prefix	Out I'face	Out label
4	2	128.89	0	9
8	3	128.89	0	10
5	2	171.69	1	7

In label	In I/F	Address Prefix	Out I'face	Out label
9	1	128.89	0	-
10	1	128.89	0	-
		...	...	



Why Multiple Labels with ATM?

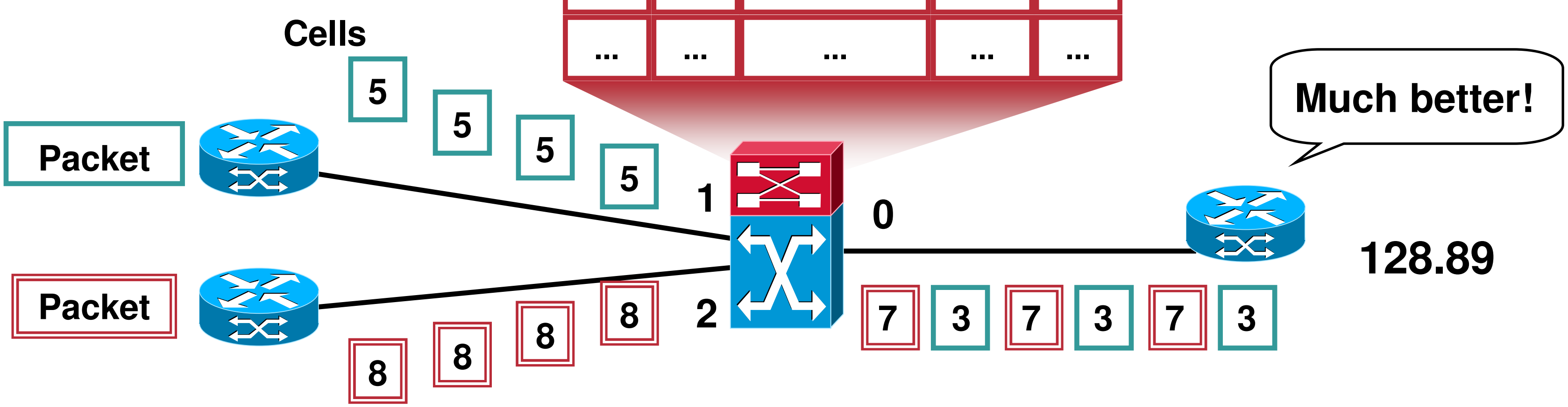
In I/F	In label	Address Prefix	Out I/F	Out label
1	5	128.89	0	3
2	8	128.89	0	3
...	...	...	...	...



- **If didn’t allocate multiple labels:**
  - **Cells of different packets would have same label (VPI/VCI)**
  - **Egress router can’t reassemble packets**

Multiple Labels

In I/F	In label	Address Prefix	Out I/F	Out label
1	5	128.89	0	3
2	8	128.89	0	7
...	...	...	...	...



- Multiple labels enables edge router to reassemble packets correctly

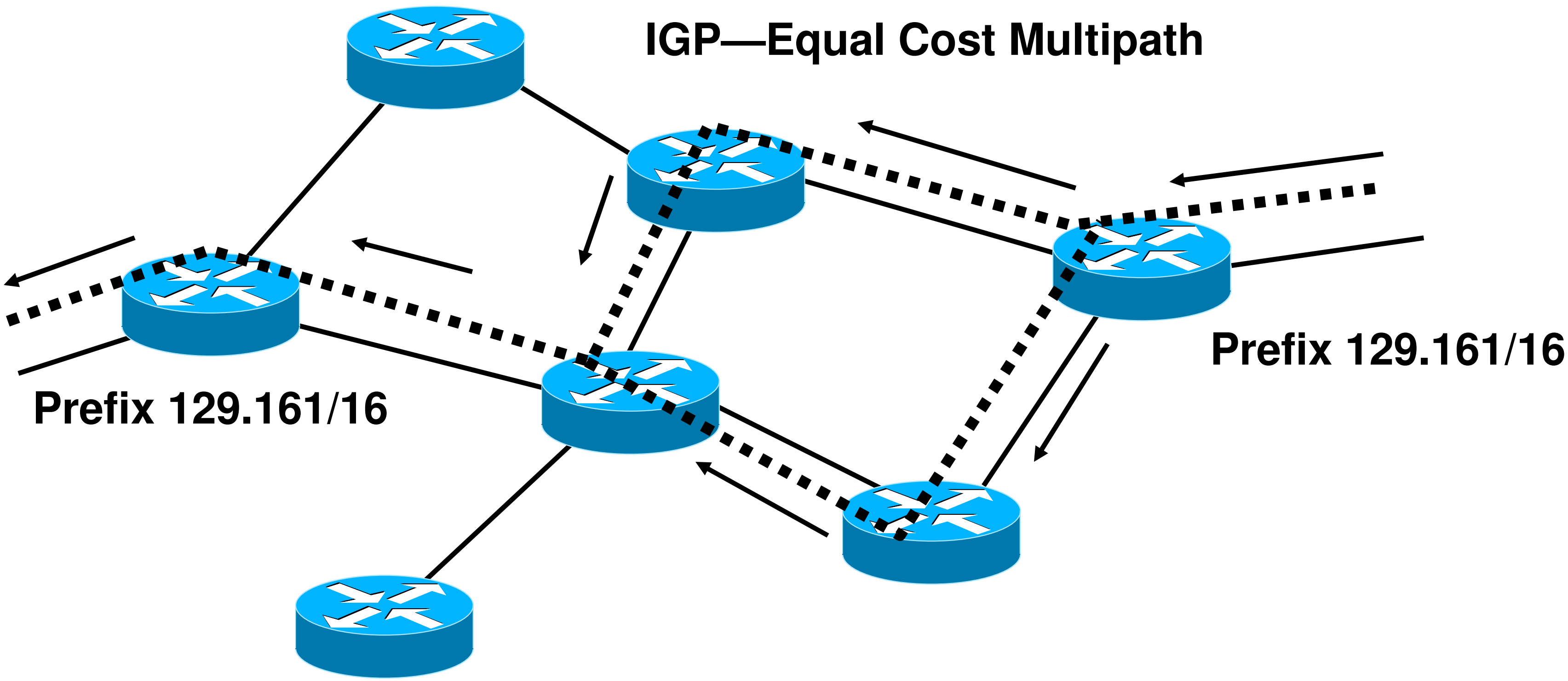
Label Distribution Protocol

- **Label Merge**

  Done by default for packet networks—unique label advertised per FEC

  Requires VC merge for ATM networks

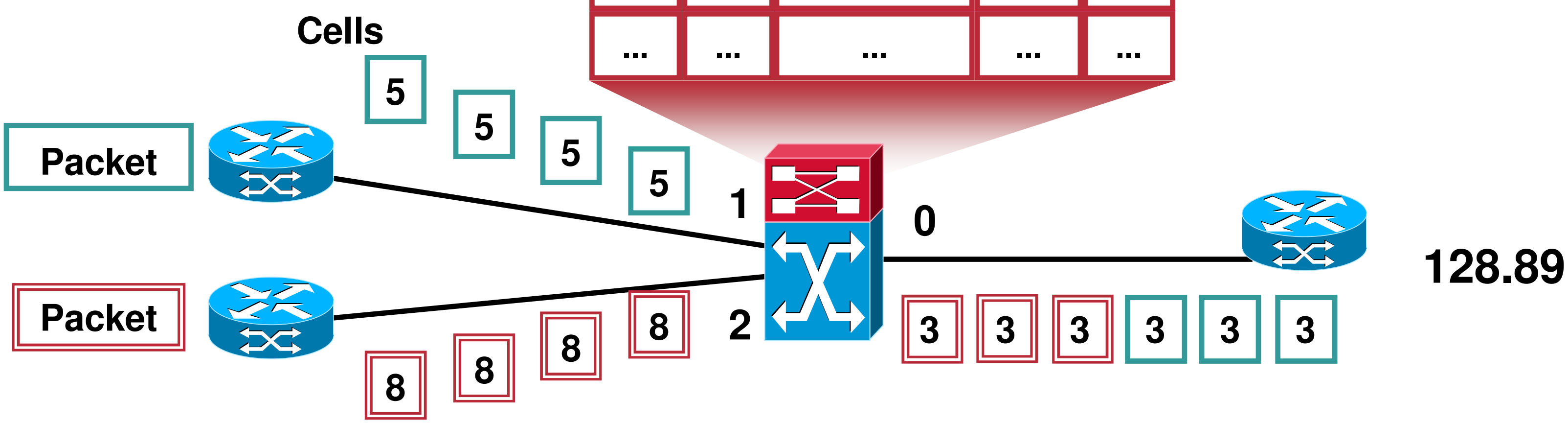
LDP—Label Merge



Labels for Prefix 129.161 Are Advertised Along both Paths

VC Merge

In I/F	In label	Address Prefix	Out I/F	Out label
1	5	128.89	0	3
2	8	128.89	0	3
...	...	...	...	...



- **With ATM switch that can merge VC's:**
  - **Can reuse outgoing label**
  - **Hardware prevents cell interleave**
  - **Fewer labels required**
  - **For very large networks**

LDP

- **Neighbor discovery**
  - Discover directly attached Neighbors—pt-to-pt links (including Ethernet)
  - Establish a session
  - Exchange prefix/FEC and label information
- **Extended Neighbor Discovery**
  - Establish peer relationship with another router that is not a neighbor
  - Exchange FEC and label information
  - May be needed to exchange service labels

TDP and LDP

- Tag Distribution Protocol—Cisco proprietary
  - Pre-cursor to LDP
  - Used for Cisco Tag Switching
- TDP and LDP supported on the same device
  - Per neighbor/link basis
  - Per target basis
- LDP is a superset of TDP
- Uses the same label/TAG
- Has different message formats

Configuring MPLS

Step 1	`Router# configure terminal`	Enables configuration mode
Step 2	`Router(config)# ip cef [distributed]`	Configures Cisco Express Forwarding
Step 3	`Router(config)# interface interface`	Specifies the interface to configure
Step 4	`Router(config-if)# mpls ip`	Configures MPLS hop-by-hop forwarding for a specified interface
Step 5	`Router(config-if)# mpls label protocol ldp`	Configures the use of LDP for a specific interface; Sets the default label distribution protocol for the specified interface to be LDP, overriding any default set by the global mpls label protocol command
Step 6	`Router# configure terminal` `Router(config)# mpls label protocol ldp`	Configures the use of LDP on all interfaces; Sets the default label distribution protocol for all interfaces to be LDP

Show Commands

Router# **show mpls interfaces**
Interface IP Tunnel Operational
Ethernet1/1/1 Yes (tdp) No No
Ethernet1/1/2 Yes (tdp) Yes No
Ethernet1/1/3 Yes (tdp) Yes Yes
POS2/0/0 Yes (tdp) No No
ATM0/0.1 Yes (tdp) No No (ATM labels)
ATM3/0.1 Yes (ldp) No Yes (ATM labels)
ATM0/0.2 Yes (tdp) No Yes

Router# **show mpls ldp discovery**
Local LDP Identifier:
118.1.1.1:0
Discovery Sources:
Interfaces:
POS2/0 (ldp): xmit/recv
LDP Id: 155.0.0.55:0
Tunnel1 (ldp): Targeted -> 133.0.0.33
Targeted Hellos:
118.1.1.1 -> 133.0.0.33 (ldp): active, xmit/recv
LDP Id: 133.0.0.33:0
118.1.1.1 -> 168.7.0.16 (tdp): passive, xmit/recv
TDP Id: 168.7.0.16:0

show mpls ip binding [**vrf** ***vpn-name***] [***network*** {***mask*** | ***length***} [**longer-prefixes**]]
[**local-label** {**atm** vpi vci **|** label [**-** ***label***]}]
[**remote-label** {**atm** vpi *vci* **|** label [**-** ***label***]}]
[**neighbor** address] [**local**]
[**interface** interface] [**generic** | **atm**]
show mpls ip binding summary

Router# **show mpls ip binding 194.44.44.0 24**
194.44.44.0/24
in label: 24
in vc label: 1/37 lsr: 203.0.7.7:2 ATM1/0.8
Active egress (vcd 56)
out label: imp-null lsr: 155.0.0.55:0 inuse
Router#

Other Label Distribution Protocols—RSVP

- Used in MPLS Traffic Engineering
- Additions to RSVP signaling protocol
- Leverage the admission control mechanism of RSVP to create an LSP with bandwidth
- Label requests are sent in PATH messages and binding is done with RESV messages
- EXPLICT-ROUTE object defines the path over which setup messages should be routed
- Using RSVP has several advantages

Other Label Distribution Protocols—BGP

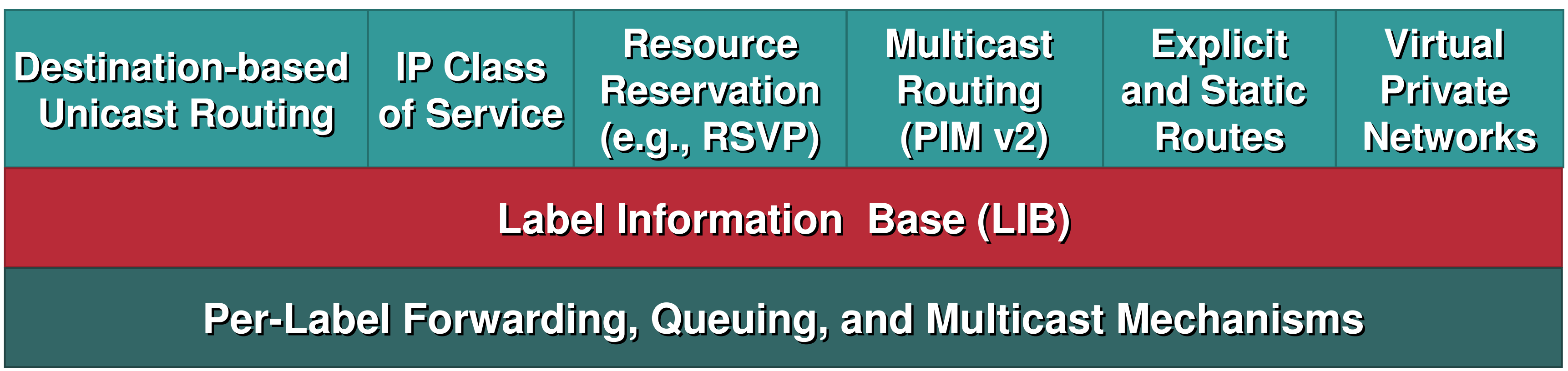
- Used in the context of MPLS VPNs
- Need multiprotocol extensions to BGP
- Routers need to be BGP peers
- Label mapping info carried as part of NLRI (Network Layer Reacheability Information)

Basic MPLS Operation - recap

- IP packets are classified in FECs
  - Forwarding Equivalence Class
- A group of IP packets which are forwarded in the same manner
  - Over the same path
  - With the same forwarding treatment
- Packet forwarding consists on
  - Assign a packet to a FEC
  - Determine the next-hop of each FEC

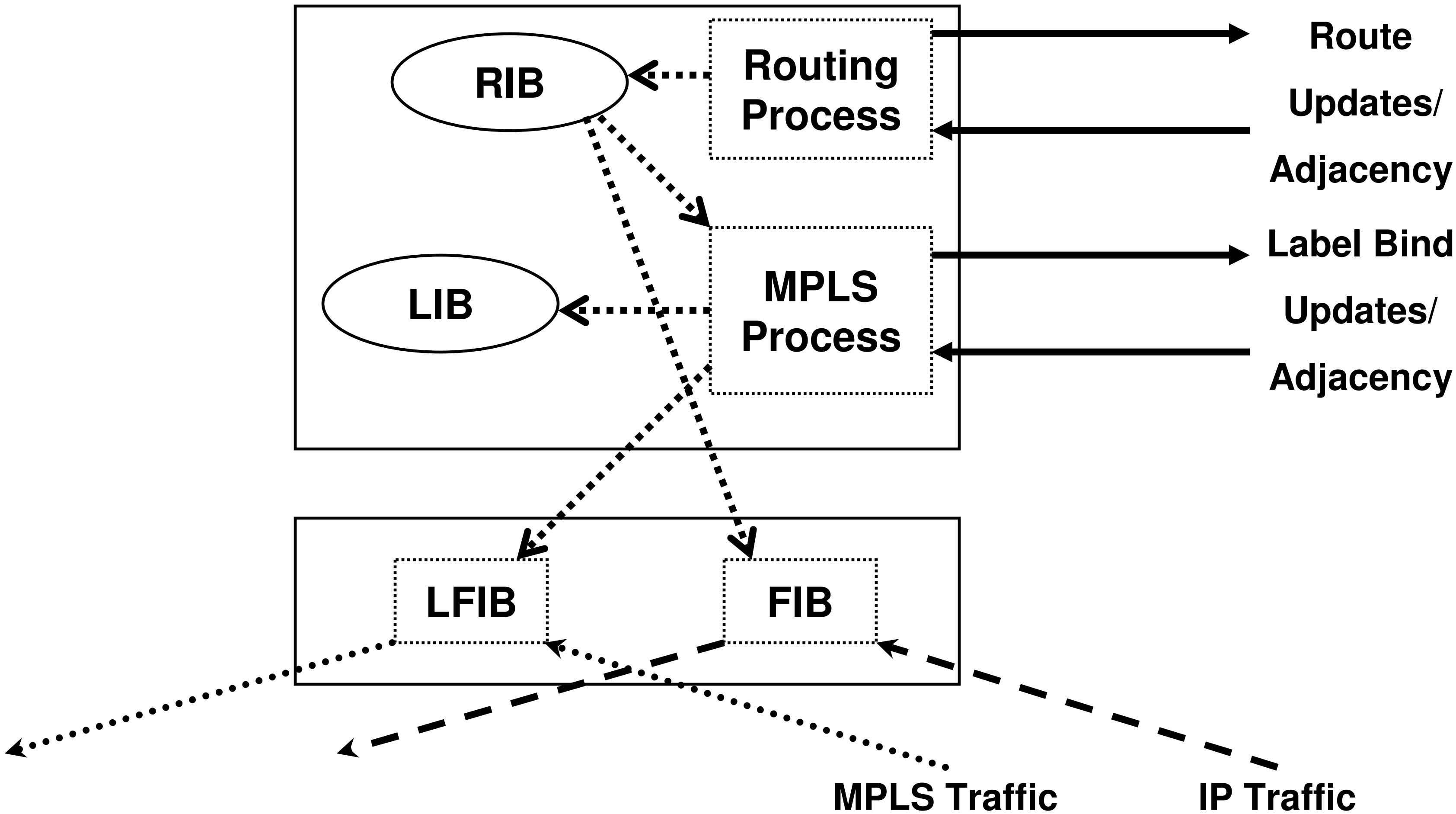
MPLS Control and Forwarding Planes

- Control plane used to distribute labels—BGP, LDP, RSVP
- Forwarding plane consists of label imposition, swapping and disposition—no matter what the control plane
- Key: There is a separation of Control Plane and Forwarding Plane
  - Basic MPLS: destination-based unicast
  - Labels divorce forwarding from IP address
  - Many additional options for assigning labels
  - Labels define destination and service



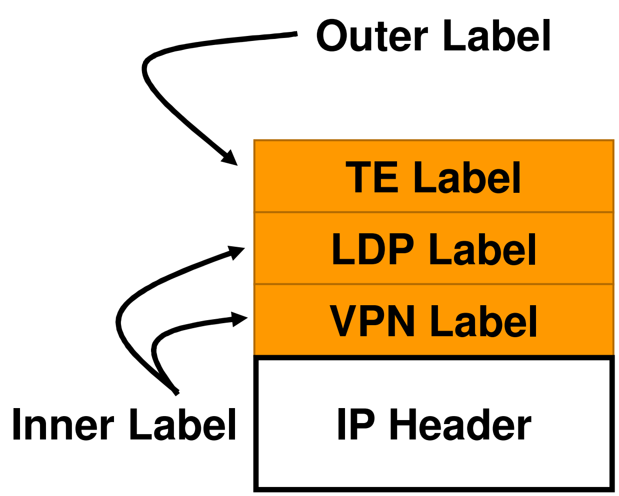
Control and Forward Plane Separation

Cisco.com



Label Stacking

- There may be more than one label in an MPLS packet
- As we know Labels correspond to forwarding equivalence classes
    - Example—There can be one label for routing the packet to an egress point and another that separates a customer A packet from Customer B
    - Inner labels can be used to designate services/FECs etc
        - E.g VPNs, Fast Re-route
- Outer label used to route/switch the MPLS packets in the network
- Last label in the stack is marked with EOS bit
- Allows building services such as
    - MPLS VPNs
    - Traffic Engineering and Fast Re-route
    - VPNs over Traffic Engineered core
    - Any Transport over MPLS

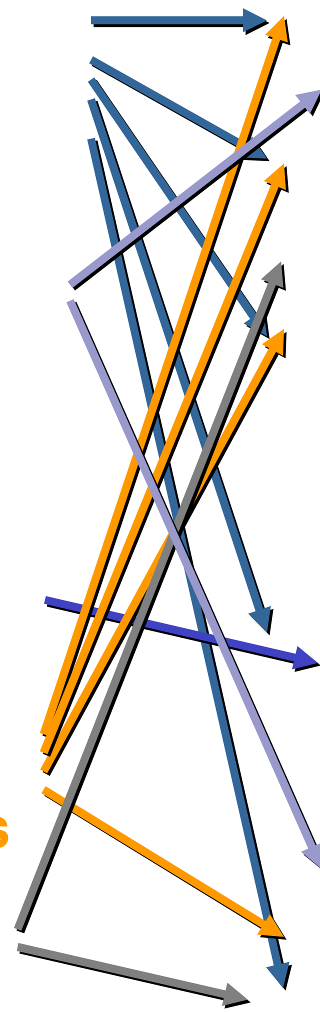


MPLS-Based Services

Azhar Sayeed

MPLS and Its Applications

- Separate forwarding information (label) from the content of IP header
- Single forwarding paradigm (label swapping)—multiple routing paradigms
- Multiple link-specific realizations of the label swapping forwarding paradigm
- Flexibility of forming FECs
- Forwarding hierarchy via label stacking



- Traffic engineering
- Fast re-route
- “Hard” QoS support
- Integration with optical cross connects
- Scalable VPN

Agenda

- MPLS and MPLS-VPN Overview
- MPLS-VPN Deployment Considerations
- Traffic Engineering
- Management Considerations and MPLS OAM
- Security Considerations
- Word About G-MPLS

MPLS and MPLS-VPN Overview

MPLS VPNs

Layer 2 and Layer 3

Monique Morrow

What Is a VPN ?

- VPN is a set of sites which are allowed to communicate with each other
- VPN is defined by a set of administrative policies
    - Policies determine both connectivity and QoS among sites
    - Policies established by VPN customers
    - Policies could be implemented completely by VPN Service Providers
        - Using BGP/MPLS VPN mechanisms

What Is a VPN (Cont.)?

- **Flexible inter-site connectivity**

    ranging from complete to partial mesh

- **Sites may be either within the same or in different organizations**

    VPN can be either intranet or extranet

- **Site may be in more than one VPN**

    VPNs may overlap

- **Not all sites have to be connected to the same service provider**

    VPN can span multiple providers

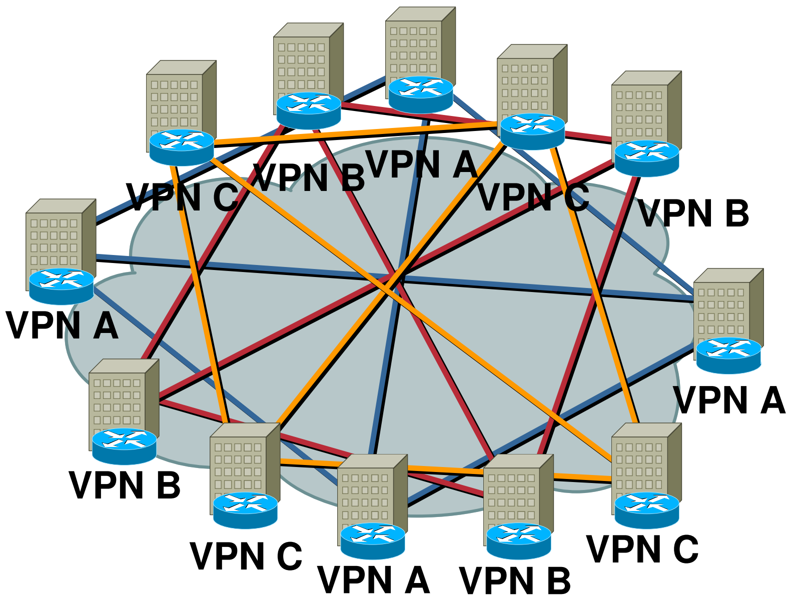
VPNs

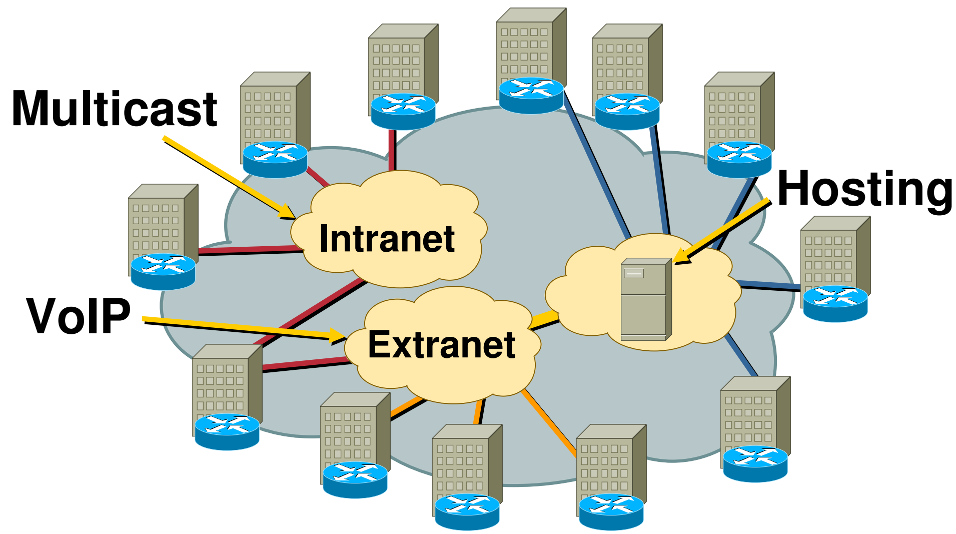
- **Layer 2 VPNs**
  - Customer End points (CPE) connected via layer 2 such as Frame Relay DLCI, ATM VC or point to point connection
  - If it connects IP routers then peering or routing relationship is between the end points
  - Multiple logical connections (one with each end point)
- **Layer 3 VPNs**
  - Customer end points peer with provider routers
  - Single peering relationship
  - No mesh of connections
  - Provider network responsible for
    - Distributing routing information to VPN sites
    - Separation of routing tables from one VPN to another

Layer 3 VPNs

Monique Morrow

Service Provider Benefits of MPLS-Based VPNs





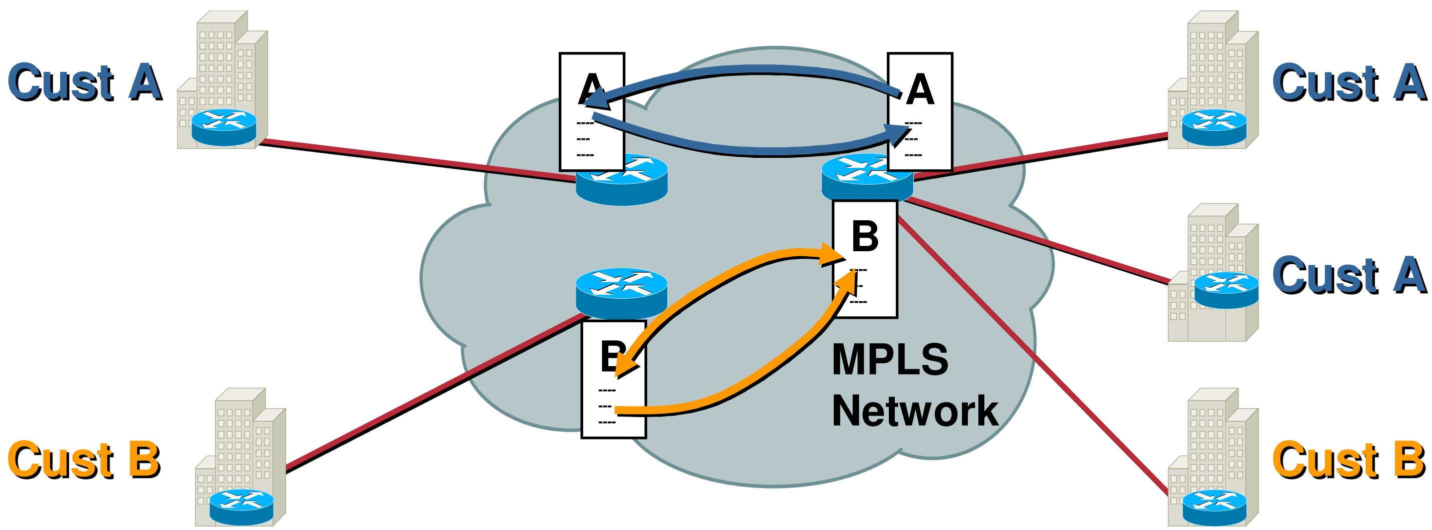
Overlay VPN

- Pushes content *outside* the network
- Costs scale exponentially
- Transport dependent
- Groups endpoints, not groups
- Complex overlay with QoS, tunnels, IP

MPLS-based VPNs

- Enables content hosting inside the network
- "Flat" cost curve
- Transport independent
- Easy grouping of users and services
- Enables QoS inside the VPNs

Using Labels to Build an IP VPN



- **The network distributes labels to each VPN**
  - **Only labels for other VPN members are distributed**
  - **Each VPN is provisioned automatically by IP routing**
- **Privacy and QoS of ATM without tunnels or encryption**
  - **Each network is as secure as a Frame Relay connection**
- **One mechanism (labels) for QoS and VPNs—no tradeoffs**

How Does It Work?

Cisco.com

- **Simple idea**
  - Use a label to designate VPN prefix
  - Route that VPN packet to egress PE advertising that prefix
    - Use the IGP label to the VPN packet to the egress node
- **How is it done?**
  - Routers need to maintain separate VPN routing tables called VRFs (Virtual Routing and Forwarding Tables)
  - Routers then export and import routes using BGP extensions to identify and separate one VPNs routes from another
  - Routers then exchange labels for VPN routes in addition to IGP routes

VRFs

- A VRF is associated to one or more interfaces on a router
- VRF is essentially a per-interface routing table and the necessary forwarding operations (CEF)
- Not virtual routers, just virtual routing and forwarding
- VRFs are IP only (no Appletalk-VRF, although in theory it's certainly possible)

VRFs

Cisco.com

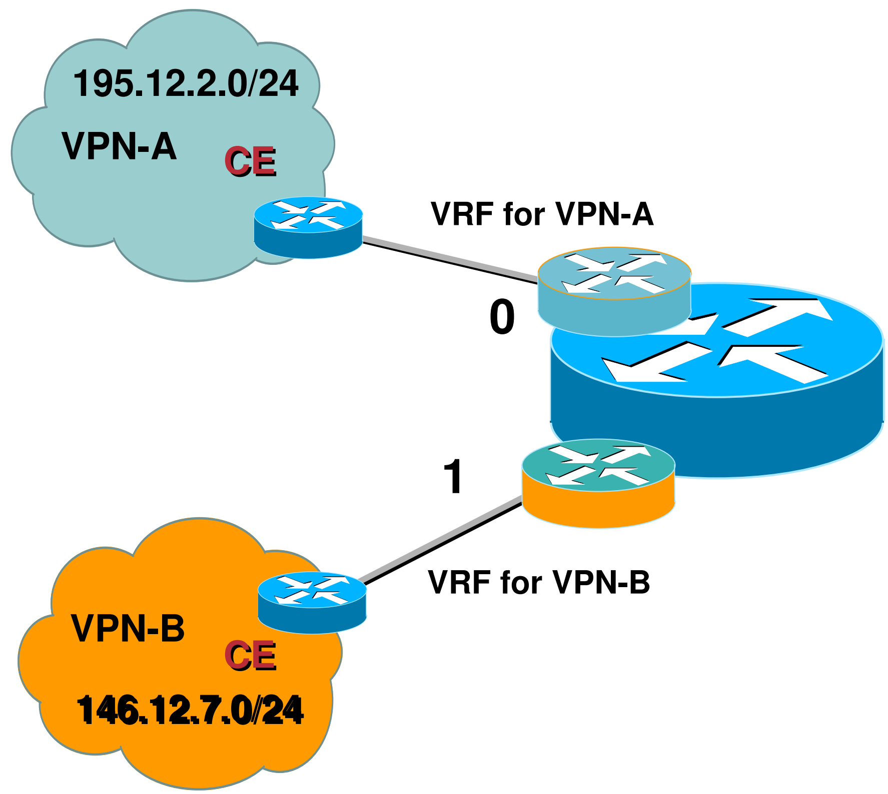
- Within a VRF, provider speaks a routing protocol with their customer
- Most protocols are supported
  - Static routes
  - RIP
  - BGP
  - EIGRP
  - OSPF
- No IS-IS support yet (have not seen the demand)
- No IGRP or EGP support either (same idea)
- Routes flow between VRF IGP/BGP and provider BGP (see VPNv4)

Virtual Routing and Forwarding Instances

- Define a VRF for interface 0
- Define a different VRF for interface 1
- Packets will never go between int. 0 and 1 unless allowed by VRF policy

  Will explain this policy in the next section

- No MPLS yet…



Carrying VPN Routes in BGP

Cisco.com

- VRFs by themselves are not 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

Cisco.com

- 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
  - So route reflectors make a bestpath decision on something other than 32-bit network + 32-bit mask

VPNv4

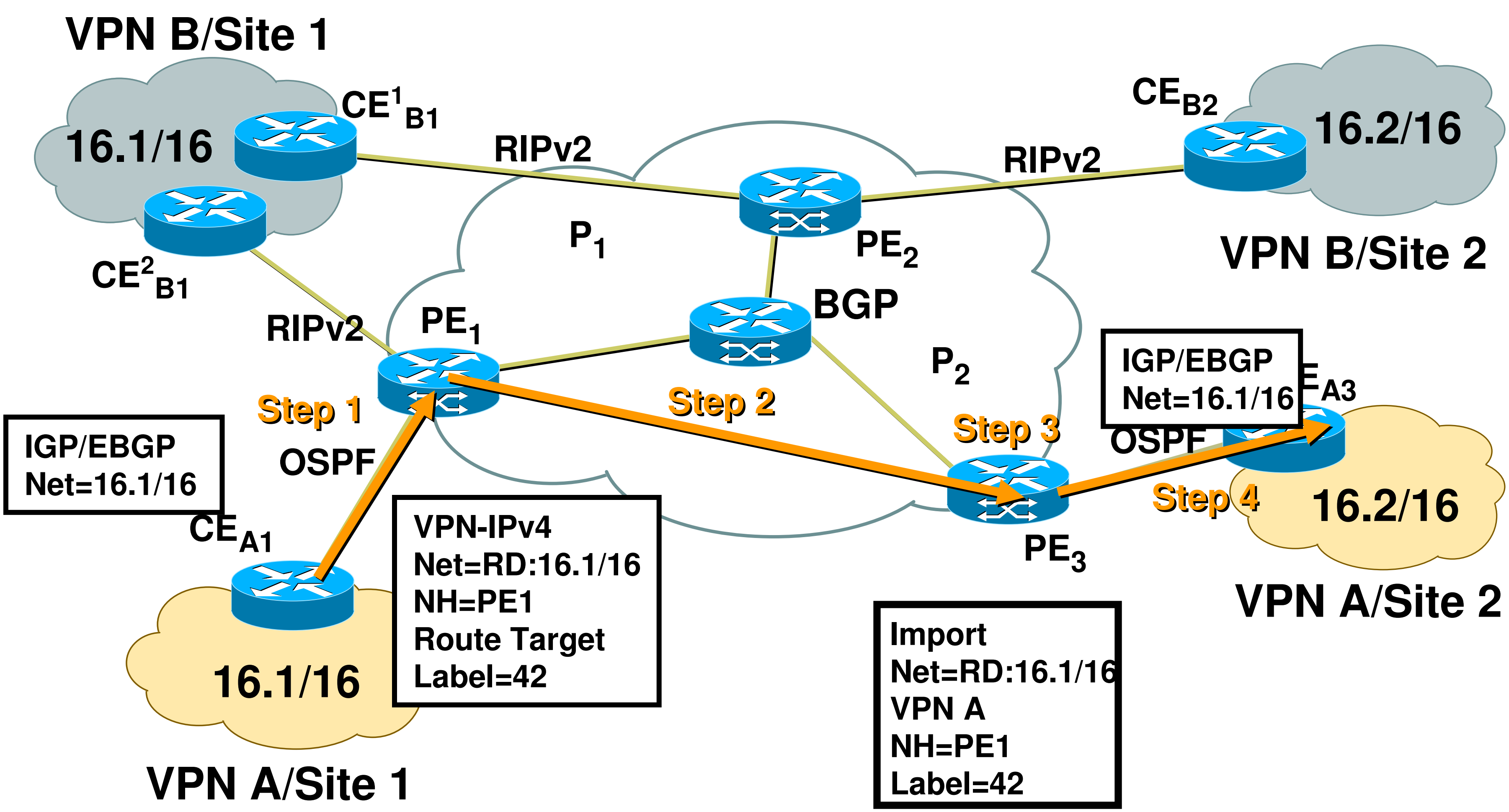
- In BGP for IP, 32-bit address + mask makes a unique announcement
- In BGP for MPLS-VPN, (64-bit RD + 32-bit address) + 32-bit mask makes a unique announcement
- Since the route encoding is different, need a different address family in BGP
- VPNv4 = VPN routes for IPv4
  - As opposed to IPv4 or IPv6 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"

Route Target

Cisco.com

- 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

Putting It All Together—Control Plane



MPLS-VPN Packet Forwarding

- Between PE and CE, regular IP packets (for now)
- Within the provider network—label stack

    Outer label: "get this packet to the egress PE"

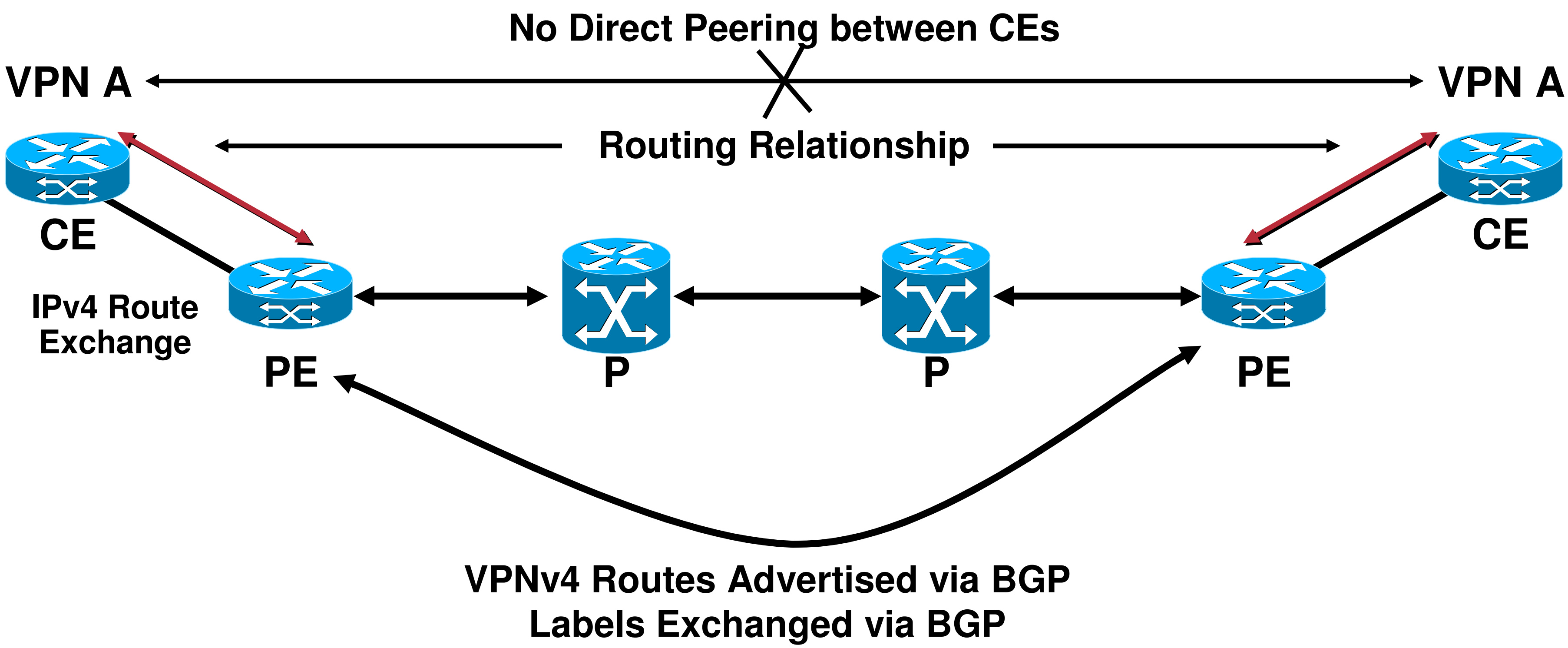
    Inner label: "get this packet to the egress CE"

Where Do Labels Come From?

Cisco.com

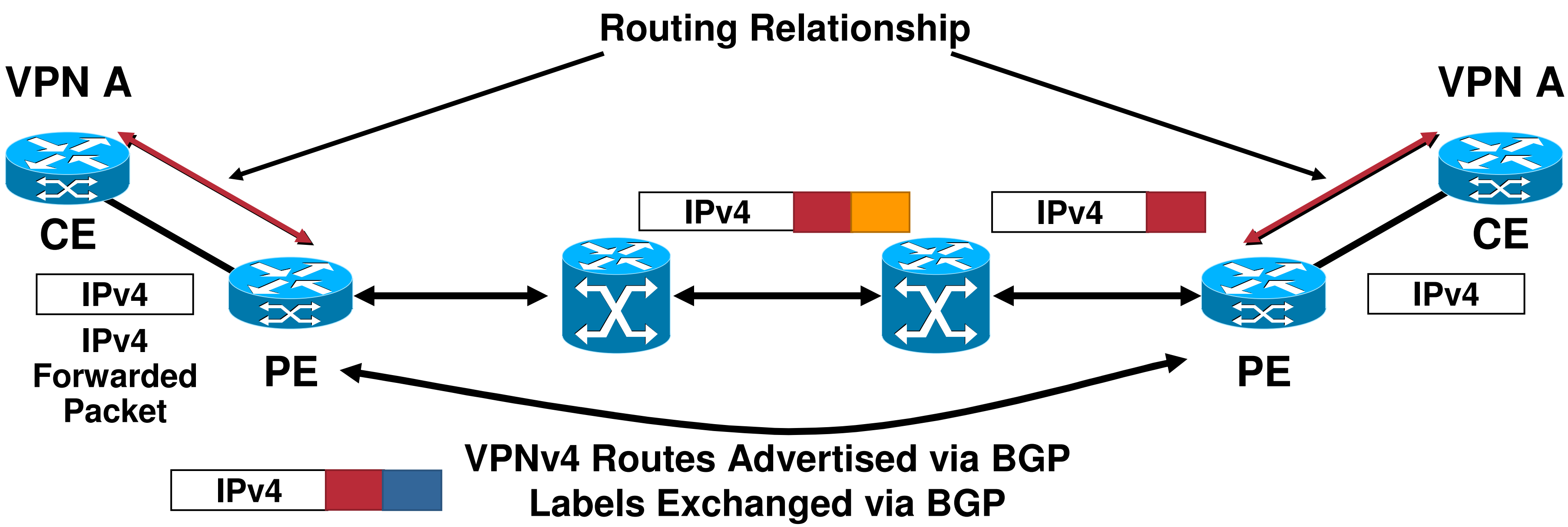
- Within a single network, can use LDP or RSVP to distribute IGP labels
- LDP follows the IGP
- RSVP (for TE) deviates from IGP shortest path
- Which IGP label distribution method you use is independent of any VPN label distribution

Control Plane Path



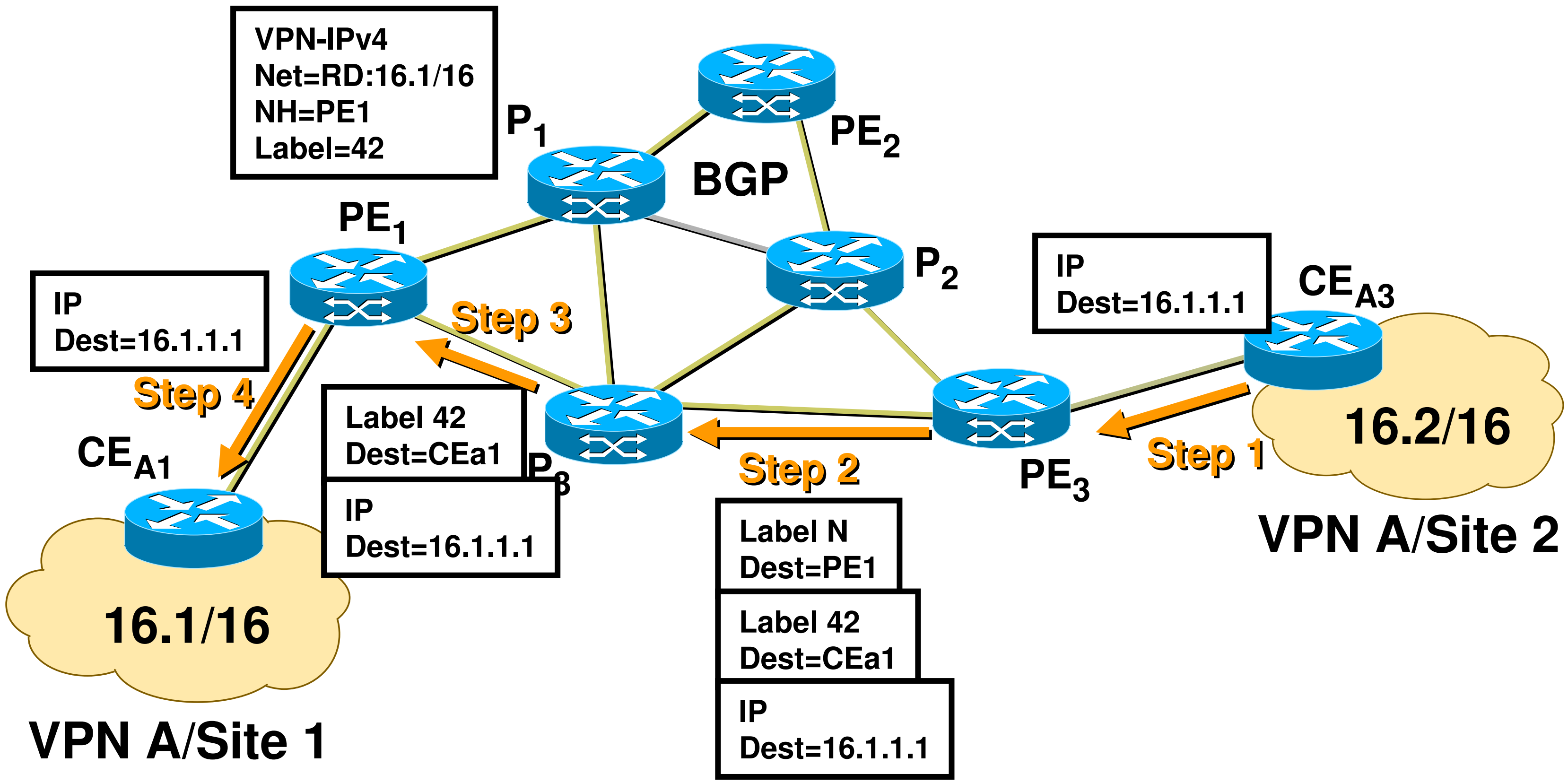
- RD—8 Byte field—assigned by provider—significant to the provider network only
- VPNv4 Address: RD+VPN Prefix
- Unique RD per VPN makes the VPNv4 address unique

Data Plane Path

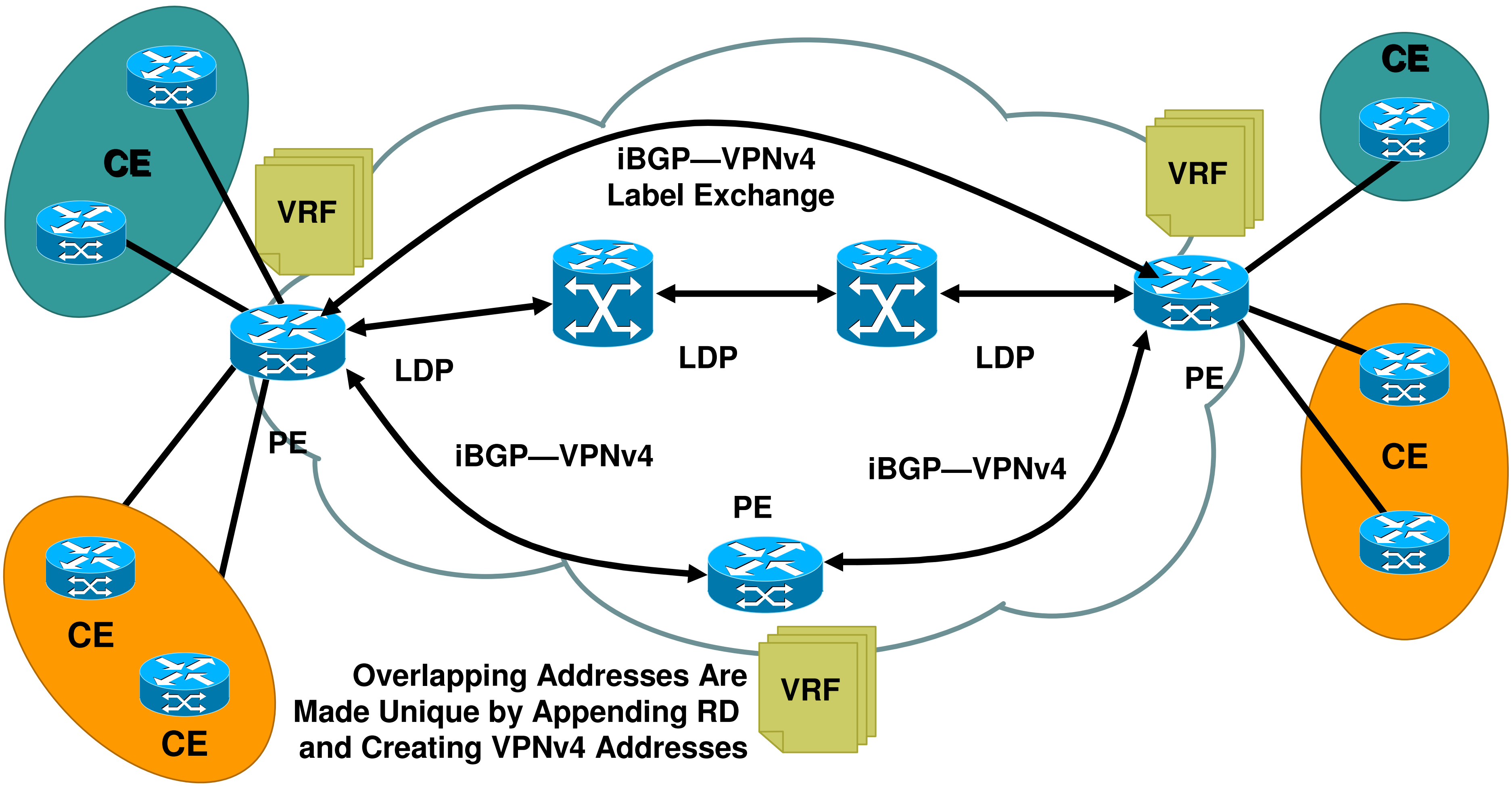


- Ingress PE is imposing 2 labels

Putting It All Together—Forwarding Plane



RFC 2547—MPLS VPNs



MPLS-VPN Deployment Considerations

Import/Export Policies

- **Full mesh:**

  All sites import X:Y and export X:Y

- **Hub and spoke:**

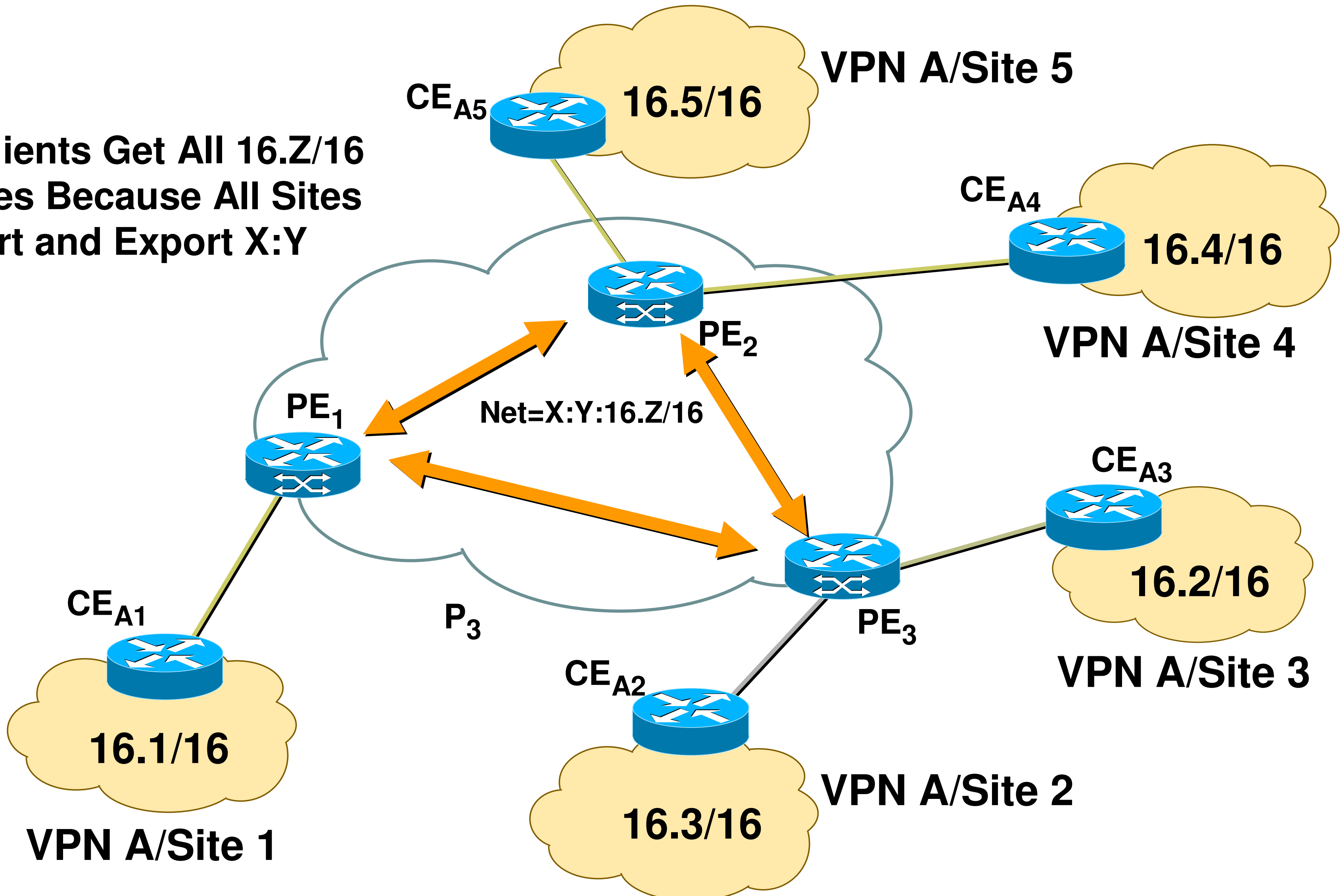
  Hub exports X:H and imports X:S

  Spokes export X:S and import X:H

Full Mesh

Cisco.com

All Clients Get All 16.Z/16 Routes Because All Sites Import and Export X:Y



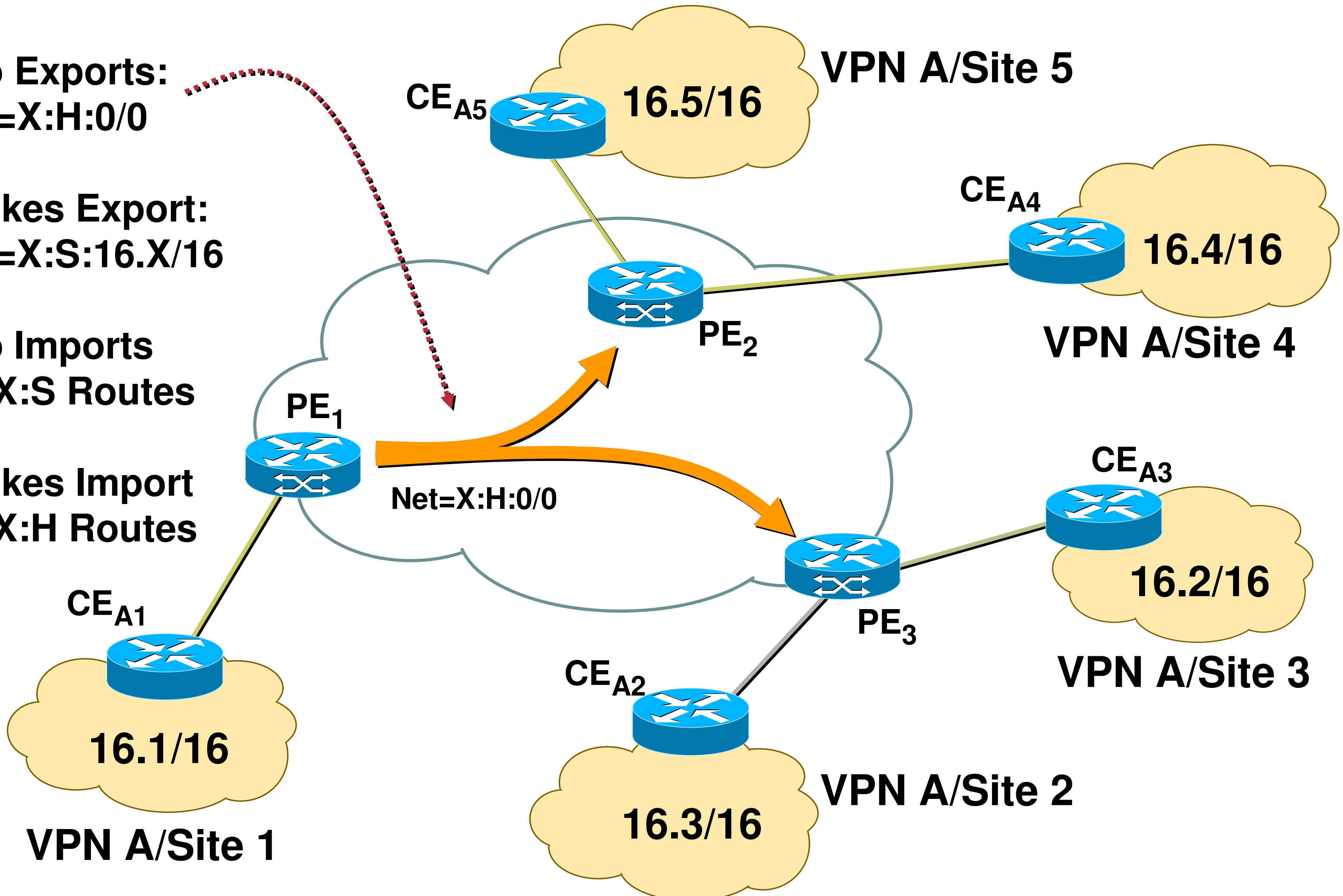
Hub and Spoke

1) Hub Exports:
   Net=X:H:0/0

2) Spokes Export:
   Net=X:S:16.X/16

3) Hub Imports
   All X:S Routes

4) Spokes Import
   All X:H Routes



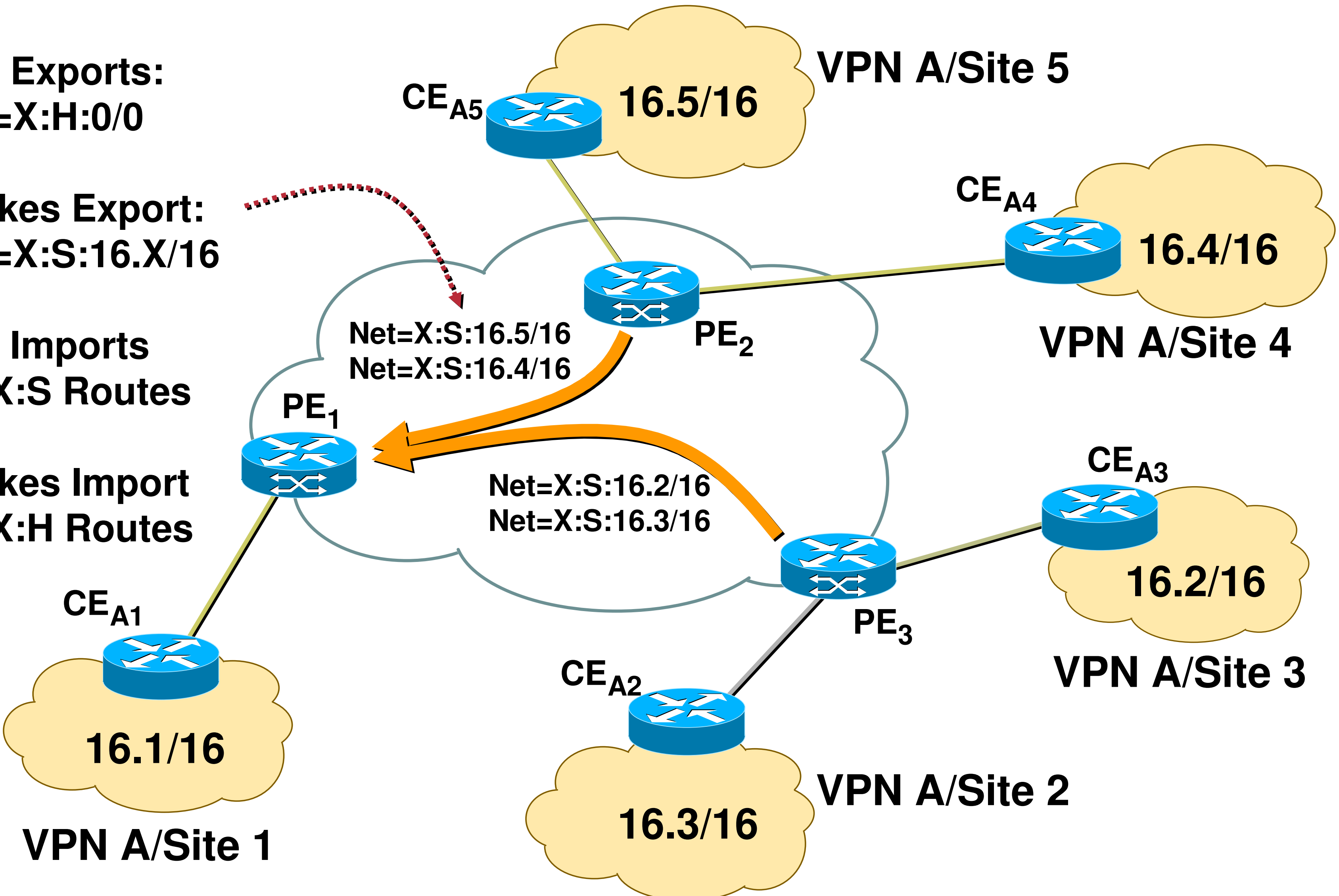
Hub and Spoke

1) Hub Exports:
   Net=X:H:0/0

2) Spokes Export:
   Net=X:S:16.X/16

3) Hub Imports
   All X:S Routes

4) Spokes Import
   All X:H Routes



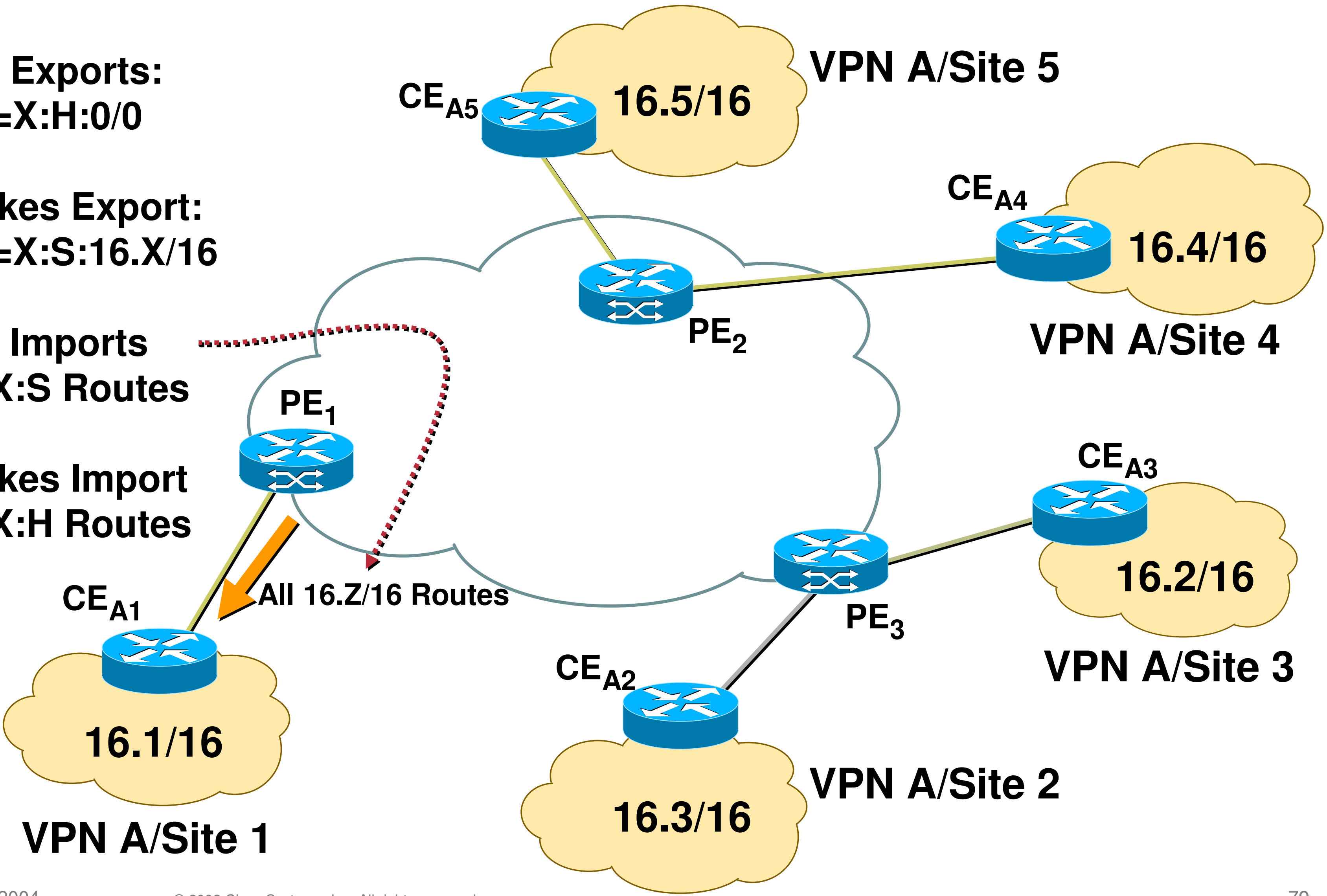
Hub and Spoke

1) Hub Exports:
   Net=X:H:0/0

2) Spokes Export:
   Net=X:S:16.X/16

3) Hub Imports
   All X:S Routes

4) Spokes Import
   All X:H Routes



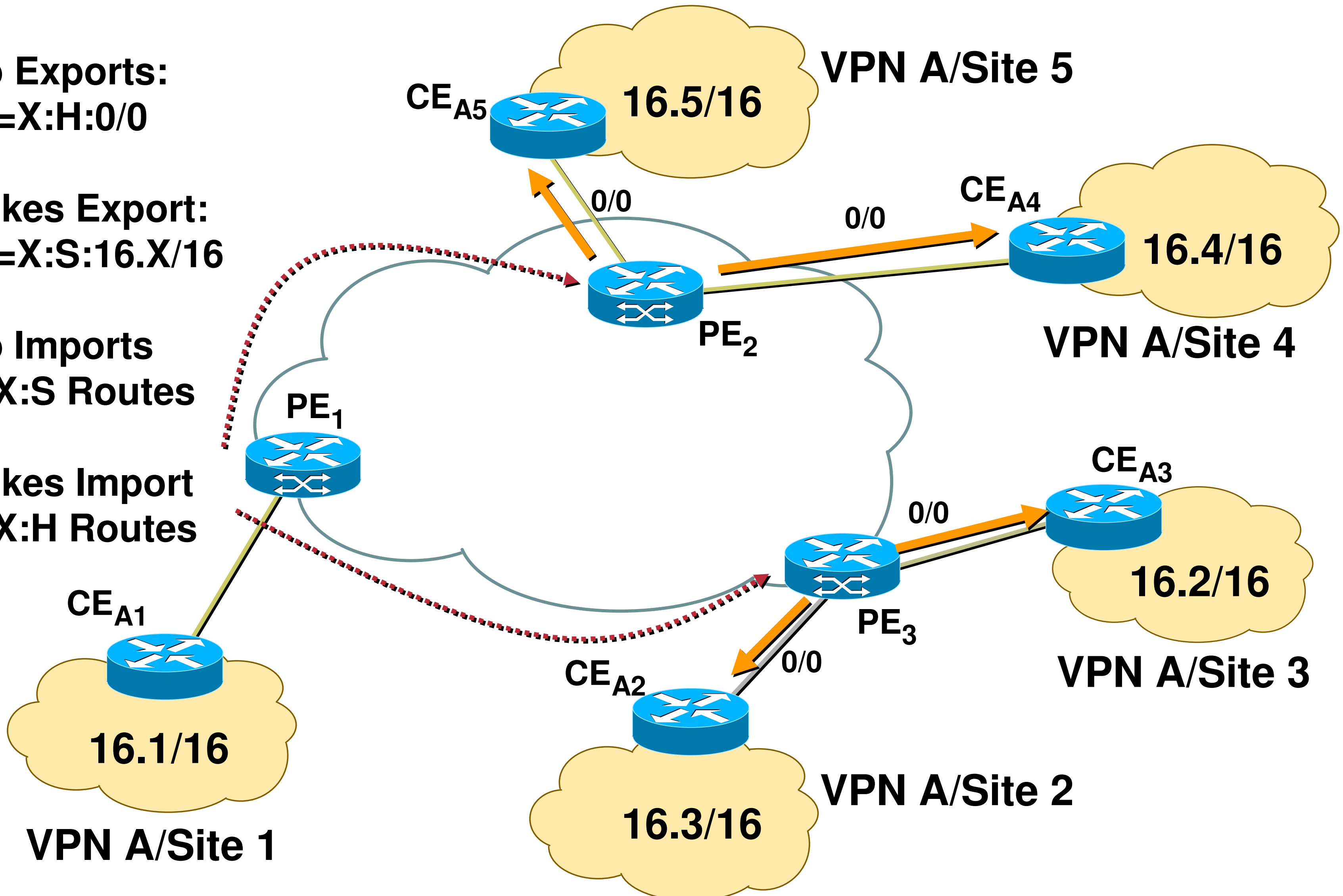
Hub and Spoke

1) Hub Exports:
   Net=X:H:0/0

2) Spokes Export:
   Net=X:S:16.X/16

3) Hub Imports
   All X:S Routes

4) Spokes Import
   All X:H Routes



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 orthogonal to MPLS-VPN

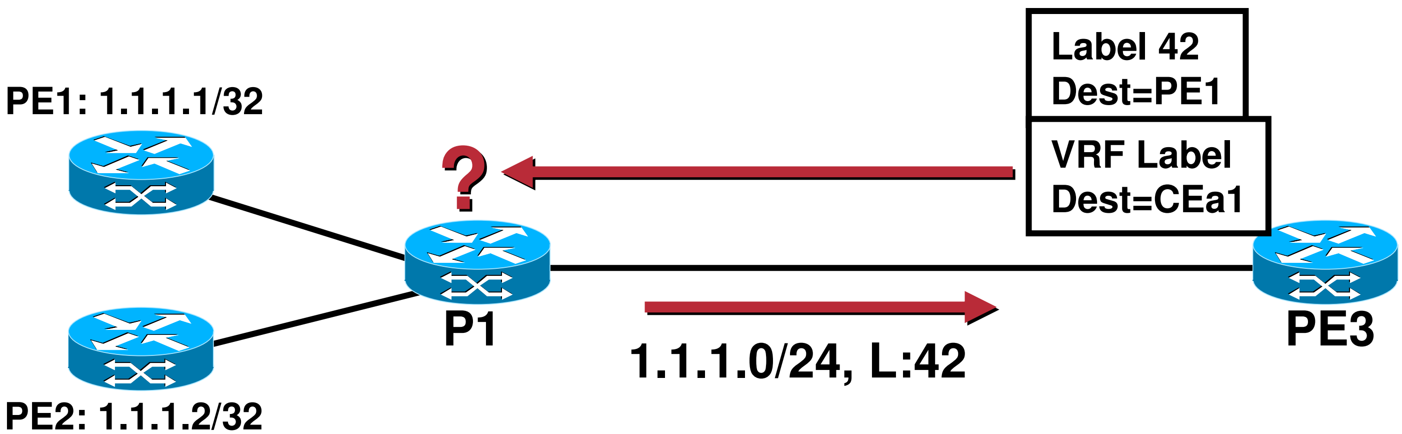
- Inner label is from BGP

  Inner label is there so the egress PE can have the same network in multiple VRFs

Things to Note

Cisco.com

- **Need /32s for all PEs if using LDP**
  - Outer label says "get me to this prefix"
  - If the prefix has a mask shorter than /32, can't guarantee we won't hit summarization at some point in the network
  - What does the summarization point do with the packet?



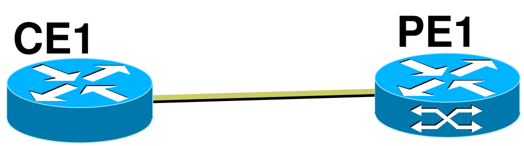
Prerequisites

`ip cef {distributed}`

`mpls ip` (on by default)

Global Config on PE

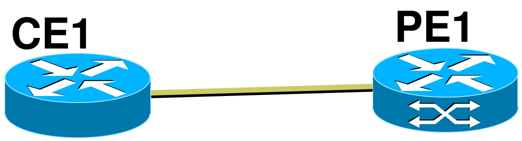
ip cef {distributed}
mpls ip (on by default)



Build a VRF

Global Config on PE

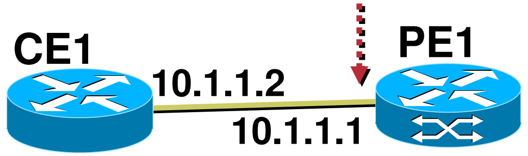
```
ip vrf foo
       rd 100:1
 route-target import 247:1
 route-target export 247:1b
```



Attach a VRF to a Customer Interface

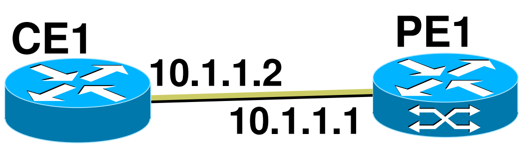
Cisco.com

```
interface Serial0
 ip vrf forwarding foo
 ip address 10.1.1.1 255.255.255.0
```



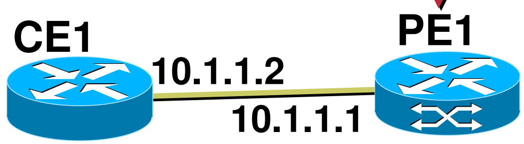
Run an IGP within a VRF—RIP


```
router rip
 address-family ipv4 vrf foo
  version 2
  no auto-summary
  network 10.0.0.0
 exit-address-family
```



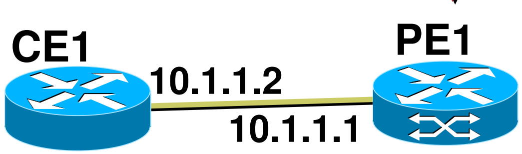
Run an IGP within a VRF—EIGRP


```
router eigrp 1
 address-family ipv4 vrf test
  network 10.1.1.0 0.0.0.255
  autonomous-system 1
  exit-address-family
```



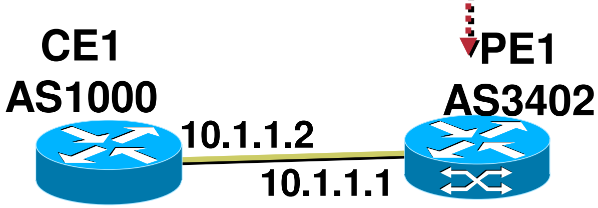
Run an IGP within a VRF—OSPF


```
router ospf 1 vrf test
 network 10.1.1.0 0.0.0.255 area 0
```



Run BGP within a VRF

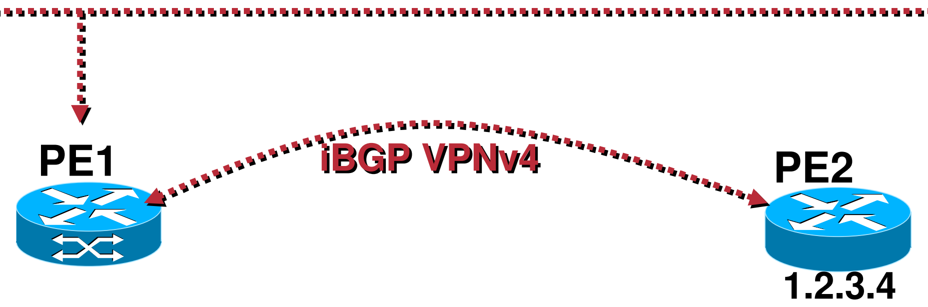

```
router bgp 3402
 address-family ipv4 vrf test
  neighbor 10.1.1.2 remote-as 1000
  neighbor 10.1.1.2 activate
  exit-address-family
```



Enable VPNv4 BGP in the Backbone

Cisco.com

```
router bgp 3402
 neighbor 1.2.3.4 remote-as 3402
 neighbor 1.2.3.4 update-source loopback 0
 address-family vpnv4
  neighbor 1.2.3.4 activate
  neighbor 1.2.3.4 send-community both
```

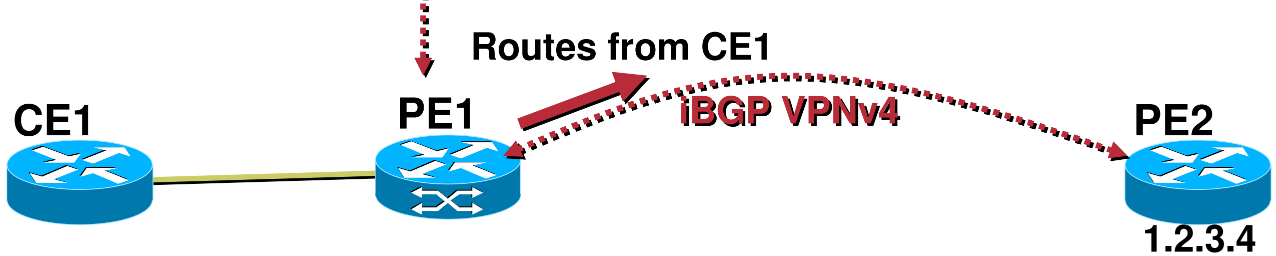


Get Routes from Customer Routing to VPNv4

Cisco.com

- If CE routing is not BGP, need to redistribute into BGP
- NOTE: this means you *need* an IPv4 VRF BGP context to get routes into the PE backbone, even if you don't have any BGP neighbors in the VRF
- IGP metric is usually carried as MED, unless changed
    - EIGRP is an exception, carries the 5-part metric as BGP extended communities

```
router bgp 34032
 neighbor 1.2.3.4 remote-as 3402
 neighbor 1.2.3.4 update-source loopback 0
 address-family ipv4 vrf test
  redistribute {rip|connected|static|eigrp|ospf}
```

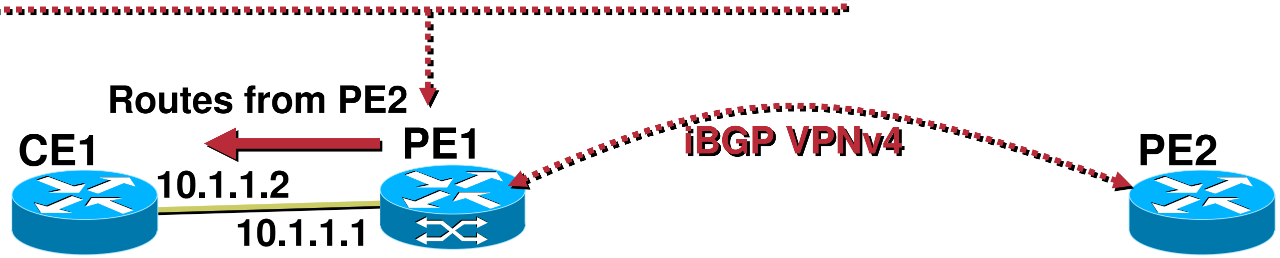


Get Routes from VPNv4 to Customer Routing

Cisco.com

- If CE routing is not BGP, need to redistribute from VPNv4 to CE routing
- Redistributing BGP into IGP makes some people nervous; don't worry about it, it's hard to screw up
  - Please note that "hard" != "impossible"…:)
- Metric is important when going from MED to RIP or EIGRP
  - Can also use default-metric or route-map

```
router rip
 address-family ipv4 vrf foo
  version 2
  redistribute bgp 3402 metric 1
  no auto-summary
  network 10.0.0.0
 exit-address-family
```



Diagnostics on the PE

- **Many commands have a 'vrf' keyword**
  - Ping, traceroute, telnet, etc
  - Pretty much every diagnostic command that makes sense

```
ping vrf test 10.1.1.1
trace vrf test 10.1.1.1
telnet 10.1.1.1 /vrf test
```

Diagnostics on the PE


```
show ip route vrf test
```

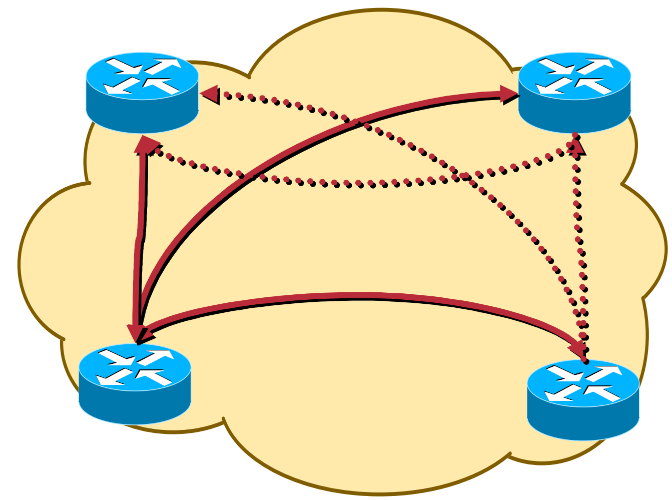
```
show ip cef vrf test
```

...etc...

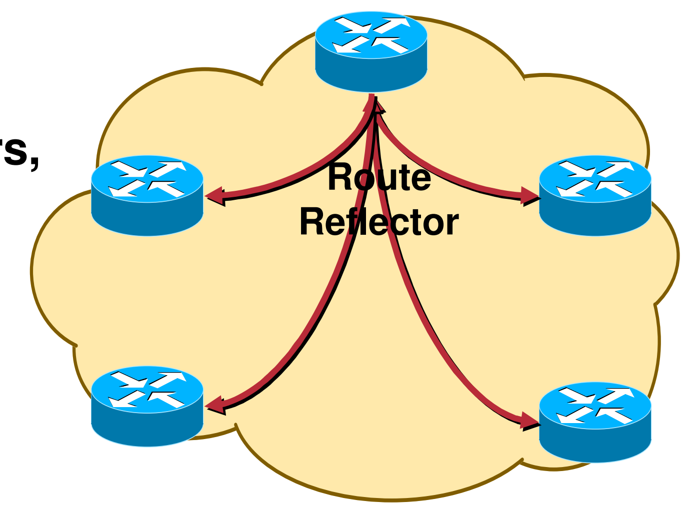
Route Reflectors

- Biggest scaling hurdle with MPLS-VPN is BGP
- Luckily, we have lots of experience scaling BGP
- Can use confederations or route reflectors
    - Confederations falling out of favor
- RRs make more sense when not every router needs all routes (i.e., Pes)
- Scaling is a little different
    - Currently ~120k Internet routes
    - Some customers are asking for 500k-1M VPNv4 routes
        - Largest in reality is closer to 200k-250k, but be prepared

Route Reflectors



- Full iBGP mesh is a lot of neighbors to maintain on every router
- N^2 provisioning when a PE is added, and VPN networks are growing constantly



- Route Reflector takes routes from neighbors, gives them to other neighbors
- Can build a dedicated RR that isn't used for forwarding, but which can hold lots of routes
- 1GB Memory, ~1,000,000 routes

Route Reflectors—Basic Configuration

Cisco.com

Client

```
neighbor 1.2.3.4 remote-as 3402
neighbor 1.2.3.4 update-source loopback0
```

Reflector

```
router bgp 3402
 [no bgp default route-target filter]
neighbor 1.2.3.6 remote-as 3402
neighbor 1.2.3.6 update-source loopback0
address-family vpnv4
 neighbor 1.2.3.6 route-reflector-client
```

On by Default If Configured with RR-clients

Route Reflectors—Peer Groups

- Use peer groups for a tremendous convergence improvement
- On the RR

```
neighbor foo peer-group
neighbor 1.2.3.6 peer-group foo
```

- ...then apply a common output policy to neighbor foo

Route Reflectors—Other Tips

- Peer-groups are such a powerful enhancement that the RR can be overwhelmed by ACKs from lots of clients
- Increase input hold-queue to hold these ACKs

```
Router(config-if)# hold-queue <x> in
```

- Default is 75, consider 500, 1,000, etc (max is 4,096)
- Memory consumed is (Qsize * ifMTU), so 1500byte MTU @1,000-packet depth = 1.5Mbyte per interface

  If you can't spare the 1.5Mb/interface, you probably shouldn't be a Route Reflector

Route Reflectors—Other Tips

- TCP MSS (max segment size) is 536 by default
- All backbone links now are MTU 1500 or higher (most ~4k)
- ‘`ip tcp path-mtu-discovery`’ to increase tcp MSS to fix in MTU
- Benefit: get BGP routes to peers faster, less protocol overhead

Advanced Services: Carrier Supporting Carrier

- **RFC3107 defines a way to exchange a label with an IPv4 (not VPNv4) BGP route**
- **This is useful to exchange label reachability for IPv4 prefixes between ASes**
- **Also used in Carrier's Carrier and Inter-AS**
- **Under IPv4 (or IPv4 vrf) address-family:**

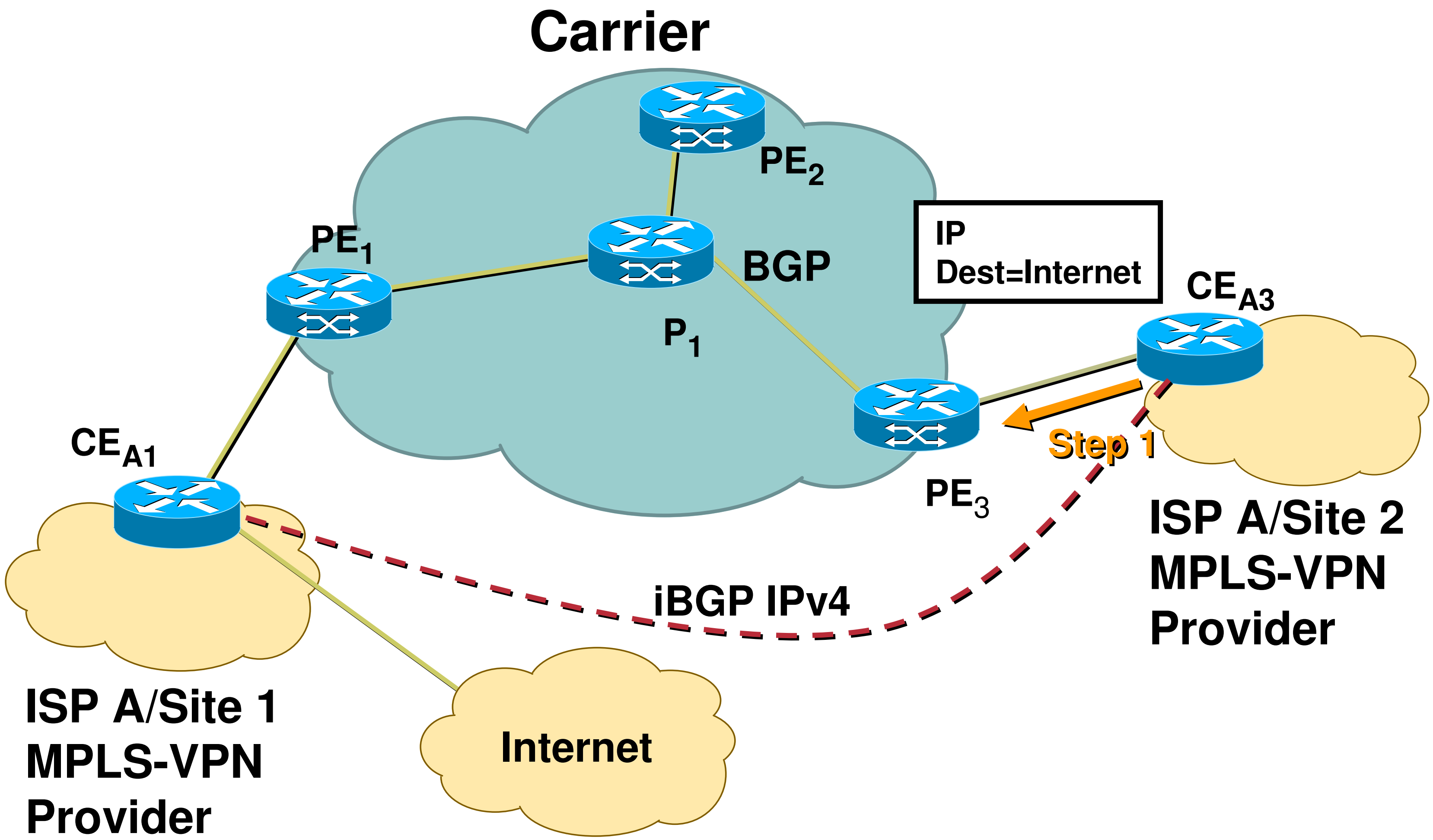
```
neighbor 1.2.3.4 send-label
```

Carrier's Carrier: The Problem

- MPLS-VPN works well for carrying customer IGPs
- Platforms, network scale to N*O(IGP) routes
- What if the CE wants the PE to carry all their BGP routes?
- Or if CE wants to run their own VPN service?

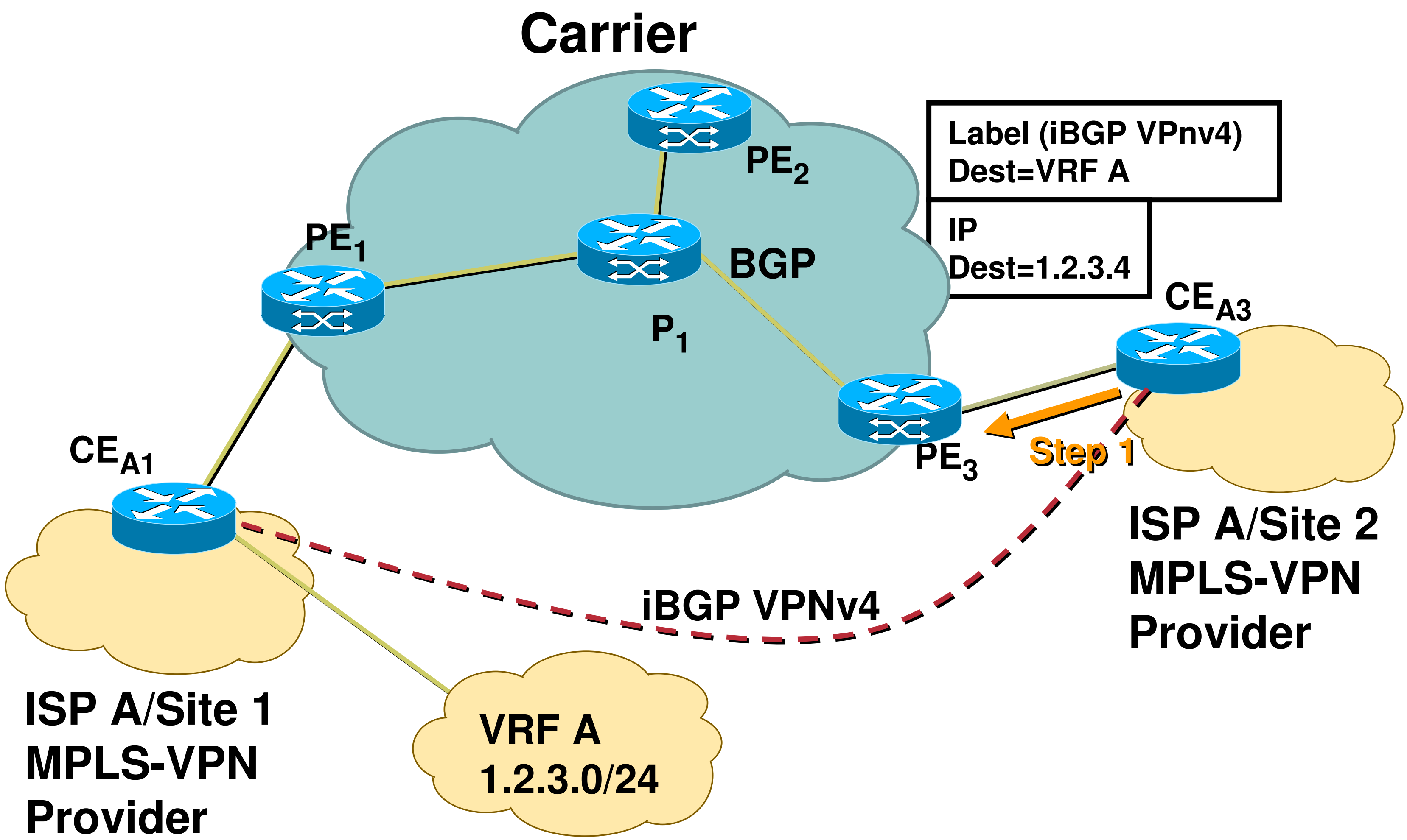
Carrier's Carrier: The Problem (Internet)

Cisco.com

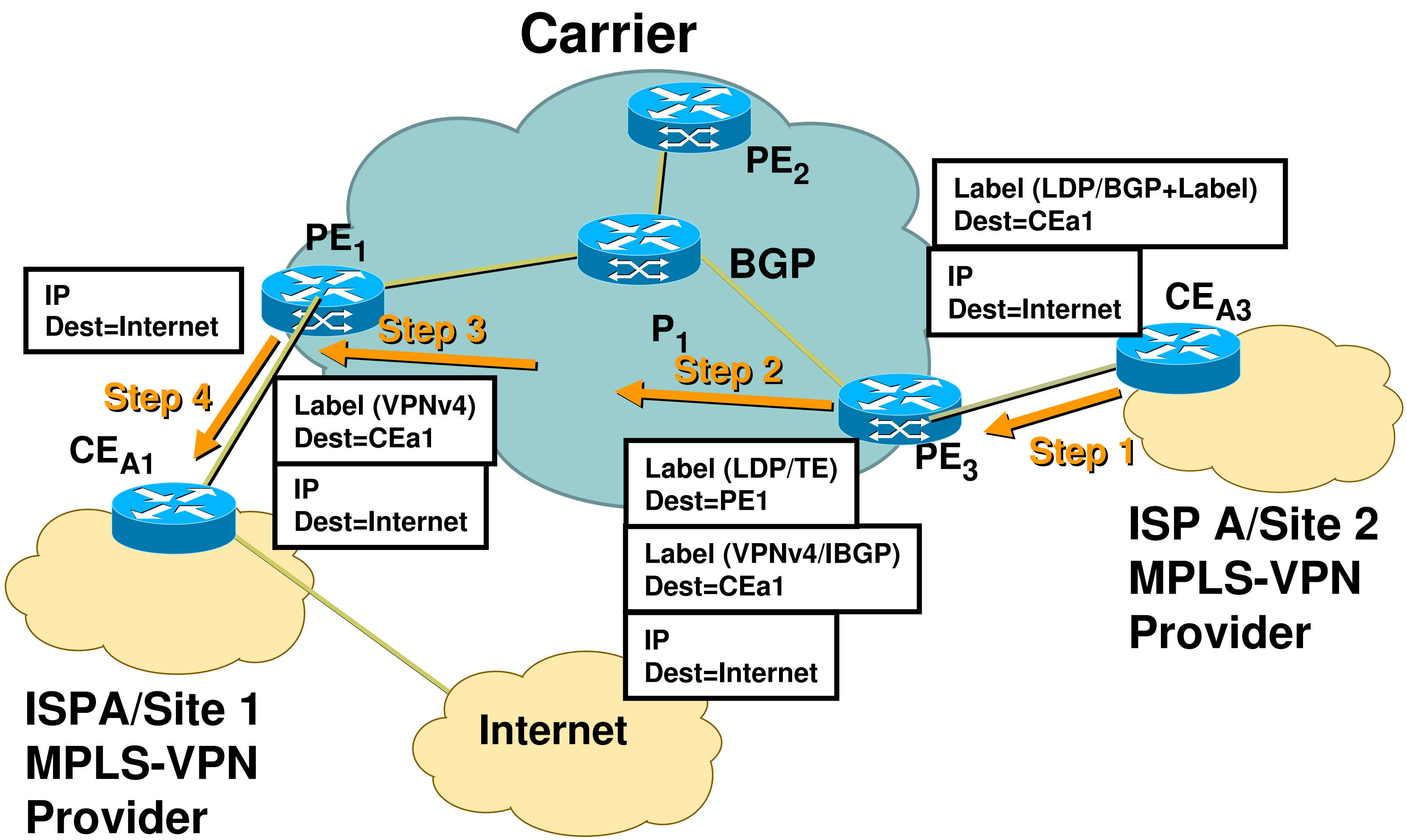


Carrier's Carrier: The Problem (VPN)

Cisco.com

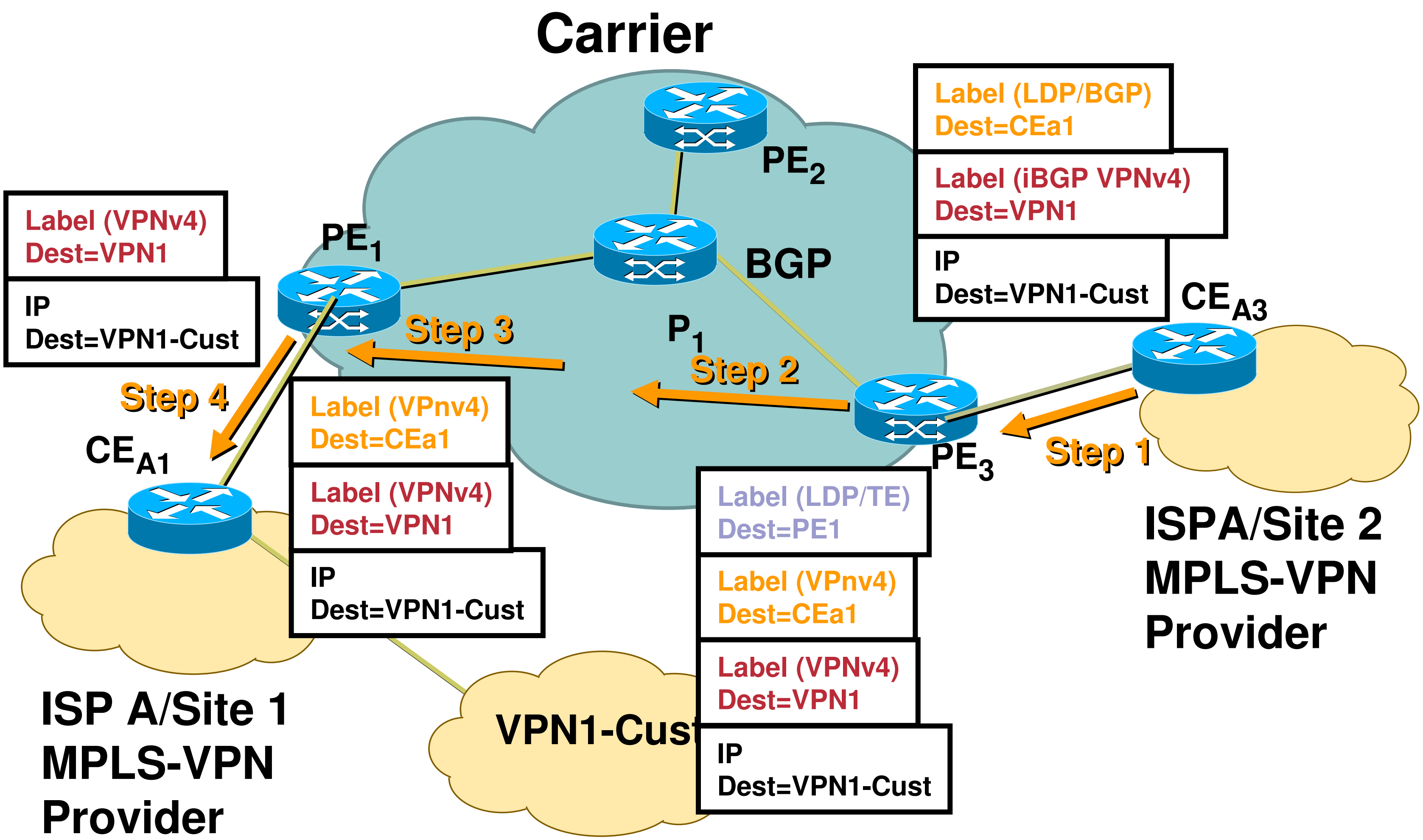


Carrier's Carrier: The Solution (Internet)



Carrier's Carrier: The Solution (VPN)

Cisco.com



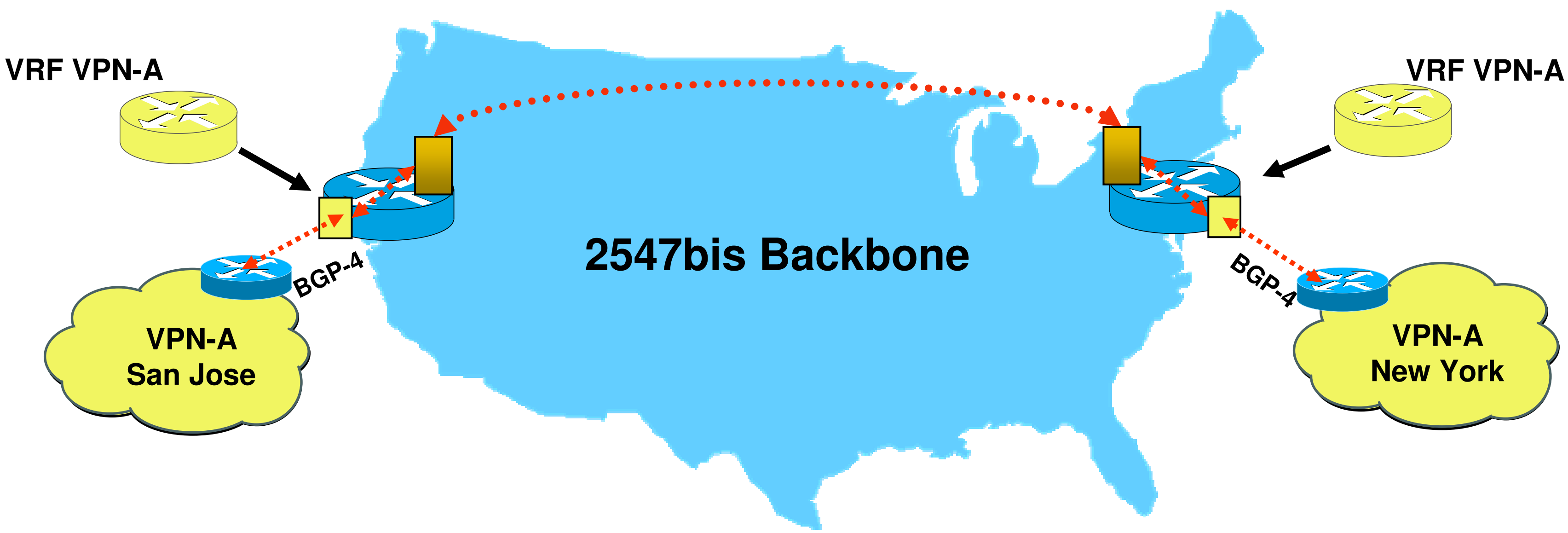
2547 Intra-AS Connectivity Model

- A VPN is a collection of sites sharing common routing information
  - same set of routes within the RIB/FIB
- A site may obtain Intranet or Extranet connectivity
  - through sharing of routing information
- A VPN can be thought of as a Closed User Group (CUG) or community of interest
- Layer-3 forwarding between VPN sites

Distribution of local routing information

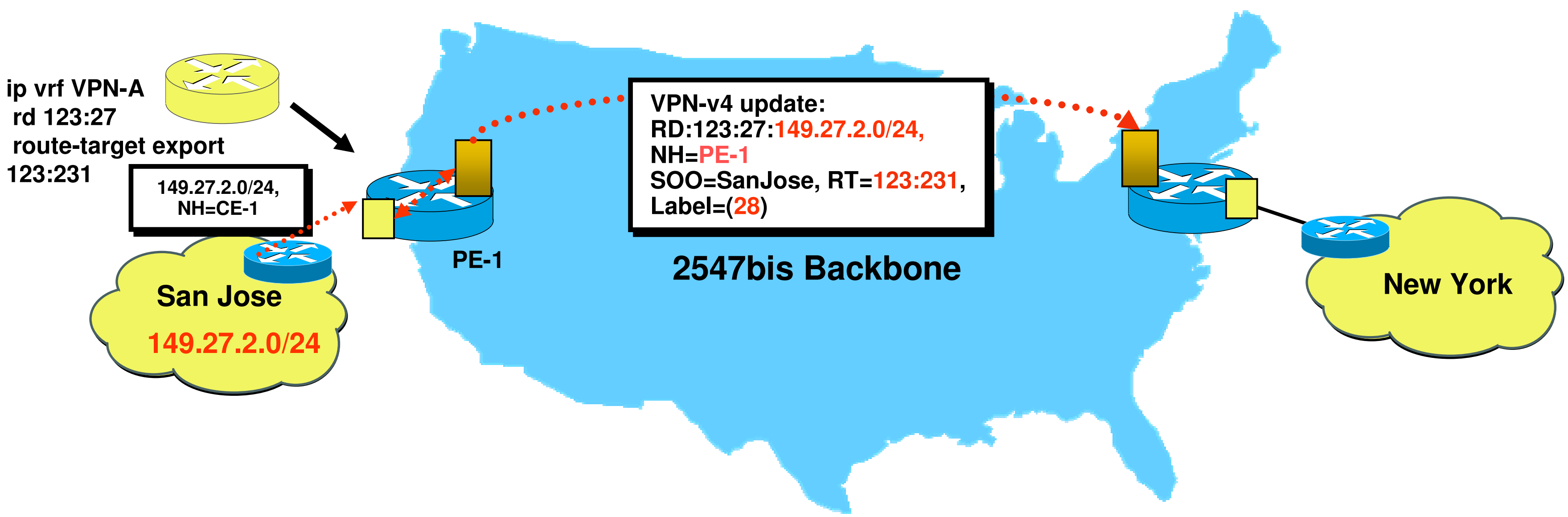
Cisco.com

- **PE routers distribute local VPN information across the 2547 backbone**
  - **through the use of MP-BGP & redistribution from VRFs**
  - **receiving PE imports routes into attached VRFs**



VRF Population of MP-BGP

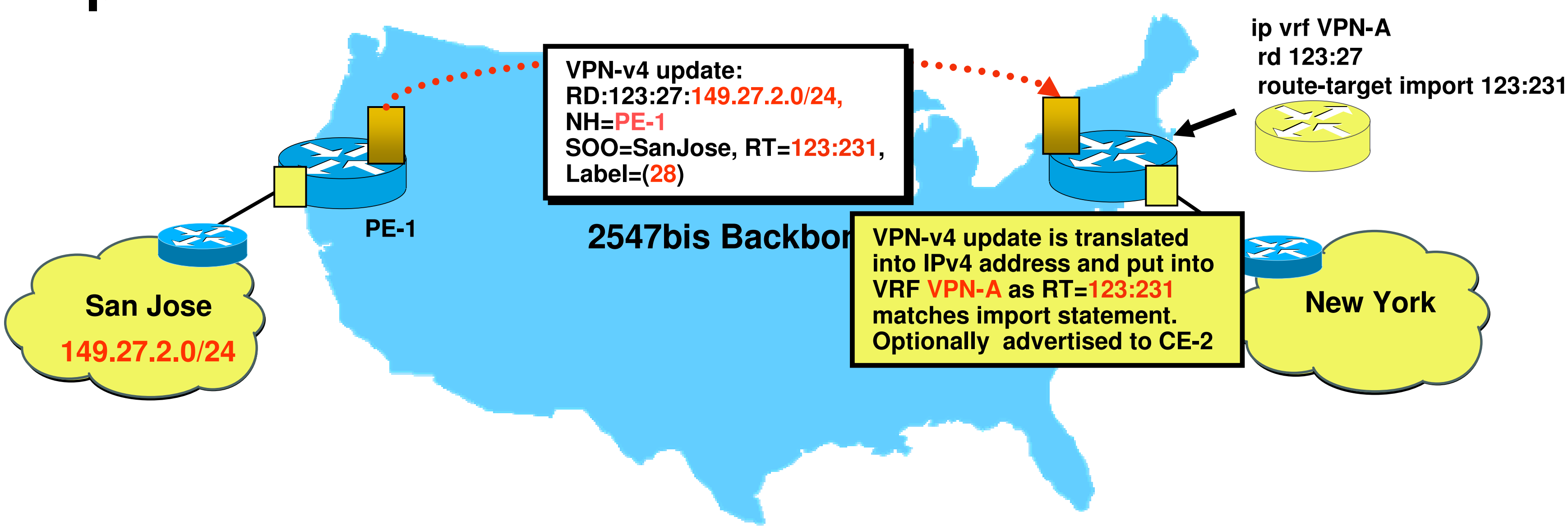
- **PE routers translate into VPNv4 routes**
  - Assign RD, SOO & RT based on configuration
  - Re-write next-hop attribute
  - Assign label based on <u>prefix, VRF and/or interface</u>
  - Send MP-BGP update to all MP-BGP peers



MP-BGP Updated Processing

Cisco.com

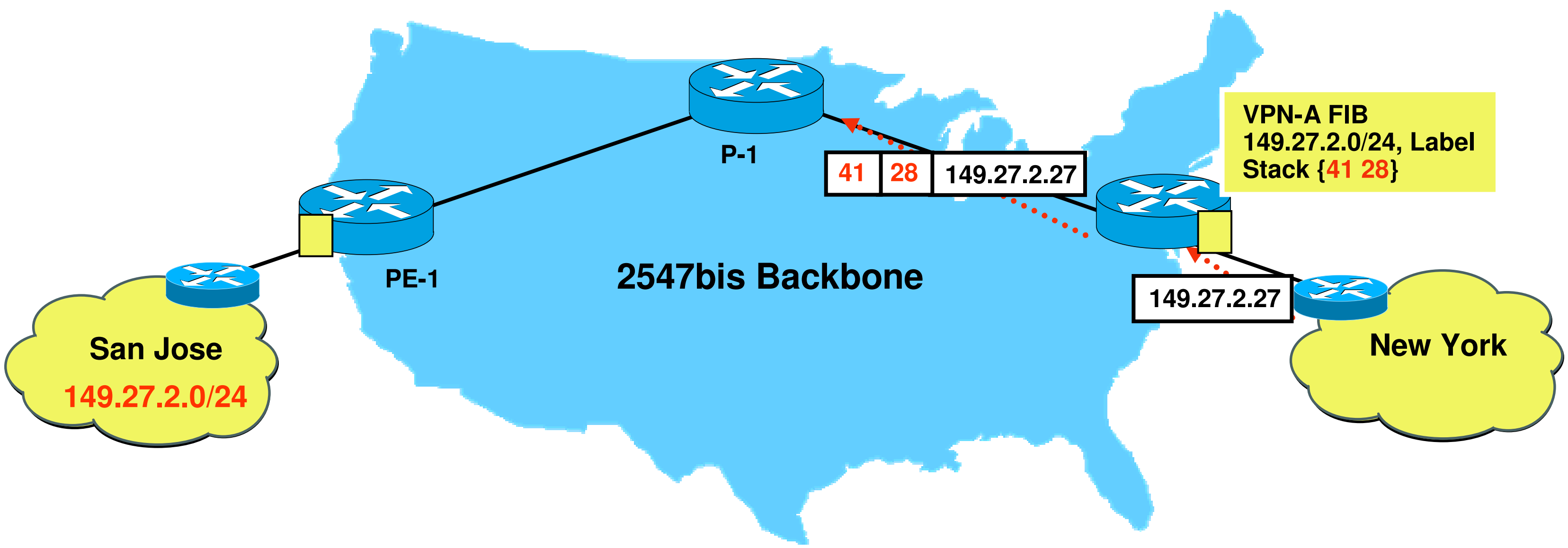
- **Receiving PE routers translate to IPv4 prefix**
  - Inserts routes into relevant VRFs identified by Route target extended-community attribute
- **Label associated with VPNv4 prefix now set on packets forwarded towards the destination**



Ingress PE Label Imposition

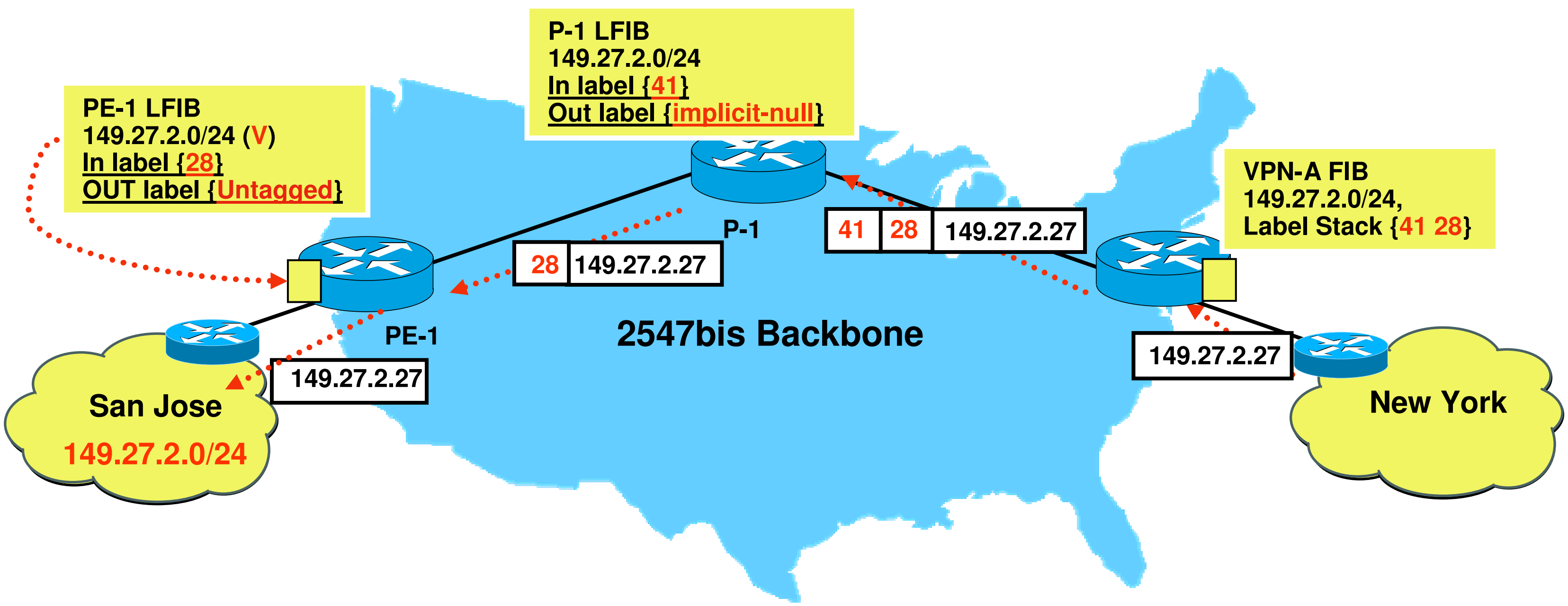
Cisco.com

- Ingress PE receives normal IPv4 packets
- PE router performs IP longest match from VPN VRF, finds BGP next-hop and imposes label stack <IGP, VPN>



Egress PE Label Disposition

- **Penultimate hop router removes top label**
- **Egress PE router uses VPN label to select outgoing interface, label is removed & <u>IP packet</u> is forwarded**



VPN Connectivity between AS#s

- **VPN sites may be geographically dispersed**

  Requiring connectivity to multiple providers, or different regions of the same provider

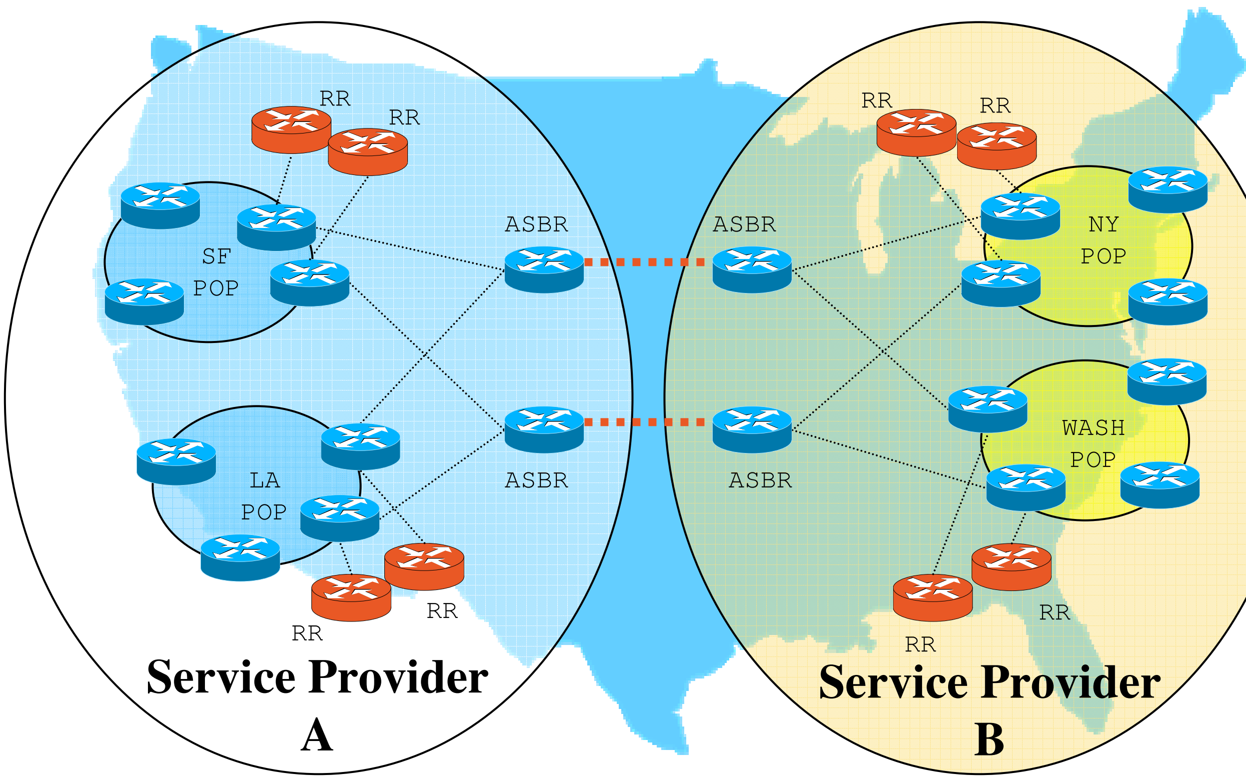
- **Transit traffic between VPN sites may pass through multiple AS#s**

  This implies that routing information MUST be exchanged across AS#s

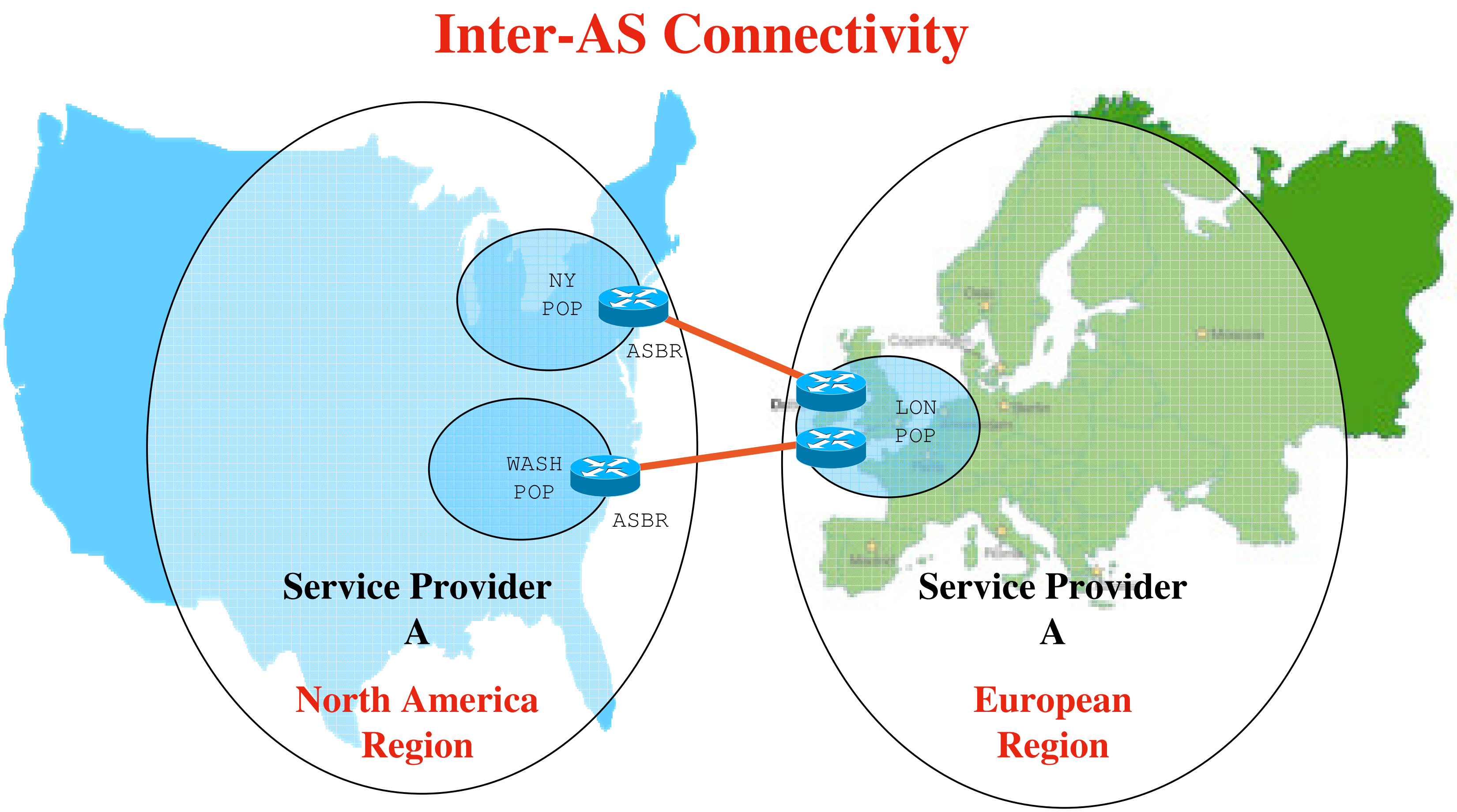
- **Distinction drawn between Inter-Provider & Inter-AS**

Inter-Provider Vs. Inter-AS

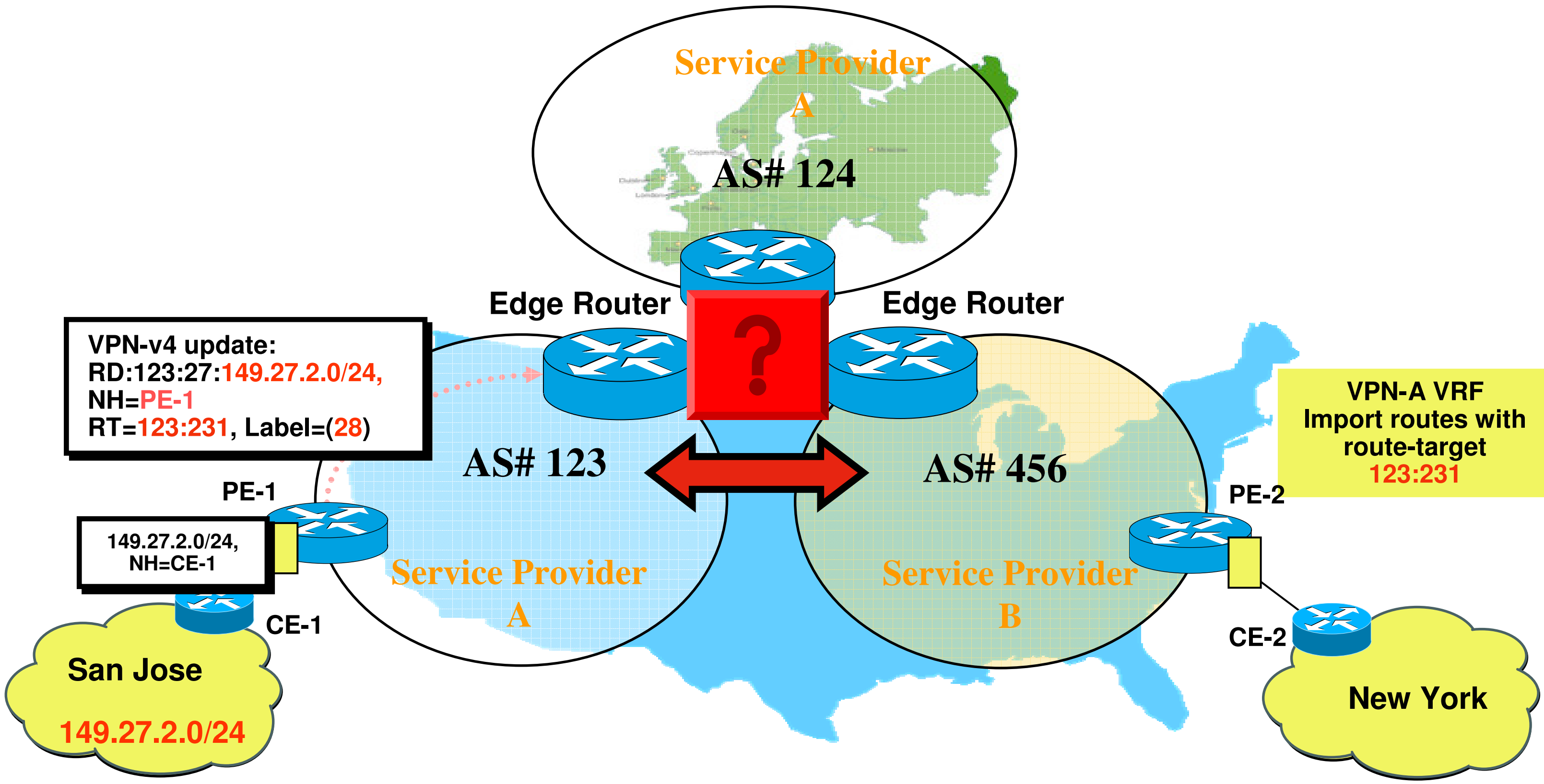
Inter-Provider Connectivity



Inter-Provider Vs Inter-AS



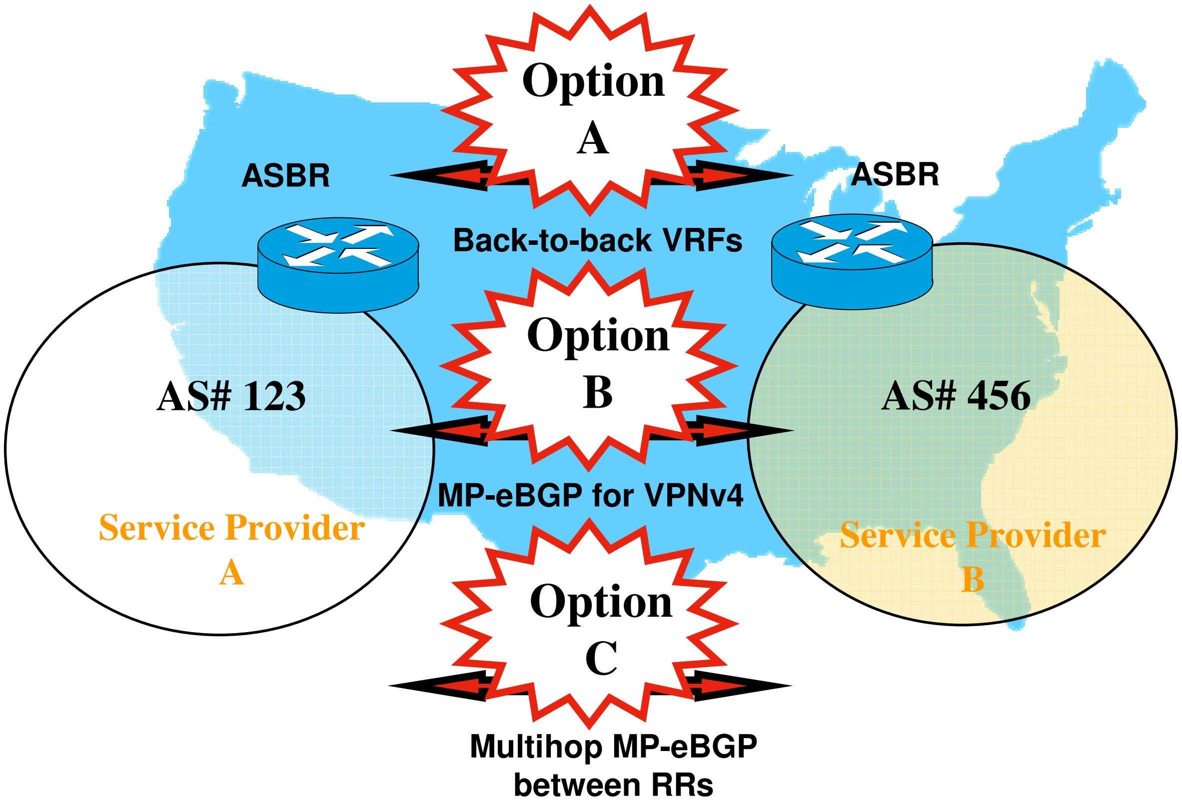
VPN Route Distribution



How to distribute VPNv4 routes between different AS's ?

VPN Route Distribution Options

Cisco.com



Several options available for route distribution

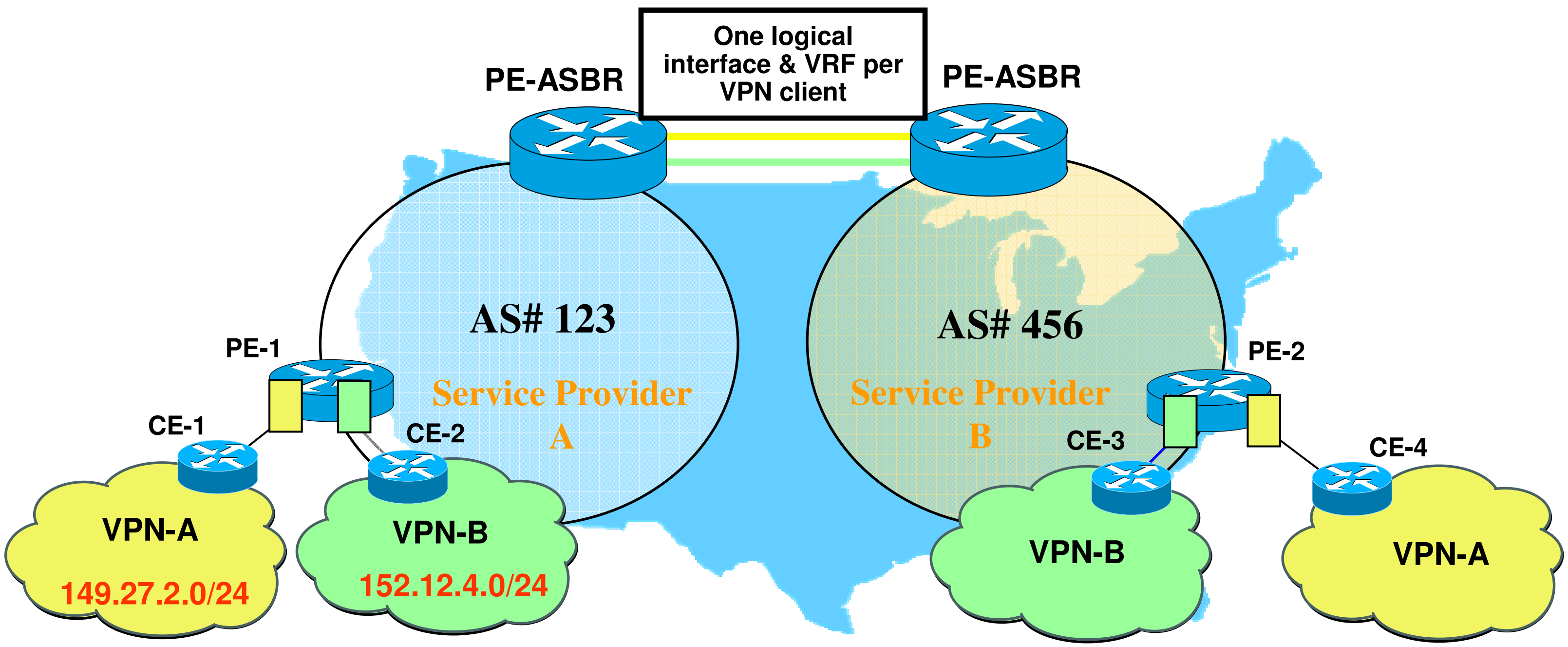
Option A – Back-to-back VRFs

Cisco.com

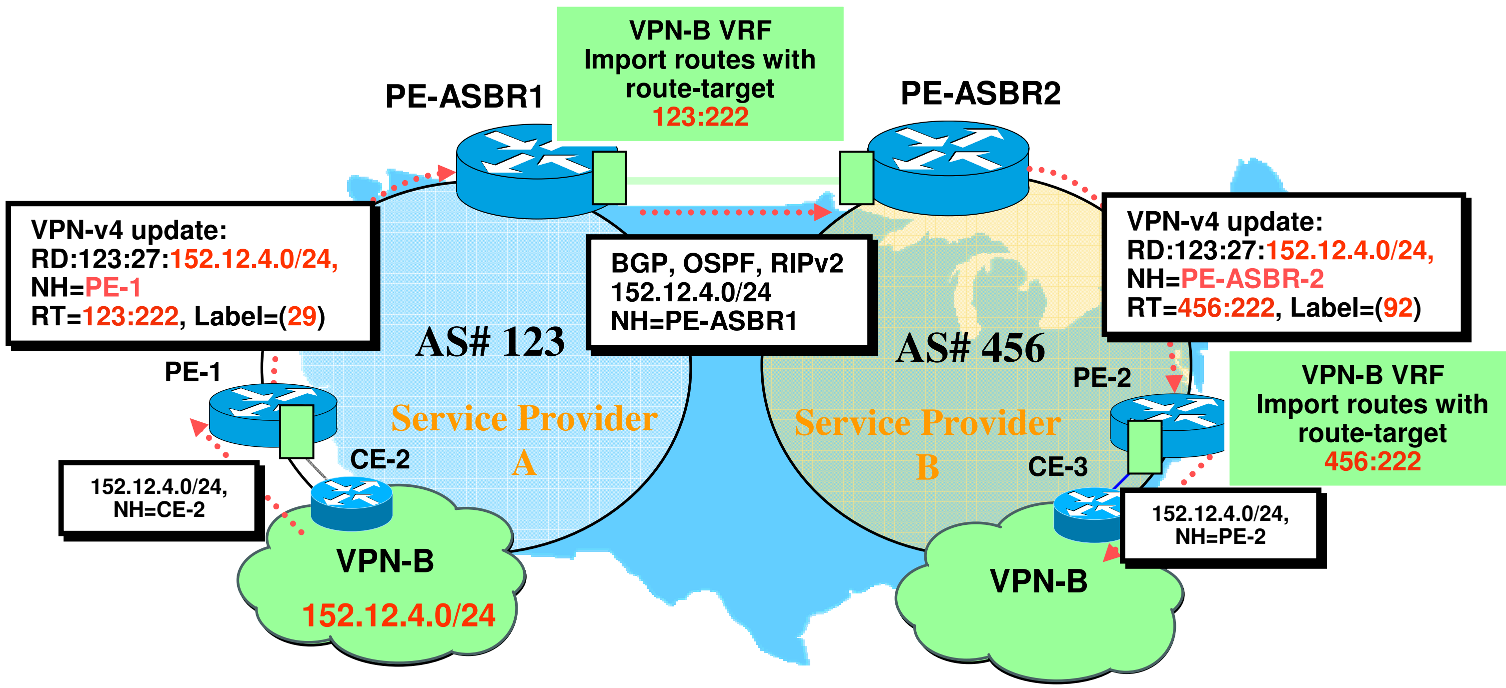
- 2547 providers exchange routes between ASBRs over VRF interfaces
  - Hence ASBR is known as a PE-ASBR
- Each PE-ASBR router treats the other as a CE router
  - Although both provider interfaces are associated with a VRF
- Provider edge routers are gateways used for VPNv4 route exchange
- PE-ASBR link may use any PE-CE routing protocol

Back-to-back VRF Connectivity Model

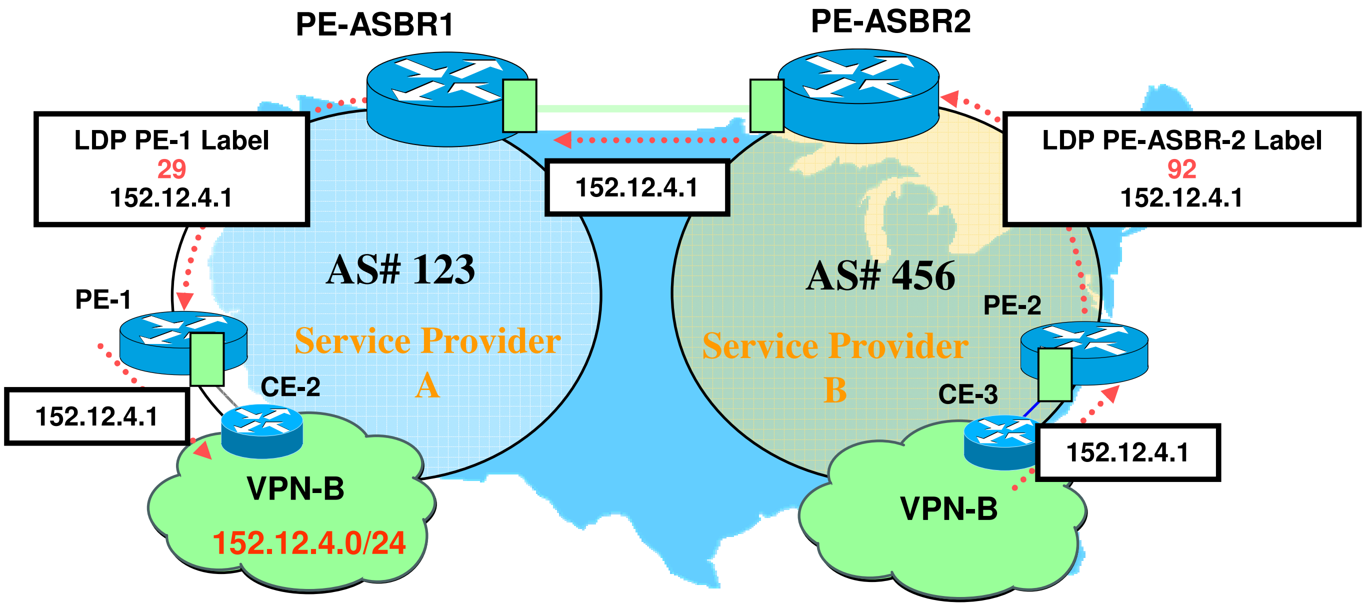
Cisco.com



Back-to-back Prefix Distribution



Back-to-back Packet Flow



Back-to-back VRFs Summary

- **Scalability is an issue with many VPNs**

  1 VRF & logical interface per VPN

  Gateway PE-ASBR must hold ALL routing information

- **PE-ASBR must filter & store VPNv4 prefixes**

- **No MPLS label switching required between providers**

  Standard IP between gateway PE-ASBRs

  No exchange of routes using External MP-BGP

  Simple deployment but limited in scope

  However, everything just works

Option B – External MP-BGP

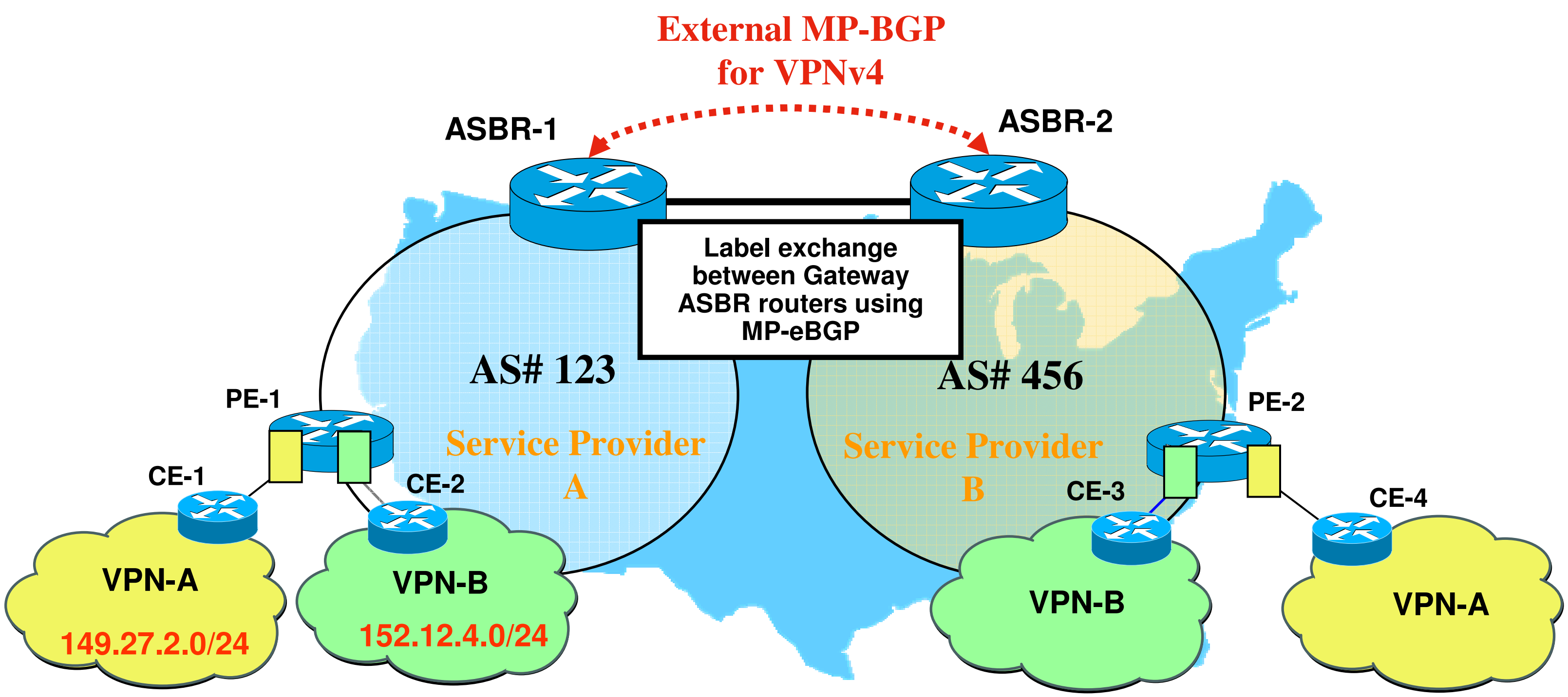
- **Gateway ASBRs exchange VPNv4 routes directly**
  - External MP-BGP for VPNv4 prefix exchange. No LDP/IGP
- **BGP next-hop set to advertising ASBR**
  - Next-hop/labels are rewritten when advertised across ASBR-ASBR link
- **ASBR stores all VPN routes that need to be exchanged**
  - But only within the BGP table. No VRFs. Labels are populated into LFIB at ASBR

Label allocation at receiving PE-ASBR

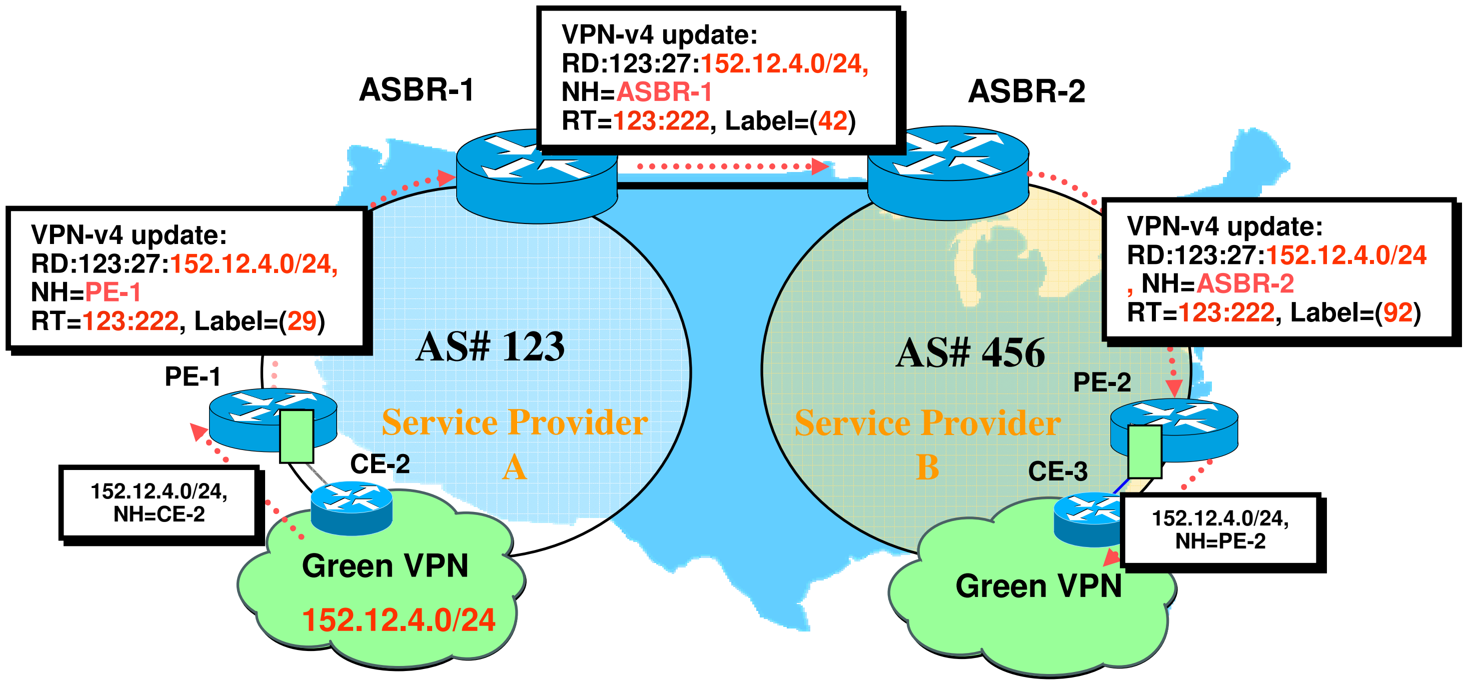
Cisco.com

- **Receiving gateway ASBR may allocate new label**
  - Controlled by configuration of next-hop-self
  - LFIB holds new label allocation
- **Receiving ASBR automatically creates a /32 host route for its ASBR neighbor**
  - Which must be advertised into receiving IGP if next-hop-self is not in operation (to maintain the LSP)

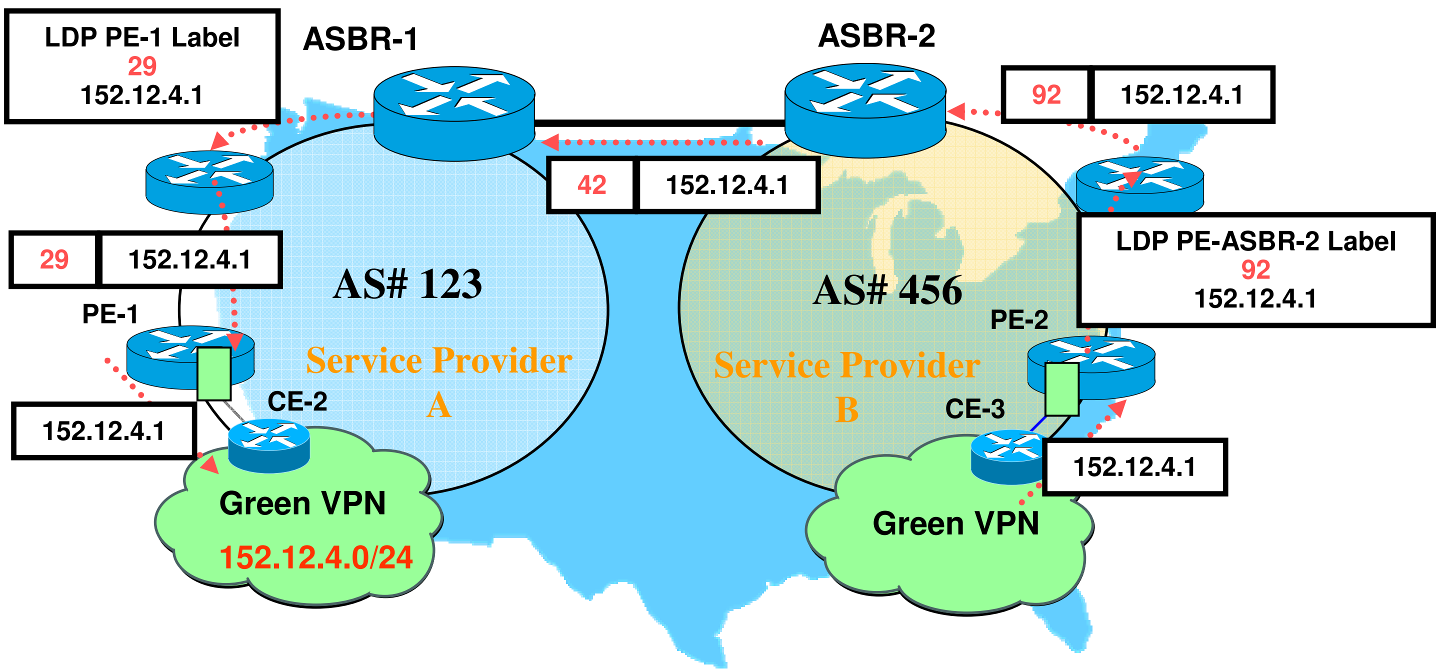
External MP-BGP Connectivity Model



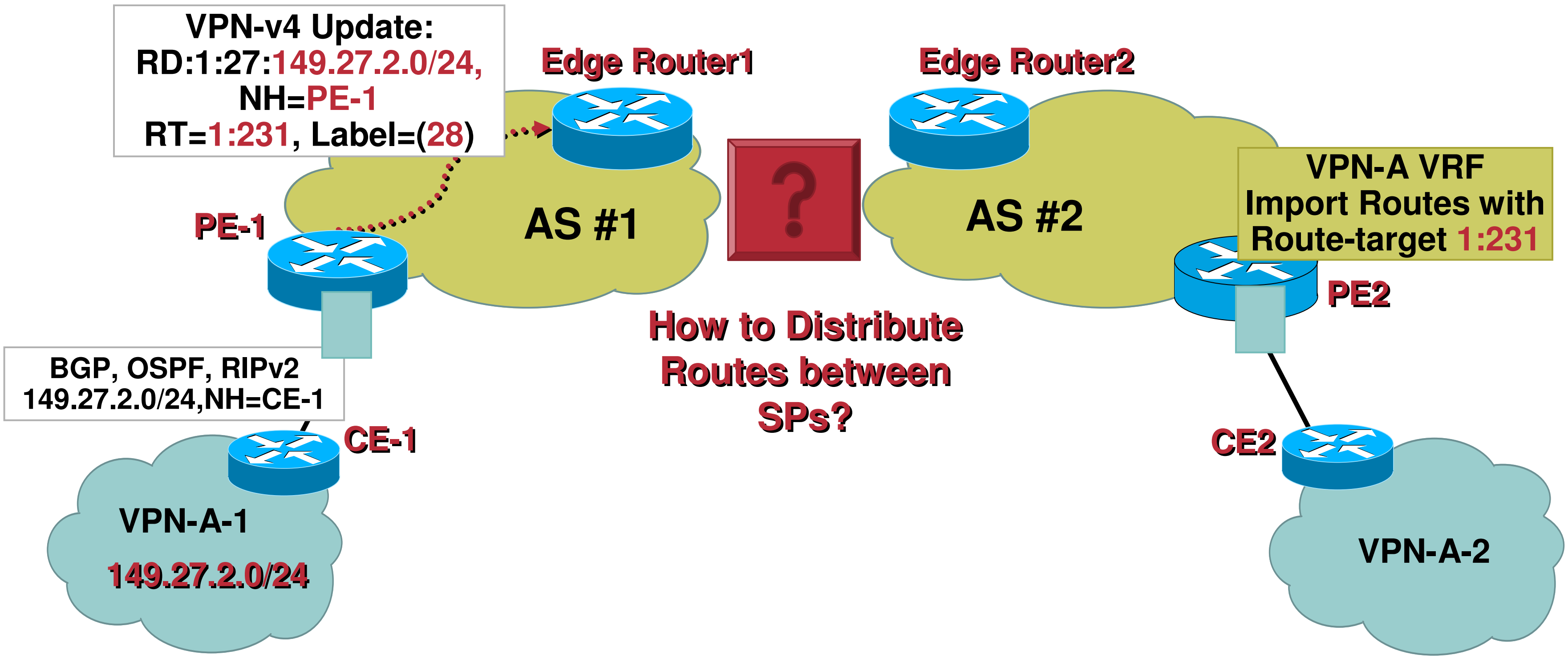
External MP-BGP Prefix Distribution



External MP-BGP Packet Flow



VPN Client Connectivity



VPN Sites Attached to Different MPLS VPN Service Providers

External MP-BGP Summary

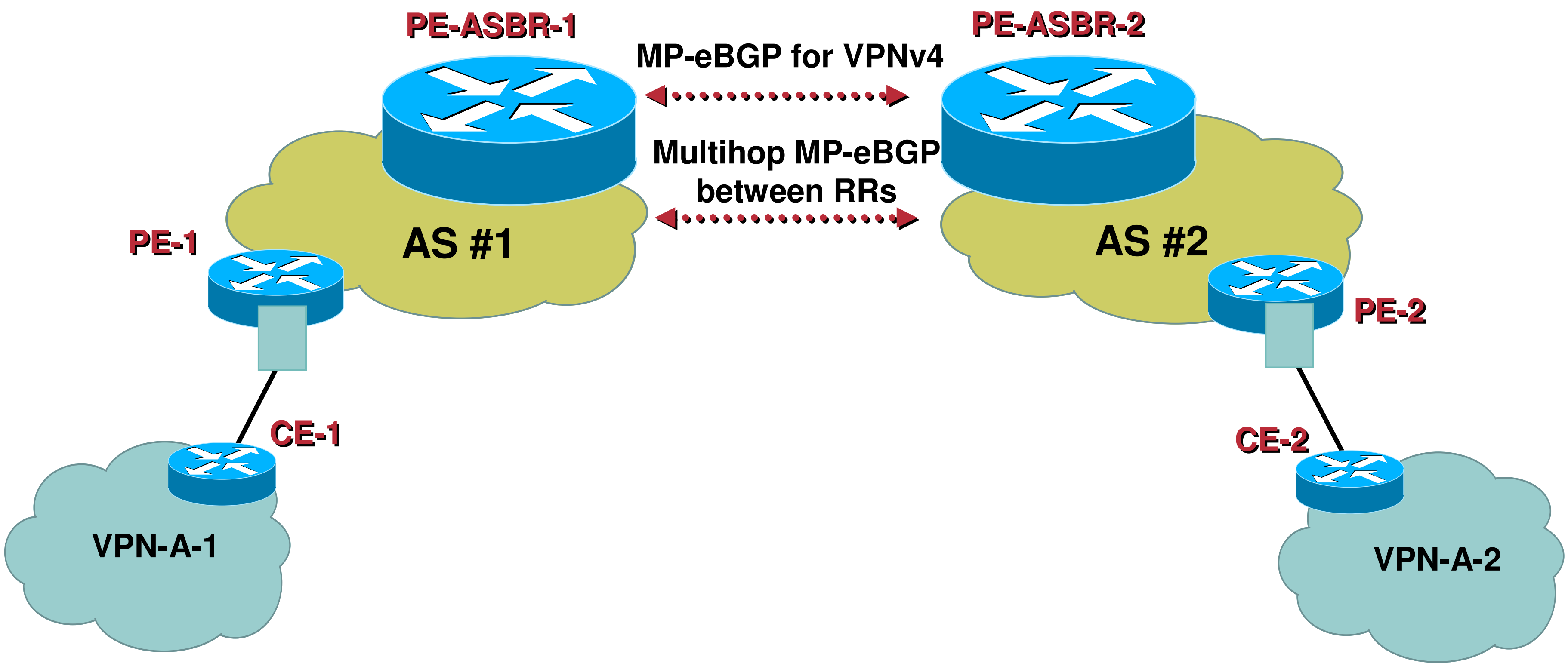
Cisco.com

- **Scalability less of an issue when compared to back-to-back VRF connectivity**
  - Only 1 interface required between ASBR routers
  - No VRF requirement on any ASBR router
- **Automatic route filtering must be disabled**
  - Hence filtering on RT values essential
  - Import of routes into VRFs is NOT required (reduced memory impact)
- **Label switching required between ASBRs**

External MP-BGP Summary (Cont).

- **Preferred option for Inter-Provider connectivity**
  - No IP prefix exchange required between providers
  - Security is tighter
  - Peering agreements specify VPN membership

VPNv4 Distribution Options



Other Options Available, These Two Are the Most Sensible

ASBR Router Protection/Filtering

Cisco.com

- **MP-eBGP session is authenticated with MD5**
  - Potentially also IPSec in the data plane
- **Routing updates filtered on ingress based on extended communities**
  - Both from internal RRs and external peerings
  - ORF used between ASBRs and RRs.
  - Maximum-prefix on MP-BGP session
- **Per-interface label space for external facing links to avoid label spoofing**

Option C – Multihop MP-eBGP between RRs

- **2547 providers exchange VPNv4 prefixes via RRs**
  - Requires multihop MP-eBGP session
- **Next-hop-self MUST be disabled on the RRs**
  - Preserves next-hop/label as allocated by originating PE router
- **Providers exchange IPv4 routes with labels between directly connected ASBRs using External BGP**
  - Only PE router BGP next-hop addresses exchanged
  - RFC3107 "Carrying Label Information in BGP-4"

RFC3107 – Carrying labels with BGP-4

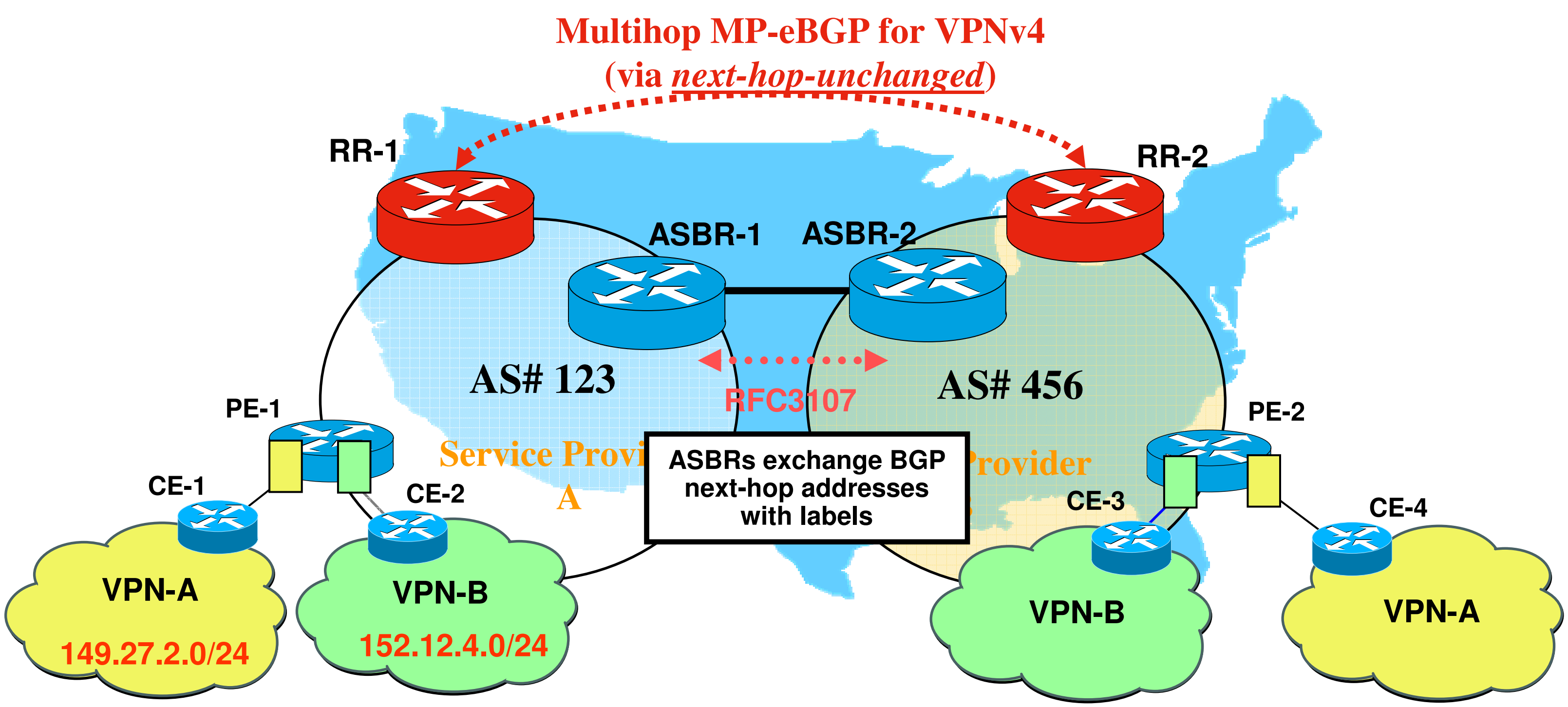

```
0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Address Family Identifier (1) |   SAFI (4)    | Next-hop Lth  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Network Address of next-hop (variable)                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| # of SNPAs    |  Network Layer Reachability Info (variable)   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Length     |                 MPLS Label                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               |             Prefix (variable)                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

MP_REACH_NLRI Attribute
(Specified in RFC 2858)

Prefix plus MPLS label
(Specified in RFC 3107)

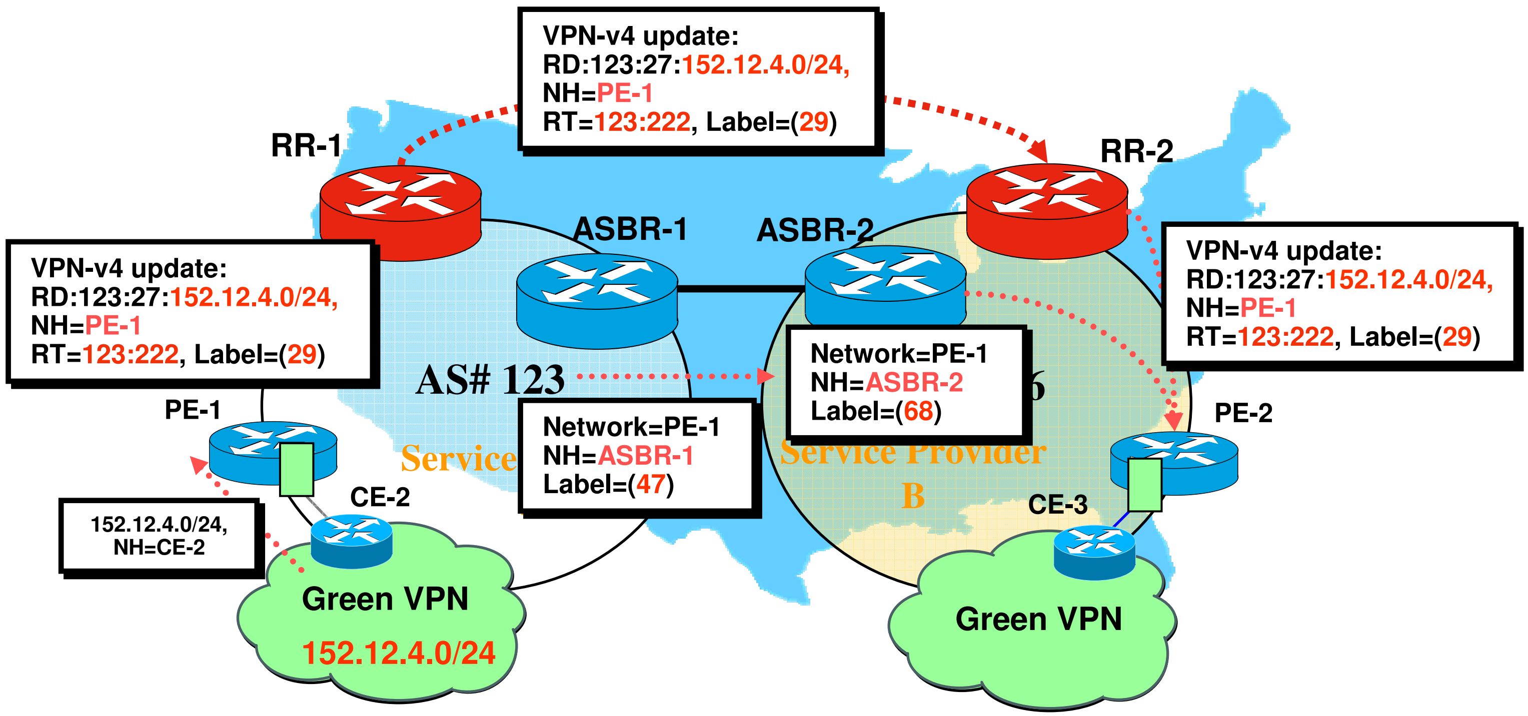
Multihop MP-eBGP Connectivity Model

Cisco.com



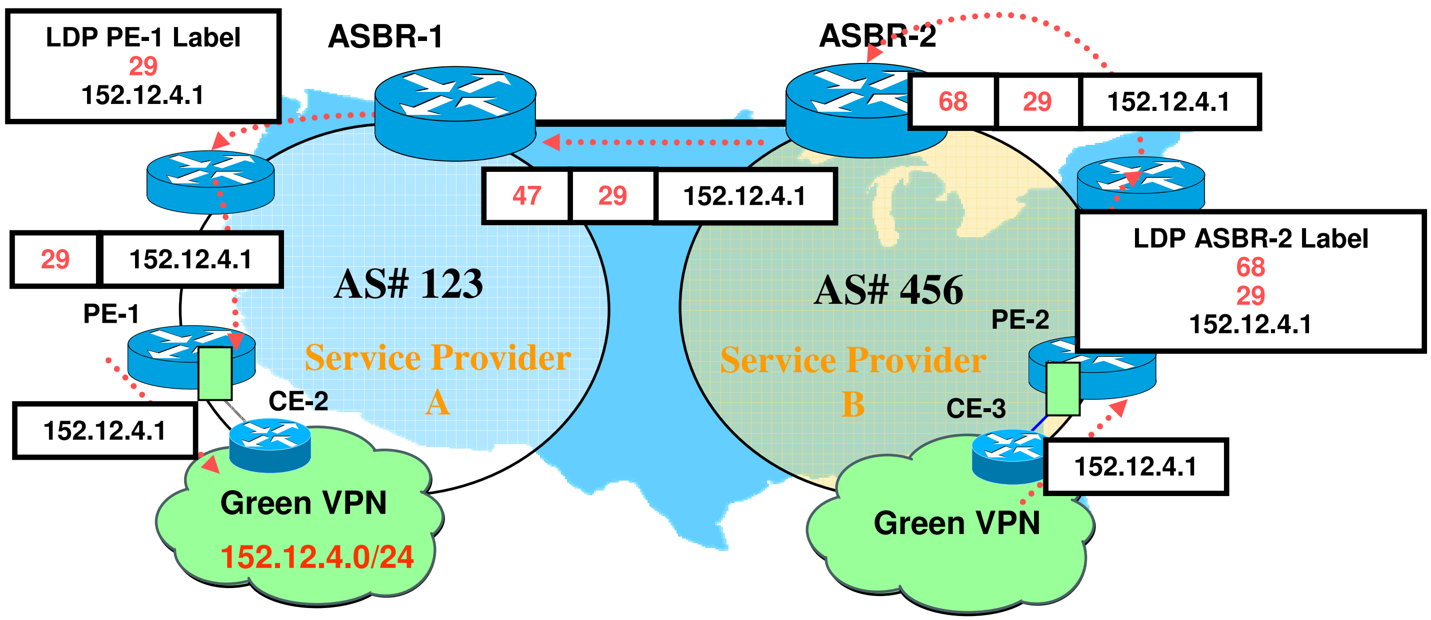
Multihop MP-eBGP Prefix Distribution

Cisco.com



Multihop MP-eBGP Packet Flow

Cisco.com



Multihop MP-eBGP Summary

- **More scalable than previous options**

  As all VPNv4 routes held on route reflectors rather than the ASBRs

- **Route reflectors hold VPNv4 information**

  Each provider utilizes route reflectors locally for VPNv4 prefix distribution

  External BGP connection added for route exchange

- **BGP next-hops across ASBR links using RFC3107**

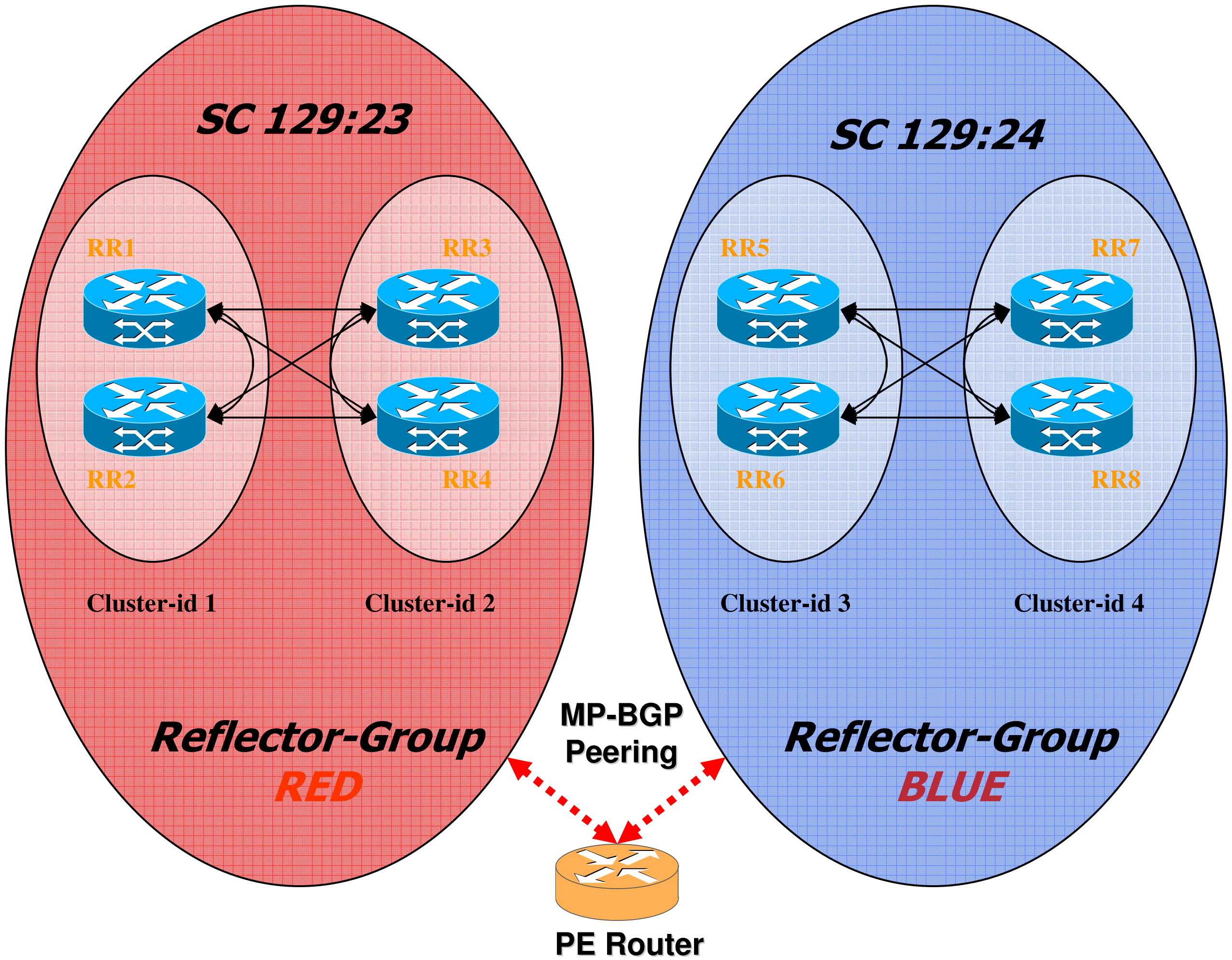
  Separation of forwarding/control planes

ASBR/RR Router Protection/Filtering

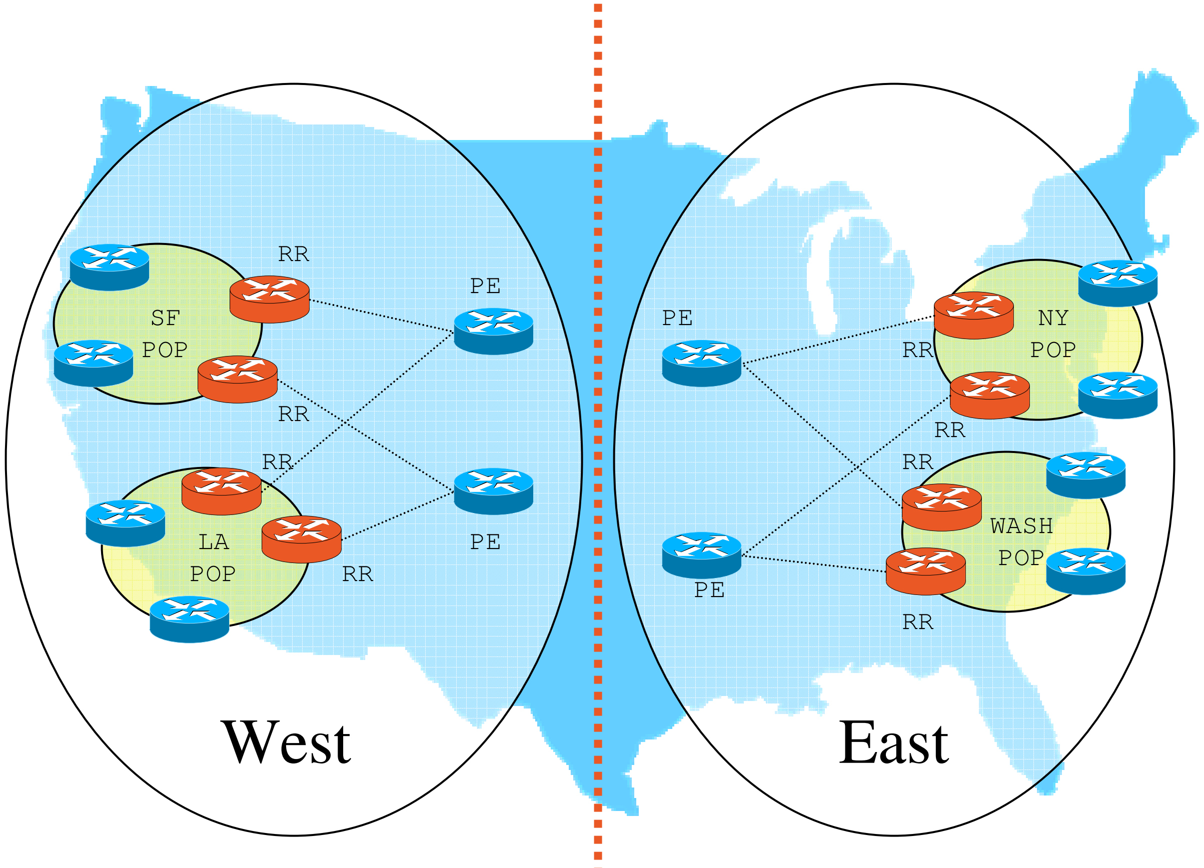
- BGP sessions are authenticated via MD5
  - Both the RFC3107 & MP-BGP sessions
  - Perhaps IPSec authentication in the data plane
- Maximum-prefix deployed on both BGP sessions
- ORF between RRs to filter on extended communities

Distribution of VPNv4 Prefix Information

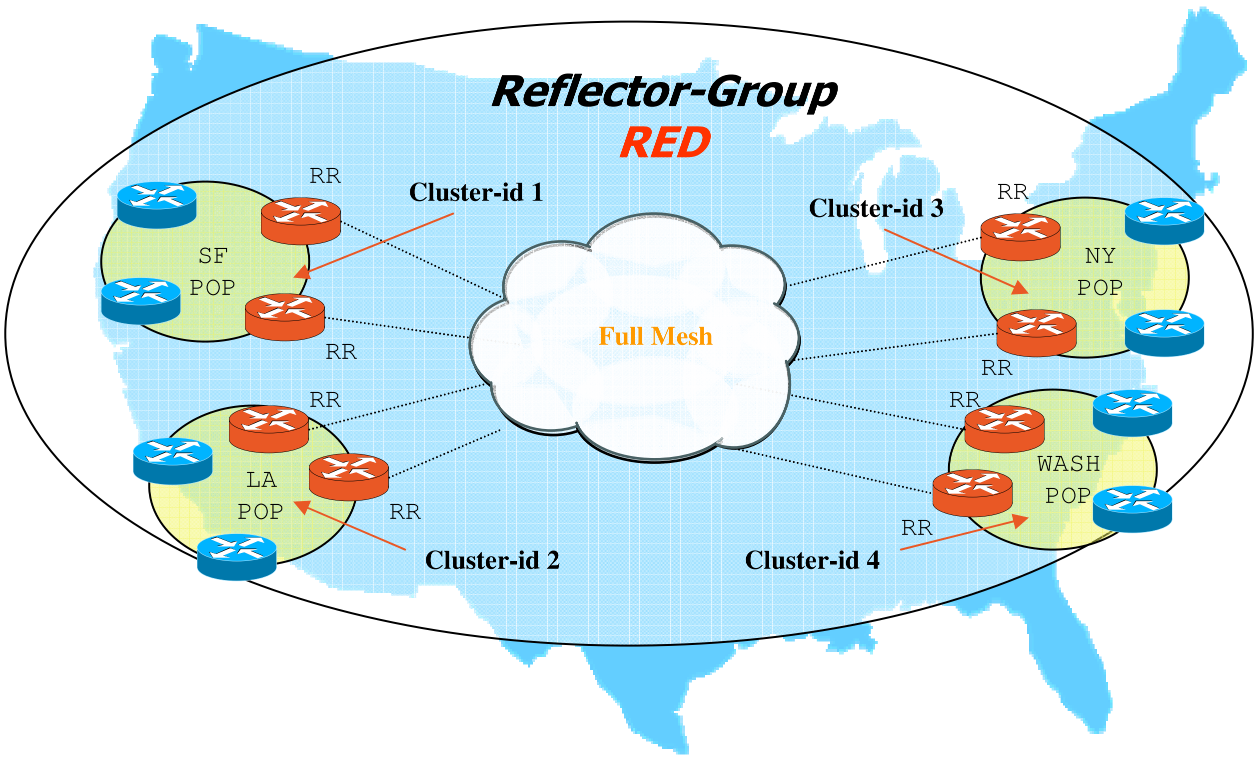
Cisco.com



Route-reflector Topology



Route-reflectors with Reflector-groups



Key Features

- **No constraints on addressing plans used by VPNs— a VPN customer may:**

  Use globally unique and routable/non-routable addresses,

  Use private addresses (RFC1918)

- **Security:**

  Basic security is comparable to that provided by FR/ATM-based VPNs without providing data encryption

  VPN customer may still use IPSec-based mechanisms

    e.g., CE- CE IPSec-based encryption

Key Features (Cont.)

Cisco.com

- **Quality of Service:**

  Flexible and scaleable support for a CoS-based networks

- **Scalability:**

  Total capacity of the system isn’t bounded by the capacity of an individual component

  Scale to virtually unlimited number of VPNs per VPN Service Provider and scale to thousands of sites per VPN

Key Features (Cont.)

- **Connectivity to the Internet:**

  VPN Service Provider may also provide connectivity to the Internet to its VPN customers

  Common infrastructure is used for both VPN and the Internet connectivity services

- **Simplifies operations and management for VPN Service Providers:**

  No need for VPN Service Providers to set up and manage a separate backbone or "virtual backbone" for each VPN

BGP/MPLS VPN—Summary

- Supports large scale VPN service
- Increases value add by the VPN Service Provider
- Decreases Service Provider cost of providing VPN services
- Mechanisms are general enough to enable VPN Service Provider to support a wide range of VPN customers

Deployment/Architecture Challenges

- **As with all technologies there are challenges**
  - Control-plane Scale
  - Filtering & route distribution
  - Security
  - Multicast
  - QOS/End-to-end SLA's
  - Integration of services e.g. Layer-2/Layer-3
  - Network Management
  - Traffic Engineering

MPLS Traffic Engineering

Azhar Sayeed

What Is MPLS Traffic Engineering?

Cisco.com

- Process of routing data traffic in order to balance the traffic load on the various links, routers, and switches in the network
- Key in most networks where multiple parallel or alternate paths are available

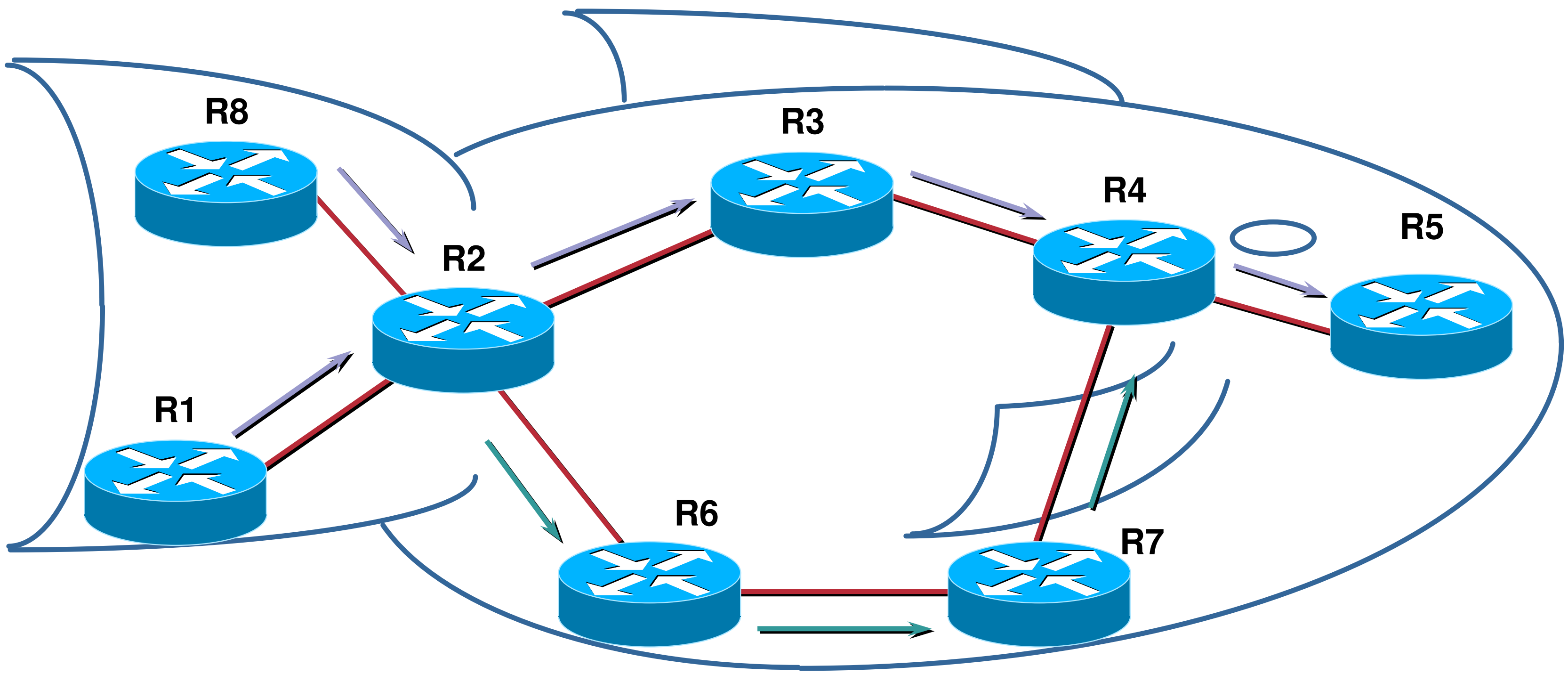
Why Traffic Engineering?

- Congestion in the network due to changing traffic patterns
  - Election news, online trading, major sports events
- Better utilization of available bandwidth
  - Route on the non-shortest path
- Route around failed links/nodes
  - Fast rerouting around failures, transparently to users
  - Like SONET APS (Automatic Protection Switching)
- Build New Services—Virtual leased line services
  - VoIP Toll-Bypass applications, point-to-point bandwidth guarantees
- Capacity planning
  - TE improves aggregate availability of the network

Background – Why Have MPLS-TE?

- IP networks route based only on destination (route)
- ATM/FR networks switch based on both source and destination (PVC, etc)
- Some very large IP networks were built on ATM or FR to take advantage of src/dst routing
- Overlay networks inherently hinder scaling (see "The Fish Problem")
- MPLS-TE lets you do src/dst routing while removing the major scaling limitation of overlay networks
- MPLS-TE has since evolved to do things other than bandwidth optimization

IP Routing and The Fish



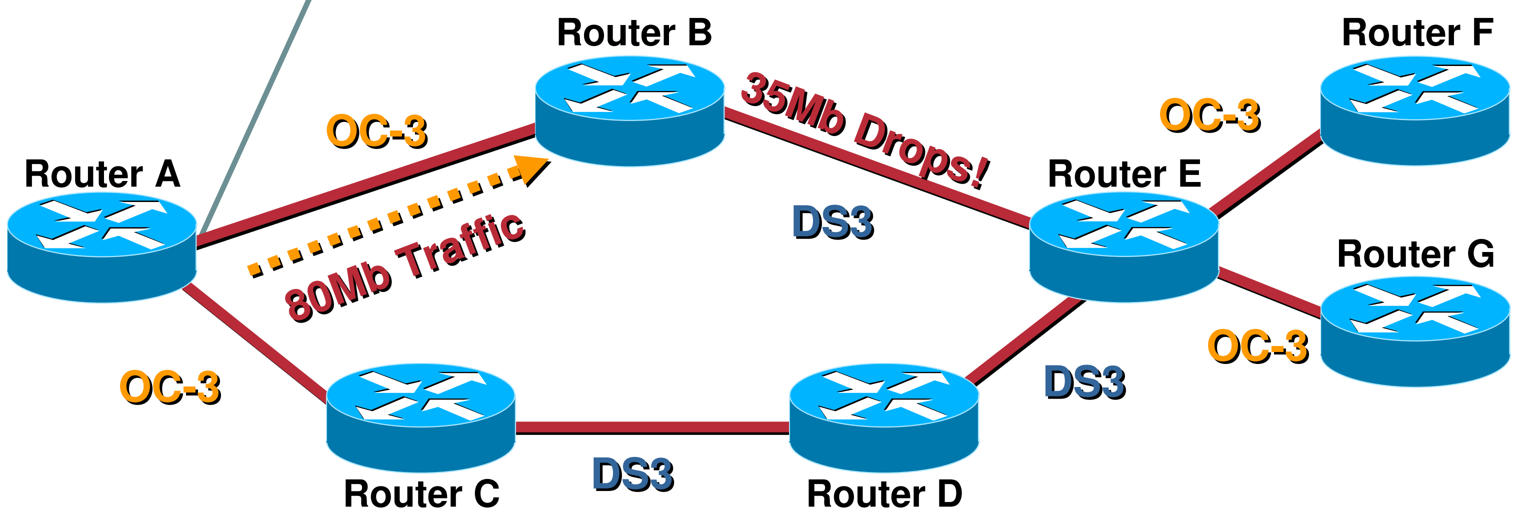
- IP (Mostly) Uses Destination-Based Least-Cost Routing
- Flows from R8 and R1 Merge at R2 and Become Indistinguishable
- From R2, Traffic to R3, R4, R5 Use Upper Route
- Alternate Path Under-Utilized

The Problem with Shortest-Path

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

- Some links are DS3, some are OC-3
- 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

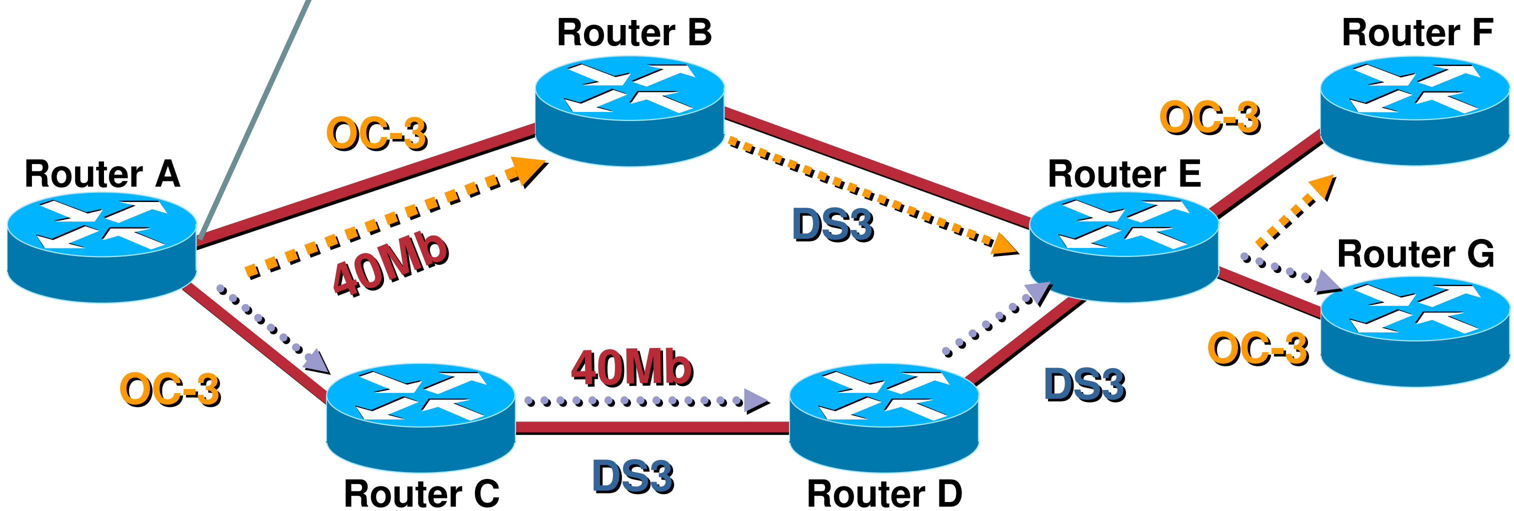


How MPLS TE Solves the Problem

Cisco.com

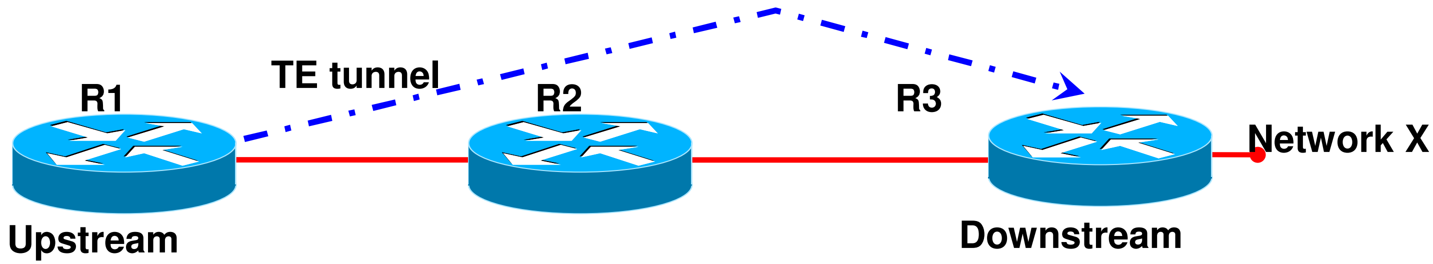
Node	Next-Hop	Cost
B	B	10
C	C	10
D	C	20
E	B	20
F	Tunnel 0	30
G	Tunnel 1	30

- Router A sees all links
- Router A computes paths on properties other than just shortest cost
- No link oversubscribed!



A terminology slide – head, tail, LSP, etc

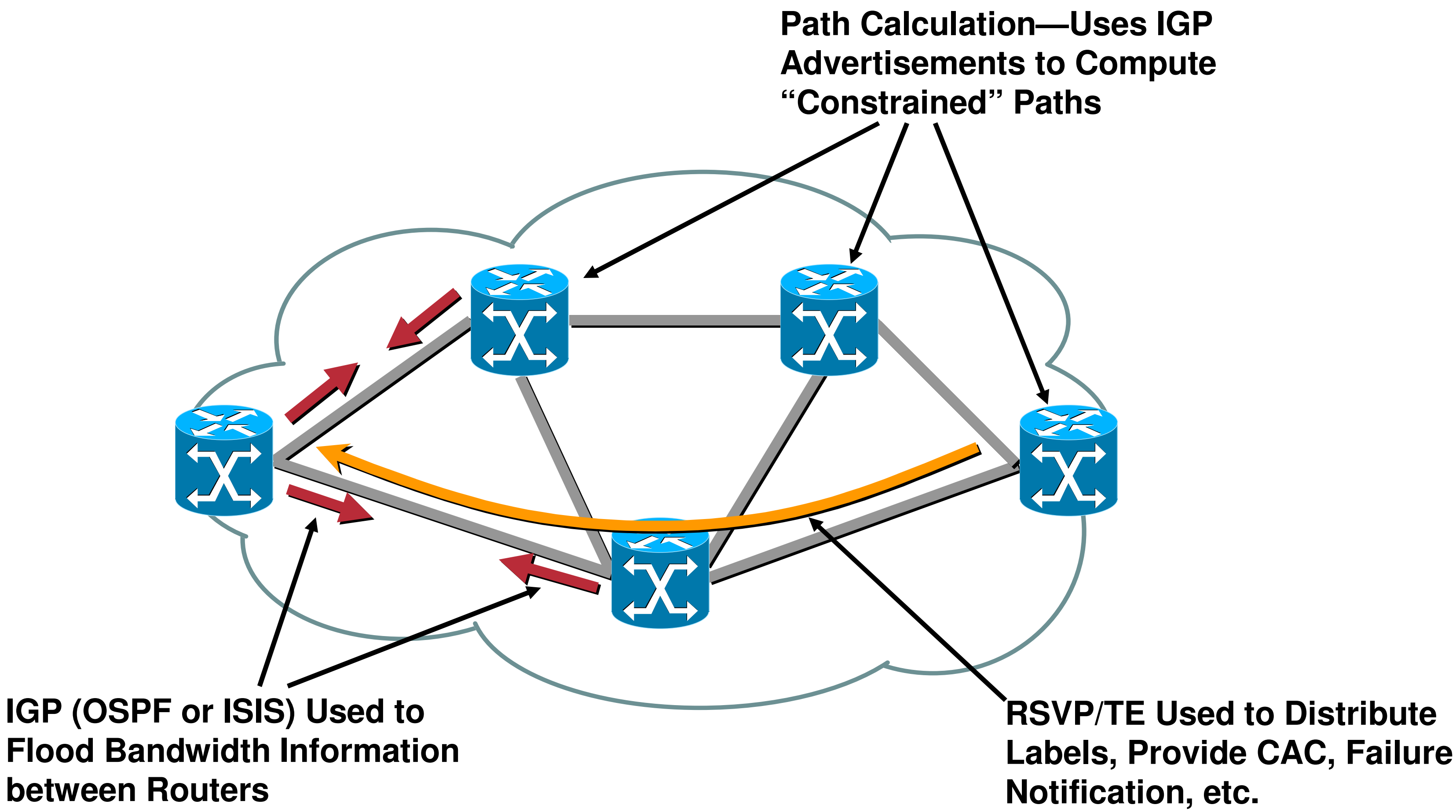
Cisco.com



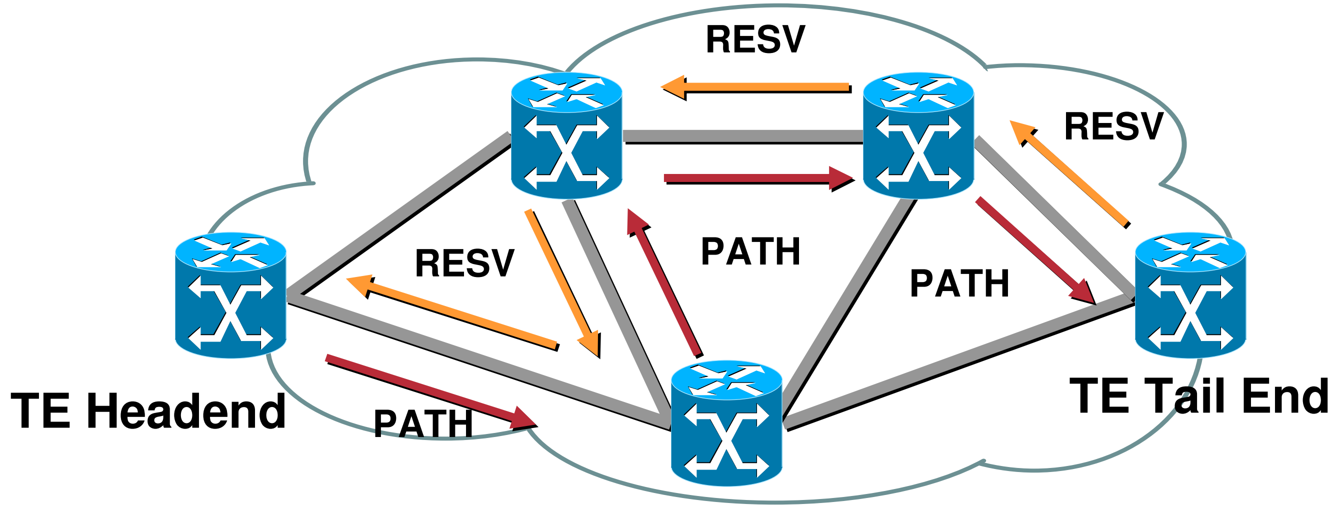
- Head-End is a router on which a TE tunnel is configured (R1)
- Tail-End is the router on which TE tunnel terminates (R3)
- Mid-point is a router thru which the TE tunnel passes (R2)
- LSP is the Label Switched Path taken by the TE tunnel, here R1-R2-R3
- Downstream router is a router closer to the tunnel tail
- Upstream router is farther from the tunnel tail (so R2 is upstream to R3's downstream, R1 is upstream from R2's downstream)

TE Fundamentals—“Building Blocks”

Cisco.com



Example



- PATH messages are sent with requested bandwidth
- RESV messages are sent with label bindings for the TE tunnel
- Tunnels can be explicitly routes
- Admission control at each hop to see if the bandwidth requirement can be met
- Packets are mapped to the tunnel via
    - Static routed
    - Autoroute
    - Policy route
- Packets follow the tunnel—LSP

Traffic Engineering

Theory

- Information Distribution
- Path Calculation
- Path Setup
- Routing Traffic Down A Tunnel

Information Distribution

- You need a link-state protocol as your IGP
  - IS-IS or OSPF
- Link-state requirement is only for MPLS-TE!
  - Not a requirement for VPNs, etc!
- Why do I need a link-state protocol?
  - To make sure info gets flooded
  - To build a picture of the entire network
- Information flooded includes Link, Bandwidth, Attributes, etc.

Information Distribution

- TE LSPs can (optionally) reserve bandwidth across the network
- Reserving bandwidth is one of the ways to find more optimal paths to a destination
- This is a control-plane reservation only
- Need to flood available bandwidth information across the network
- IGP extensions flood this information
  - -OSPF uses Type 10 (area-local) Opaque LSAs
  - -ISIS uses new TLVs
  - -Some other information flooded, not important now

Path Calculation

Cisco.com

- Once available bandwidth information is flooded, router may calculate a path from head to tail.
  - -Path may already be preconfigured on the router, will talk about that later
- TE Headend does a "Constrained SPF" (CSPF) calculation to find the best path
- CSPF is just like regular IGP SPF, except
  - -Takes required bandwidth into account
  - -Looks for best path from a head to a single tail, not to all devices
- N tunnel tails, N CSPFs
- In practice, there has been zero impact from CSPF CPU utilization on even the largest networks

Path Setup

- Once the path is calculated, need to signal it across the network.
- Why? 2 reasons:
  1. Reserve any bandwidth, so that other LSPs can't overload the path
  2. Establish an LSP for loop-free forwarding along an arbitrary path
     - Like ATM VC/FR DLCI
     - See "The Fish Problem", later

Path Setup

- PATH messages = from head to tail

  (think "call setup") carries LABEL_REQUEST

- RESV messages = from tail to head

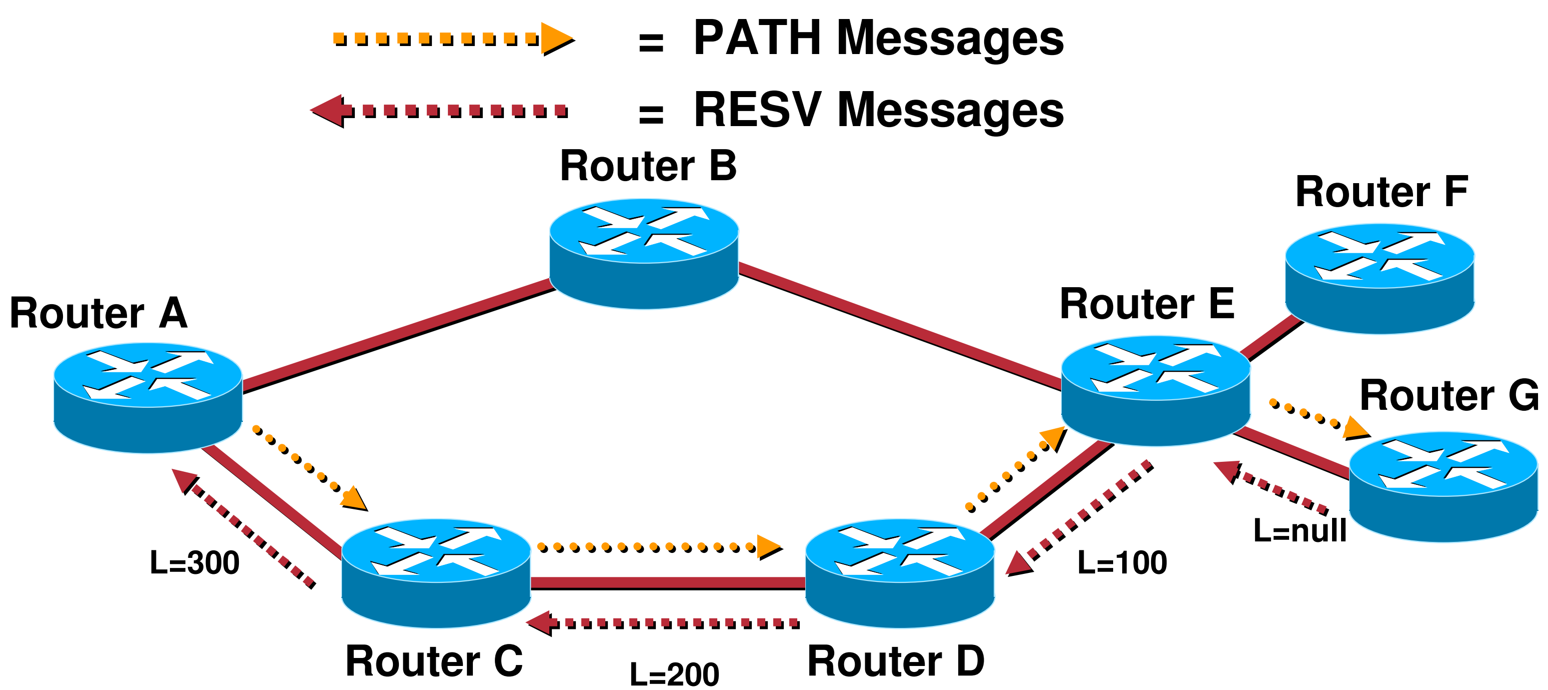
  (think "call ACK") carries LABEL

- Other RSVP message types exist for LSP teardown and error signalling

Path Setup

Cisco.com

- PATH message: “Can I have 40Mb along this path?”
- RESV message: “Yes, and here’s the label to use”
- LFIB is set up along each hop



Path Setup

- **Once RESV reaches headend, tunnel interface comes up**
- **Errors along the way are handled appropriately (tunnel does not come up, message gives point of failure and reason for failure)**

Path Setup

Fundamental points here:

- You can use MPLS-TE to forward traffic down a path other than that determined by your IGP cost
- You can determine these arbitrary paths per tunnel headend

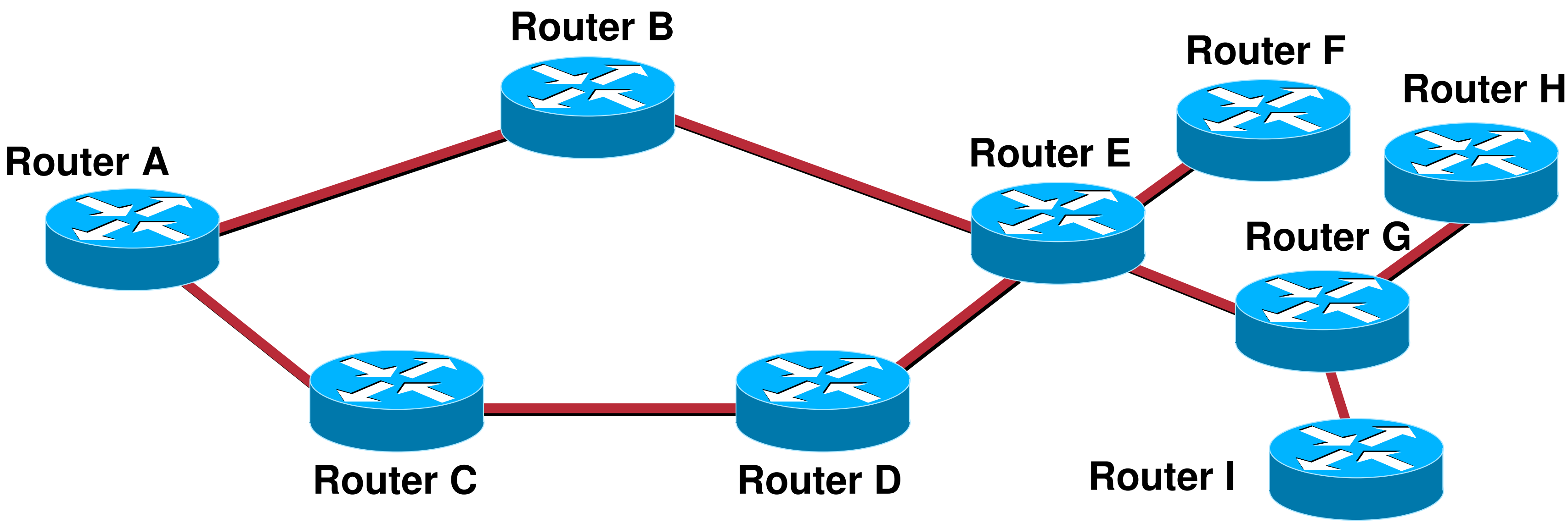
Routing Traffic Down A Tunnel

- Once RESV reaches headend, tunnel interface comes up
- How to get traffic down the tunnel?
  1. Autoroute
  2. Forwarding adjacency
  3. Static routes
  4. Policy routing

Autoroute

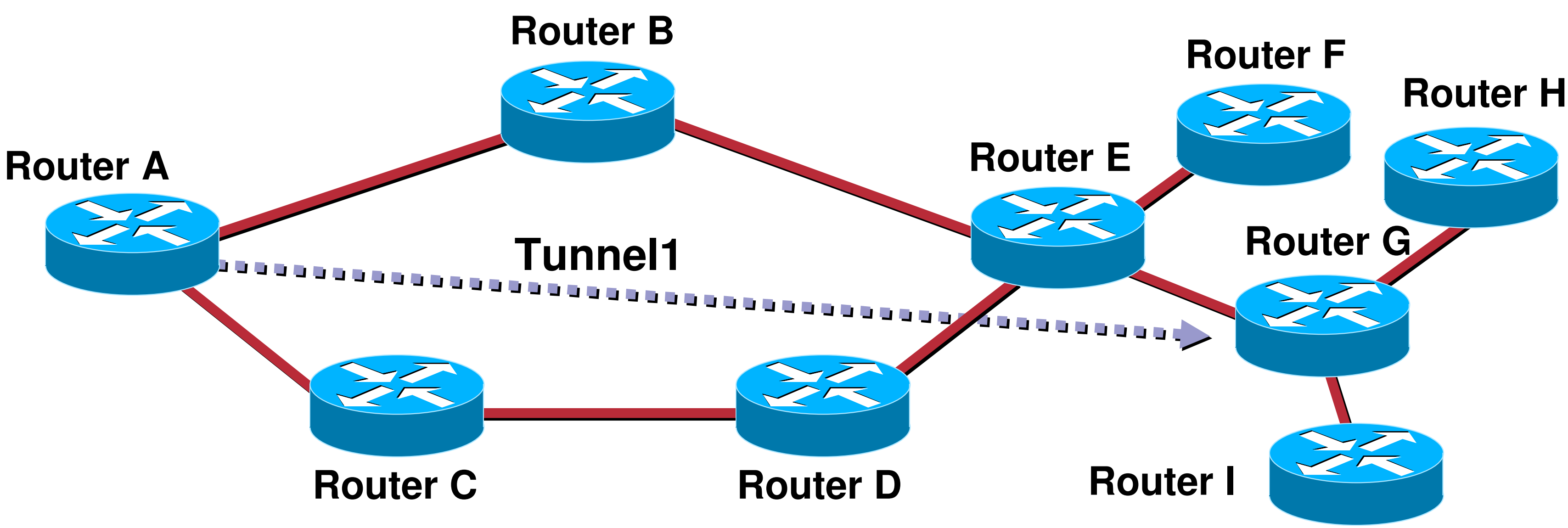
- Tunnel is treated as a directly connected link to the tail
- IGP adjacency is NOT run over the tunnel!
  - Unlike an ATM/FR VC
- Autoroute limited to single area/level only

This Is the Physical Topology



Autoroute

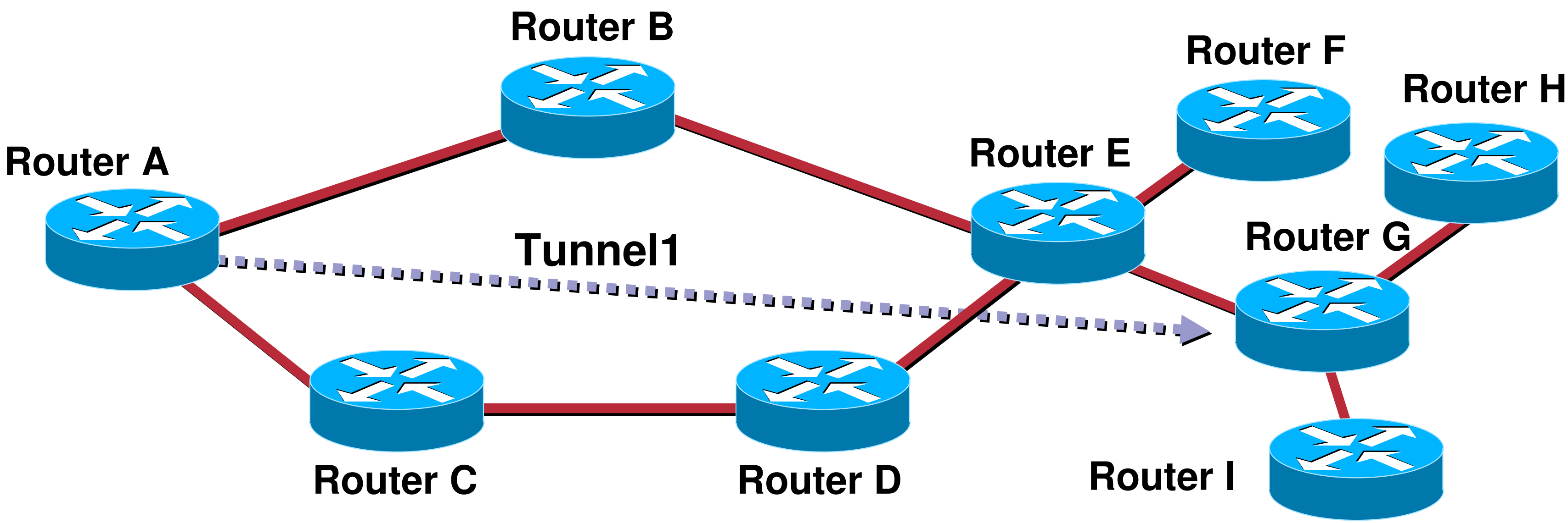
- This is Router A's logical topology
- By default, other routers don't see the tunnel!



Autoroute

Node	Next-Hop	Cost
B	B	10
C	C	10
D	C	20
E	B	20
F	B	30
G	Tunnel 1	30
H	Tunnel 1	40
I	Tunnel 1	40

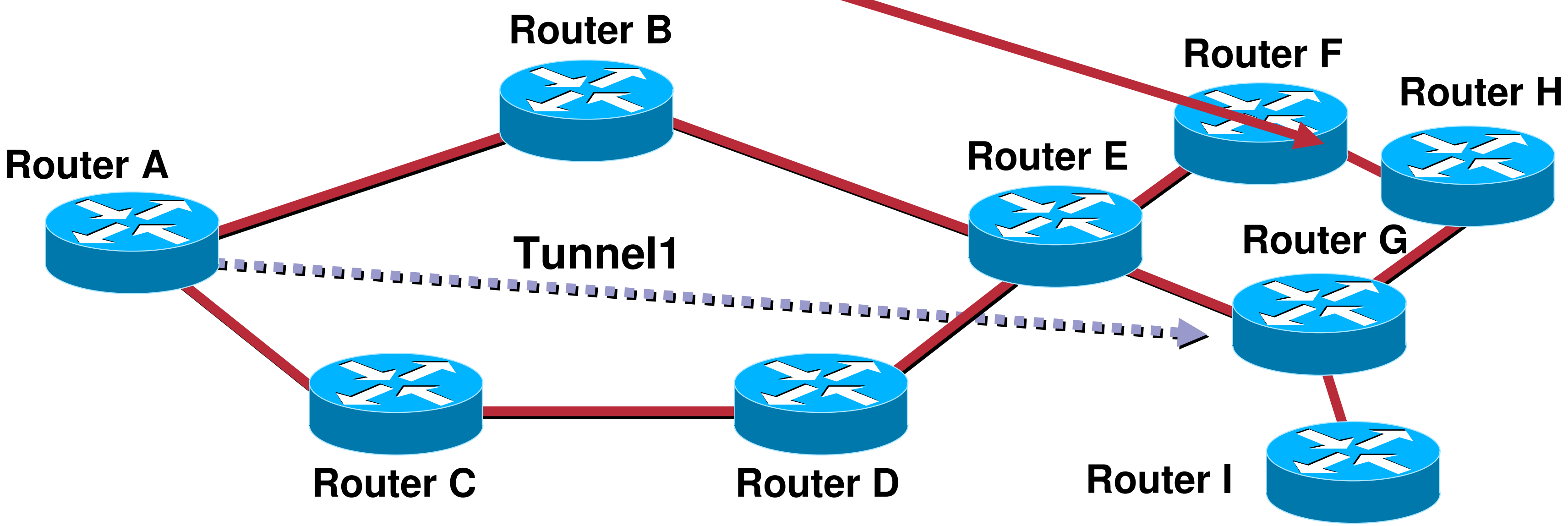
- Router A's routing table, built via auto-route
- Everything "behind" the tunnel is routed via the tunnel



Autoroute

Node	Next-Hop	Cost
B	B	10
C	C	10
D	C	20
E	B	20
F	B	30
G	Tunnel 1	30
H	Tunnel 1 & B	40
I	Tunnel 1	40

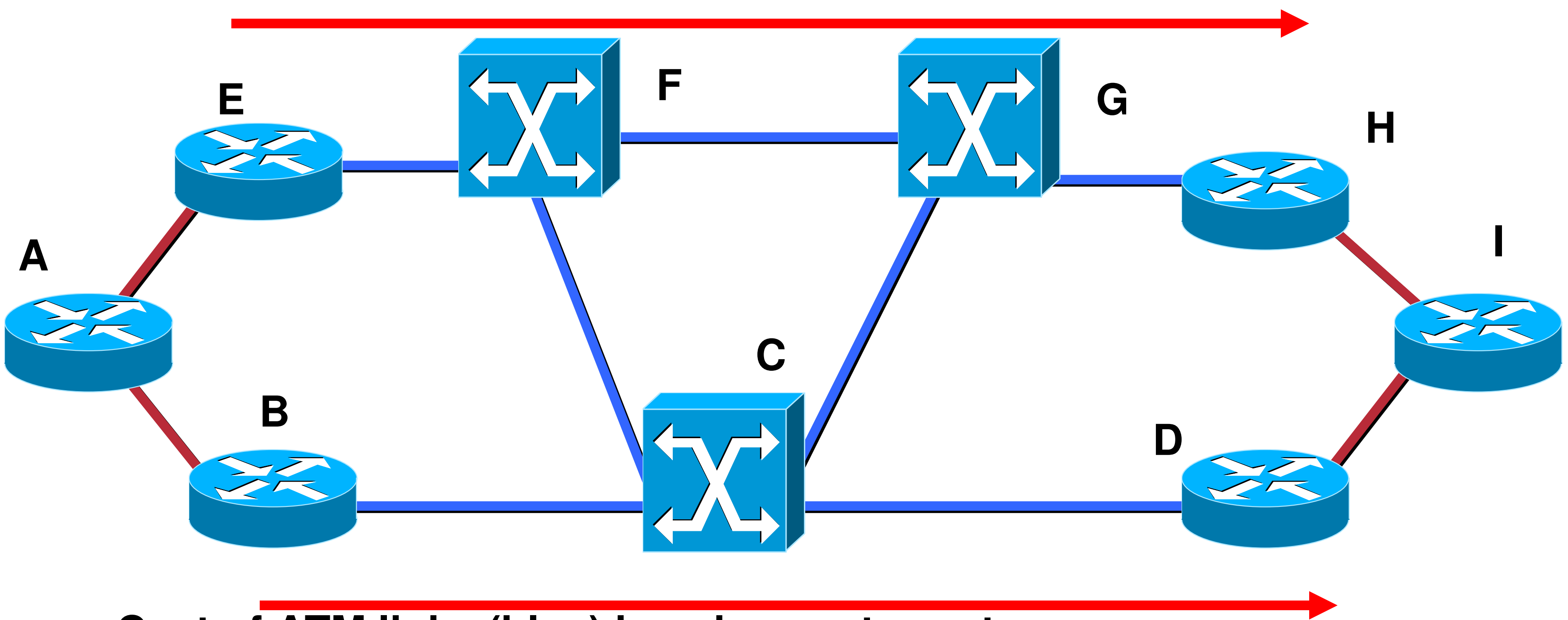
- If there was a link from F to H, Router A would have 2 paths to H (A->G->H and A->B->E->F->H)
- Nothing else changes



Forwarding Adjacency

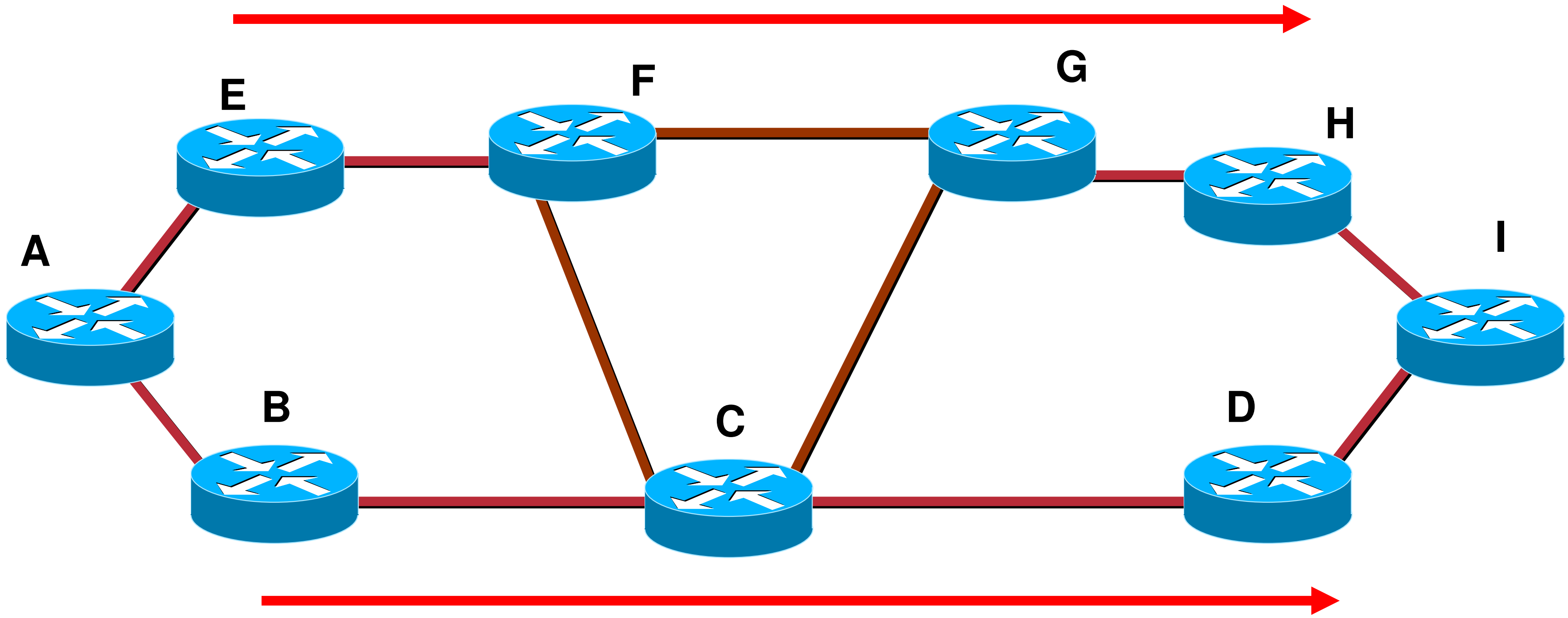
- **With autoroute, the LSP is not advertised into the IGP**
- **This is the right behavior if you're adding TE to an IP network, but maybe not if you're migrating from ATM/FR to TE**
- **Sometimes advertising the LSP into the IGP as a link is necessary to preserve the routing outside the ATM/FR cloud**

ATM Model



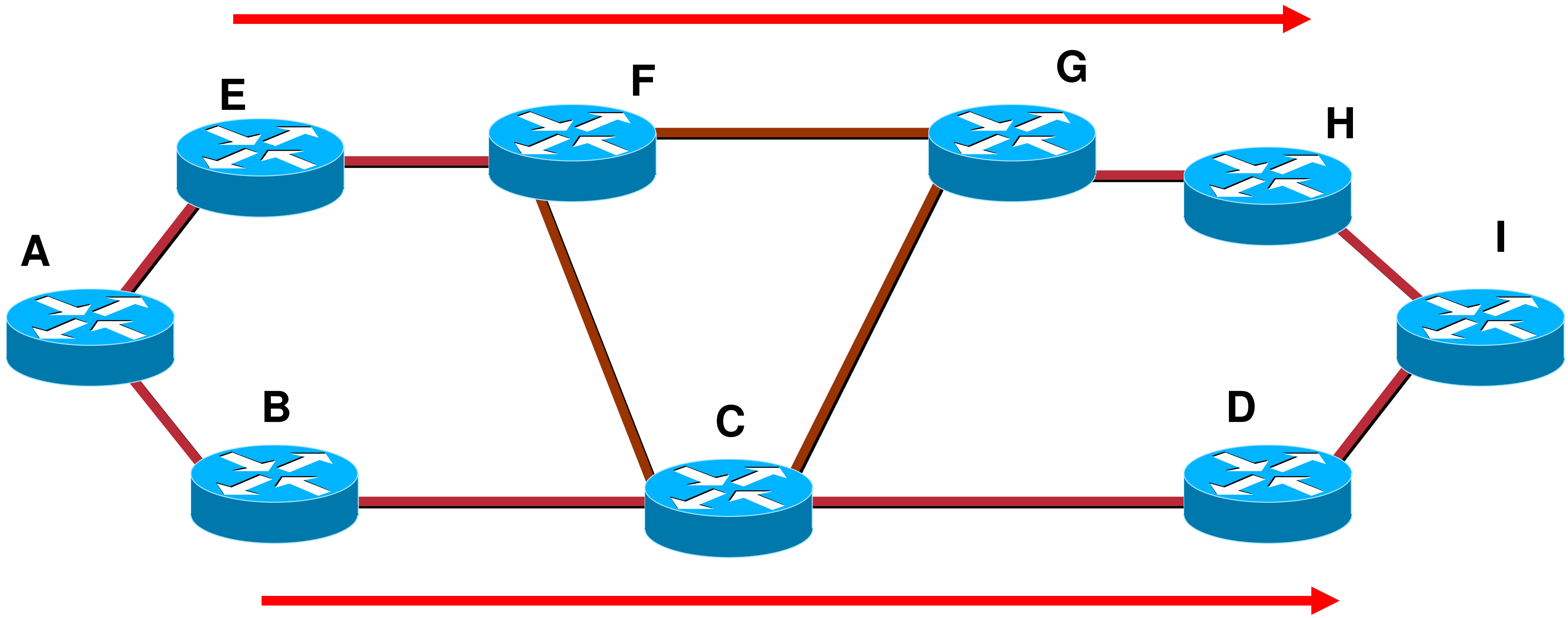
- Cost of ATM links (blue) is unknown to routers
- A sees two links in IGP—E->H and B->D
- A can load-share between B and E

Before FA



- All links have cost of 10
- A's shortest path to I is A->B->C->D->I
- A doesn't see TE tunnels on {E,B}, alternate path never gets used!
- Changing link costs is undesirable, can have strange adverse effects

F-A Advertises TE Tunnels in the IGP



- With forwarding-adjacency, A can see the TE tunnels as links
- A can then send traffic across both paths
- This is desirable in some topologies (looks just like ATM did, same methodologies can be applied)

Unequal Cost Load Balancing

- IP routing has equal-cost load balancing, but not unequal cost*
- Unequal cost load balancing difficult to do while guaranteeing a loop-free topology

*EIGRP Has ‘Variance’, but That’s Not As Flexible

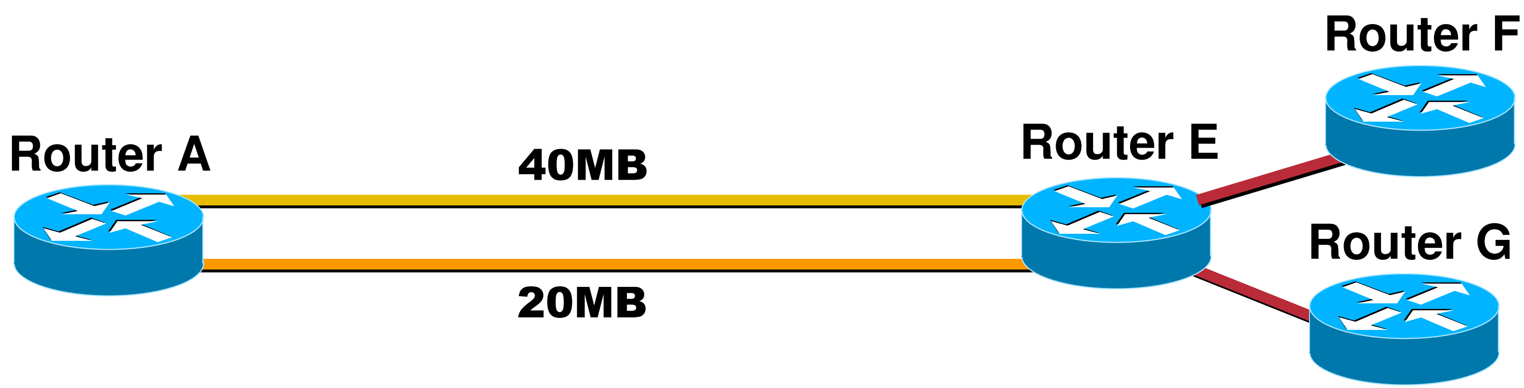
Unequal Cost Load Balancing

Cisco.com

- Since MPLS doesn't forward based on IP header, permanent routing loops
  don't happen
- 16 hash buckets for next-hop, shared in rough proportion to configured tunnel bandwidth or load-share value

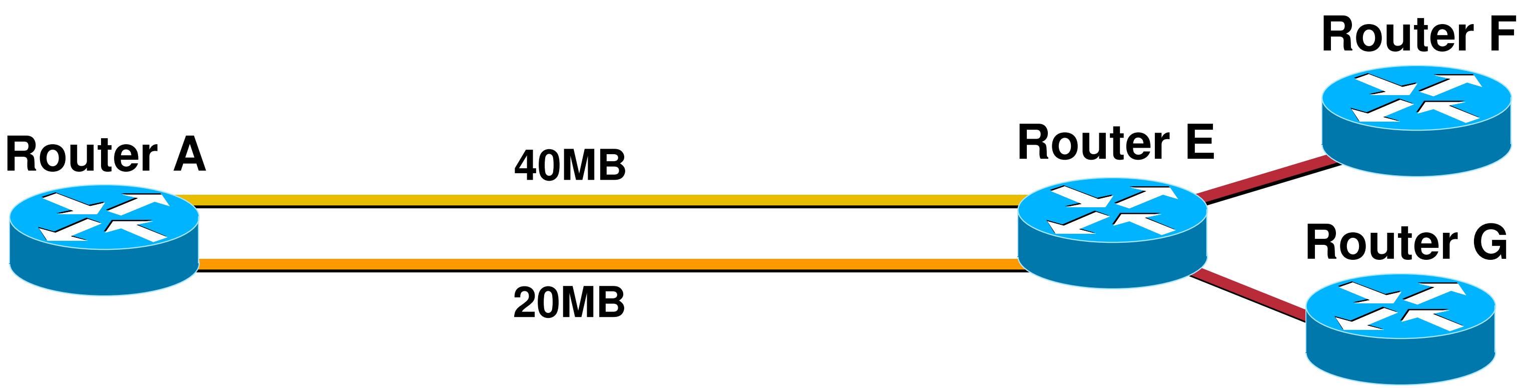
Unequal Cost: Example 1

Cisco.com




```
gsr1#show ip route 192.168.1.8
Routing entry for 192.168.1.8/32
  Known via "isis", distance 115, metric 83, type level-2
  Redistributing via isis
  Last update from 192.168.1.8 on Tunnel0, 00:00:21 ago
  Routing Descriptor Blocks:
  * 192.168.1.8, from 192.168.1.8, via Tunnel0
      Route metric is 83, traffic share count is 2
    192.168.1.8, from 192.168.1.8, via Tunnel1
      Route metric is 83, traffic share count is 1
```

Unequal Cost: Example 1




```
gsr1#sh ip cef 192.168.1.8 internal
.........
Load distribution: 0 1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 (refcount 1)
  Hash  OK  Interface                 Address          Packets  Tags imposed
  1     Y   Tunnel0                   point2point            0    {23}
  2     Y   Tunnel1                   point2point            0    {34}
.........
```

Note That the Load Distribution Is 11:5—Very Close to 2:1, but Not Quite!

Practice

- **Prerequisites (global config)**

```
ip cef {distributed}
mpls traffic-eng tunnels
```

Practice

- **Build a tunnel interface (headend)**

```
interface Tunnel0
  tunnel mode mpls traffic-eng
 ip unnumbered loopback0
 tunnel destination <RID of tail>
```

Information Distribution

OSPF

```
mpls traffic-eng tunnels
mpls traffic-eng router-id loopback0
mpls traffic-eng area <x>
```

ISIS

```
mpls traffic-eng tunnels
mpls traffic-eng router-id loopback0
mpls traffic-eng level-<x>
metric-style wide
```

Information Distribution

on each physical interface

`mpls traffic-eng tunnels`

(optional) `ip rsvp bandwidth {x}`

Path Calculation

Cisco.com

EITHER

```
int Tunnel0
 tunnel mpls traffic-eng path-option <num> dynamic
```

OR

```
int Tunnel0
 tunnel mpls traffic-eng path-option <num> explicit
  name foo
```

Path Calculation

Global config:

```
ip explicit-path name foo
 next-address 1.2.3.4 {loose}
 next-address 1.2.3.8 {loose}
```

(etc)

Path Calculation

Cisco.com

Global config:

```
ip explicit-path name foo
 next-address 1.2.3.4 {loose}
 next-address 1.2.3.8 {loose}
```

(etc)

Path Calculation

- **Can have several path options, to be tried successively**

```
tunnel mpls traffic-eng path-option 10
  explicit name foo

tunnel mpls traffic-eng path-option 20
  explicit name bar

tunnel mpls traffic-eng path-option 30
  dynamic
```

Path Setup

Cisco.com

- **Nothing to configure to explicitly enable path setup**
- **`mpls traffic-eng tunnels` (from before) implicitly enables RSVP on the physical i/f**

Routing Traffic Down A Tunnel

Cisco.com

Autoroute:

`tunnel mpls traffic-eng autoroute announce`

Forwarding adjacency:

`tunnel mpls traffic-eng forwarding-adjacency`

then

` isis metric <x> level-<y>`

or

` ip ospf cost <x>`

on tunnel interface

Static routes


```
ip route <prefix> <mask> Tunnel0
```

Policy routing


```
access-list 101 permit tcp any any eq www
interface Serial0
 ip policy route-map foo
route-map foo
 match ip address 101
 set interface Tunnel0
```

Summary Config

Cisco.com

```
ip cef (distributed}
mpls traffic-eng tunnels

interface Tunnel0
 tunnel mode mpls traffic-eng
 ip unnumbered Loopback0
 tunnel destination <RID of tail>
 tunnel mpls traffic-eng autoroute announce
 tunnel mpls traffic-eng path-option 10 dynamic
```

Summary Config

(in IGP)

```
mpls traffic-eng tunnels
mpls traffic-eng router-id Loopback0
mpls traffic-eng area <x>
mpls traffic-eng level-<x>
metric-style wide
```

OSPF

ISIS

(

physical interface)

```
interface POS0/0
 mpls traffic-eng tunnels
 ip rsvp bandwidth <kbps>
```

Tips

- Some of the more useful ones:

1. To advertise implicit-null from Tail-end

   mpls traffic-eng signalling advertise implicit-null

2. To interpret explicit-null at PHP (hidden command)

   mpls traffic-eng signalling interpret explicit-null

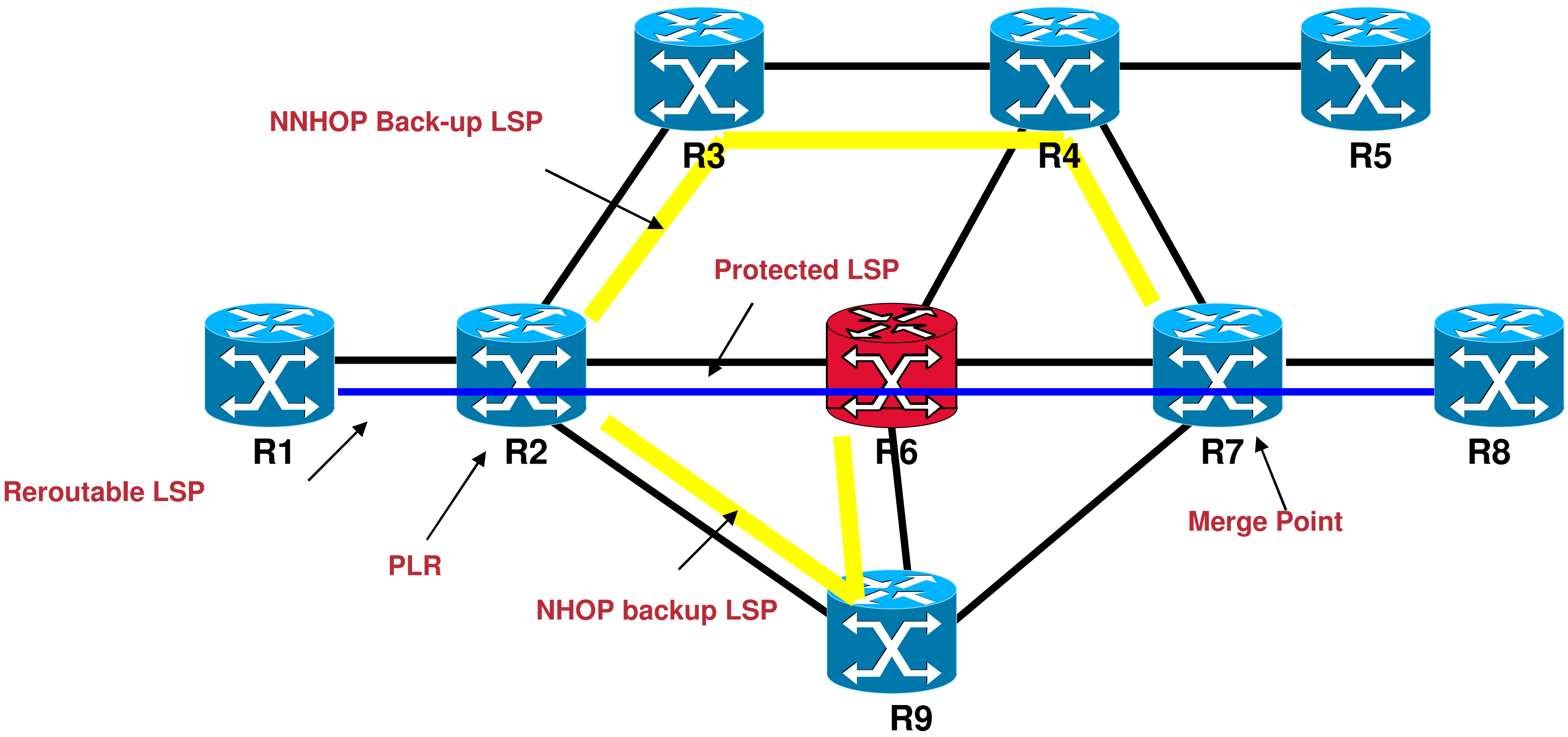
3. To automatically consider any new links as they come up

   mpls traffic-eng reoptimize events link-up

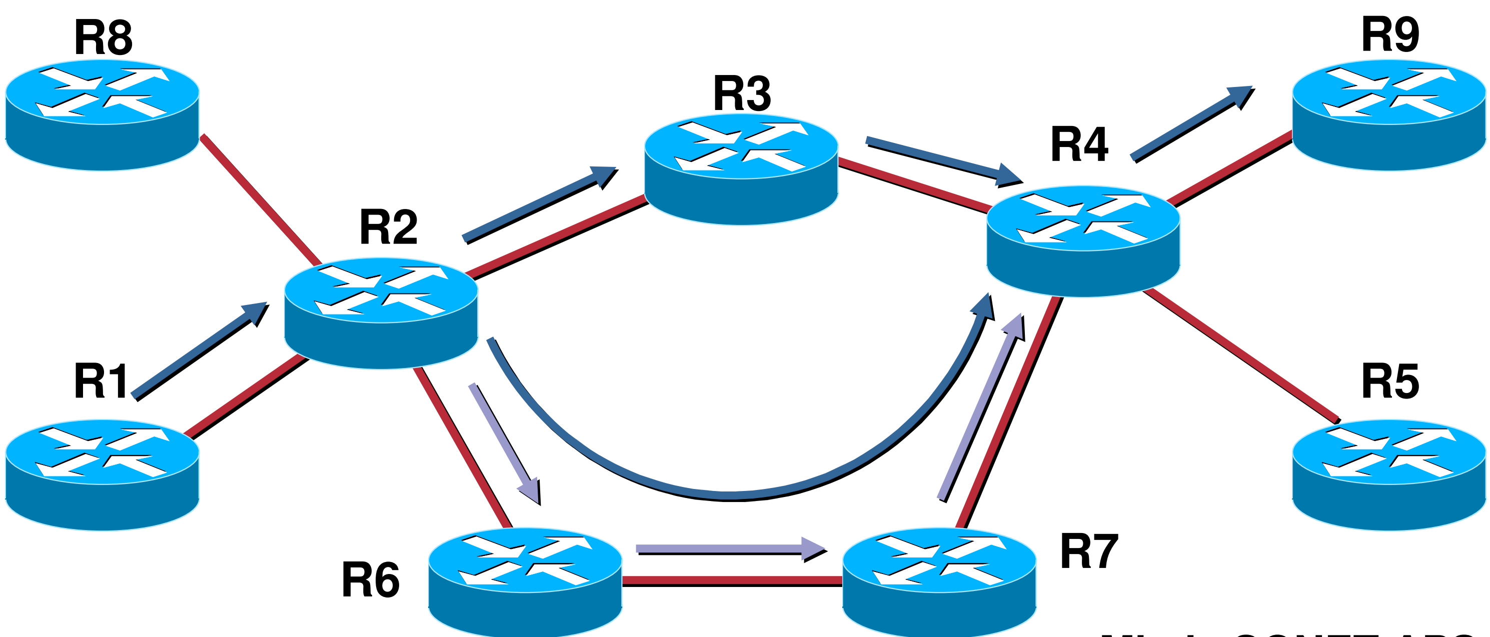
Fast ReRoute

- Fundamental point from earlier: “you can use MPLS-TE to forward traffic down a path other than that determined by your IGP cost”
- FRR builds a path to be used in case of a failure in the network
- Minimize packet loss by avoiding transient routing loops

Terminology



Applications of MPLS TE – MPLS Fast Re-Route



Mimic SONET APS
Re-route in 50ms or Less

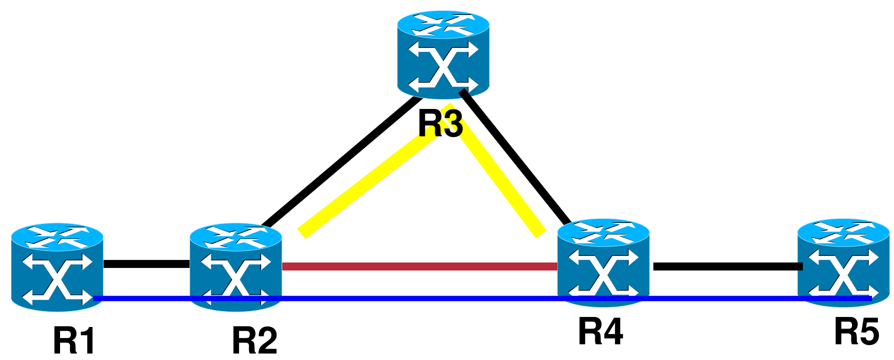
- Multiple hops can be by-passed; R2 swaps the label which R4 expects before pushing the label for R6
- R2 locally patches traffic onto the link with R6

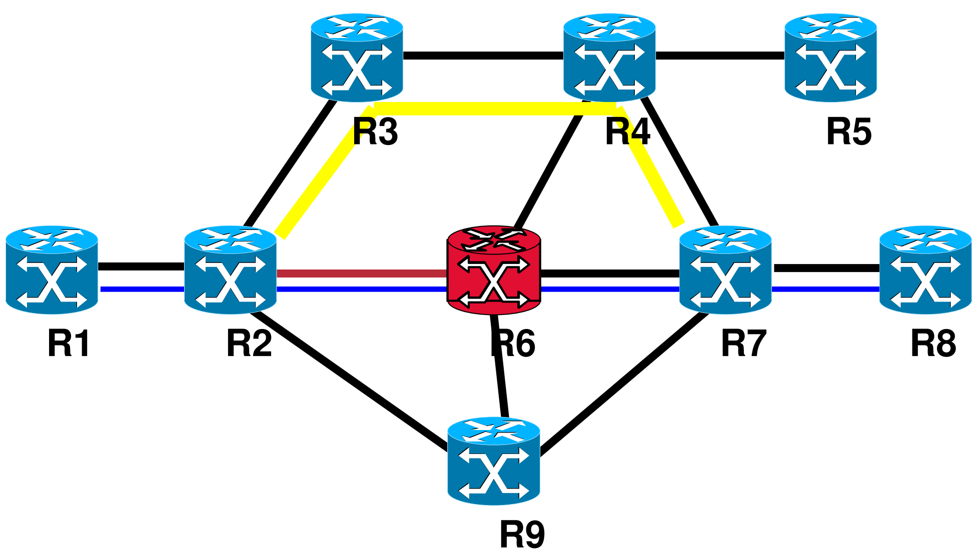
Fast ReRoute

Cisco.com

MPLS Fast Reroute local repair

- **Link protection: the backup tunnel tail-head (MP) is one hop away from the PLR**





- **Node protection: the backup tunnel tail-end (MP) is two hops away from the PLR.**

IP failure recovery

Cisco.com

For IP to recover from a failure, several things need to happen:

Thing	Time
Link Failure Detection	usec-msec
Failure Propagation + SPF	- hundreds of msec with aggressive tuning (400ms for 500 pfx) - sec (5-10) with defaults
Local forwarding rewrite	<100ms
TOTAL:	~500ms-10sec

FRR failure recovery

Cisco.com

Since FRR is a local decision, no propagation needs to take place.

Thing	Time
Link Failure Detection	usec-msec
Failure Propagation+SPF	0
Local forwarding rewrite	<100ms
TOTAL:	<100ms (often <50ms, <10ms with properly greased skateboard)

Caveats

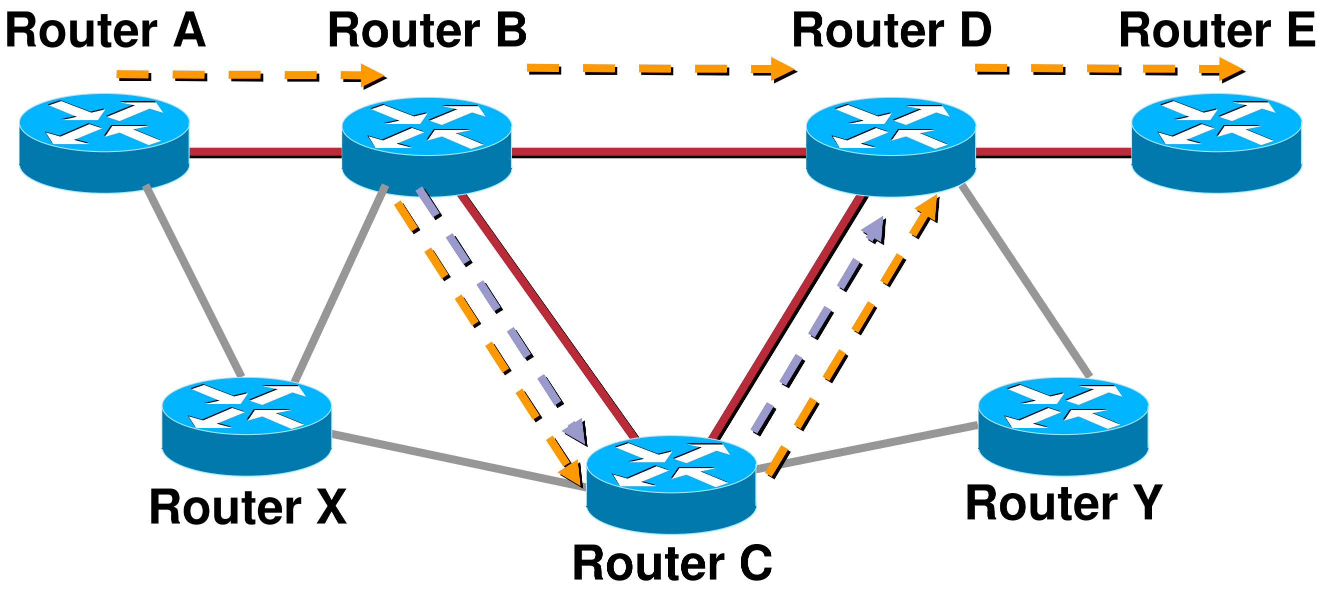
- As always, your mileage may vary. One slide does not do IP or FRR justice.
- Local failure recovery is always faster than distributed failure recovery
- What meets your needs? What makes more sense for your network? etc,..

FRR Procedures

1. pre-establish backup paths
2. failure happens, protected traffic is switched onto backup paths
3. after local repair, tunnel headends are signalled to recover if they want. No time pressure here, failure is being protected against
4. protection is in place for hopefully ~10-30+ seconds. during that time, data gets through.

Link Protection

Cisco.com

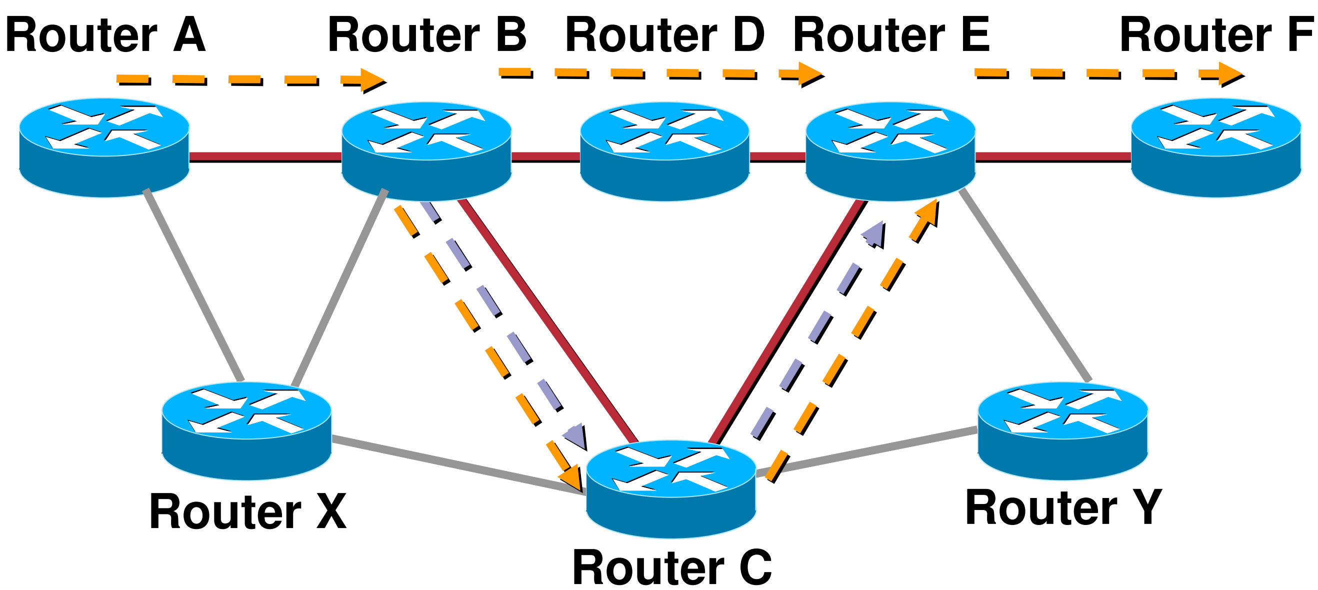


- Primary Tunnel: A -> B -> D -> E
- Backup Tunnel: B -> C -> D (Pre-provisioned)
- Recovery = ~50ms

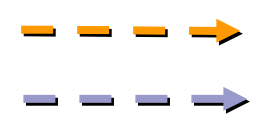
*Actual time varies—well below 50ms in lab tests, can also be higher

Node Protection

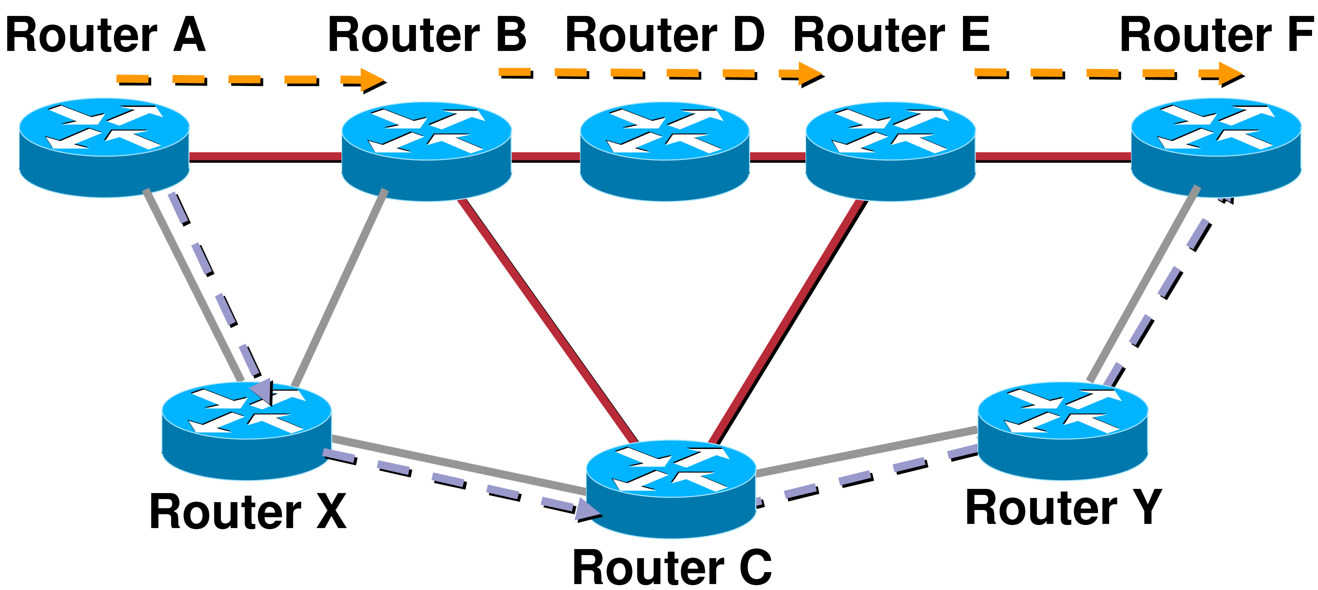
Cisco.com



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



Path Protection



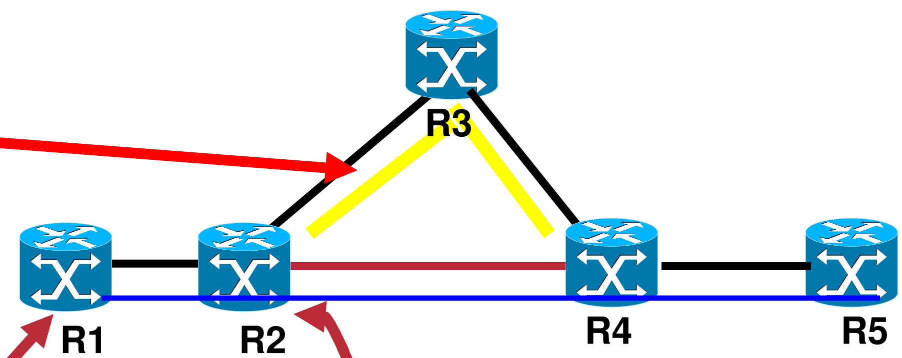
- Primary Tunnel: A -> B -> D -> E -> F
- BackUp Tunnel: A ->X -> C -> Y -> F (Pre-provisioned)
- Recovery = >100ms

FRR Configuration

Cisco.com

1) configure protection tunnel on R2

```
interface Tunnel0
 .. dest R4
 .. explicit-path R2-R3-R4
 .. NO autoroute!!!
```



2) protect an interface

```
interface POS0/0
 mpls traffic-eng backup-path Tunnel0
```

3) headend requests protection

```
interface Tunnel0
 .. dest R4
 .. etc ...
tunnel mpls traffic-eng fast-reroute
```

FRR Tips

- Bandwidth protection vs. connectivity protection is the big one
- Do not want to reserve bandwidth on the protection tunnel, this is wasteful
- Either use TBPro (see later) or backup bandwidth on the protection tunnel (yellow tunnel in previous slide)

```
tunnel mpls traffic-eng backup-bw <kbps>
```

- Allows backup to be a little smart about where it protects primary tunnels
- Only really useful if protecting 1 interface with >1 tunnels
- Offline calculation can be much smarter, but there’s operational tradeoffs

Design and Scaling

- Designing with primary tunnels
- Designing with backup tunnels

Designing with primary tunnels

- Full mesh (strategic TE)
  - Mesh of TE tunnels between a level of routers
  - Typically P<->P, can be PE<->PE in smaller networks
  - O(N^2) LSPs
- As-needed (tactical TE)
  - Put a tunnel in place to work around temporary congestion due to unforseen shift in traffic demand
  - Need to keep an eye on your tunnels

Strategic TE (full mesh)

- **Supported scalability numbers:**

  600 tunnel headends per node

  10,000 midpoints per node

- **Largest numbers deployed today:**

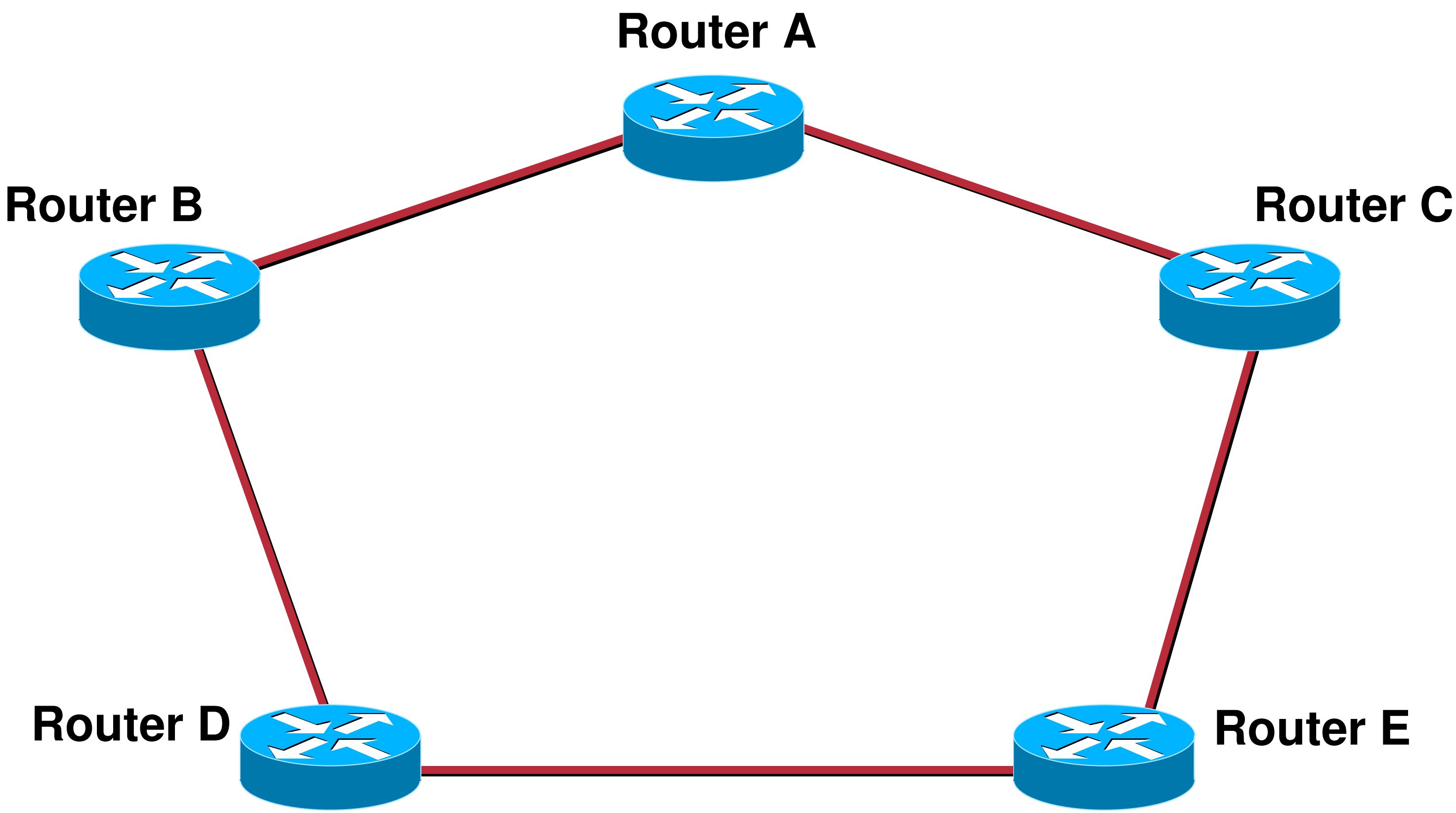
  100 routers full mesh = ~10,000 tunnels in the network

  As many as 2,000-3,000 at certain midpoints

  Plenty of room to grow!

Strategic

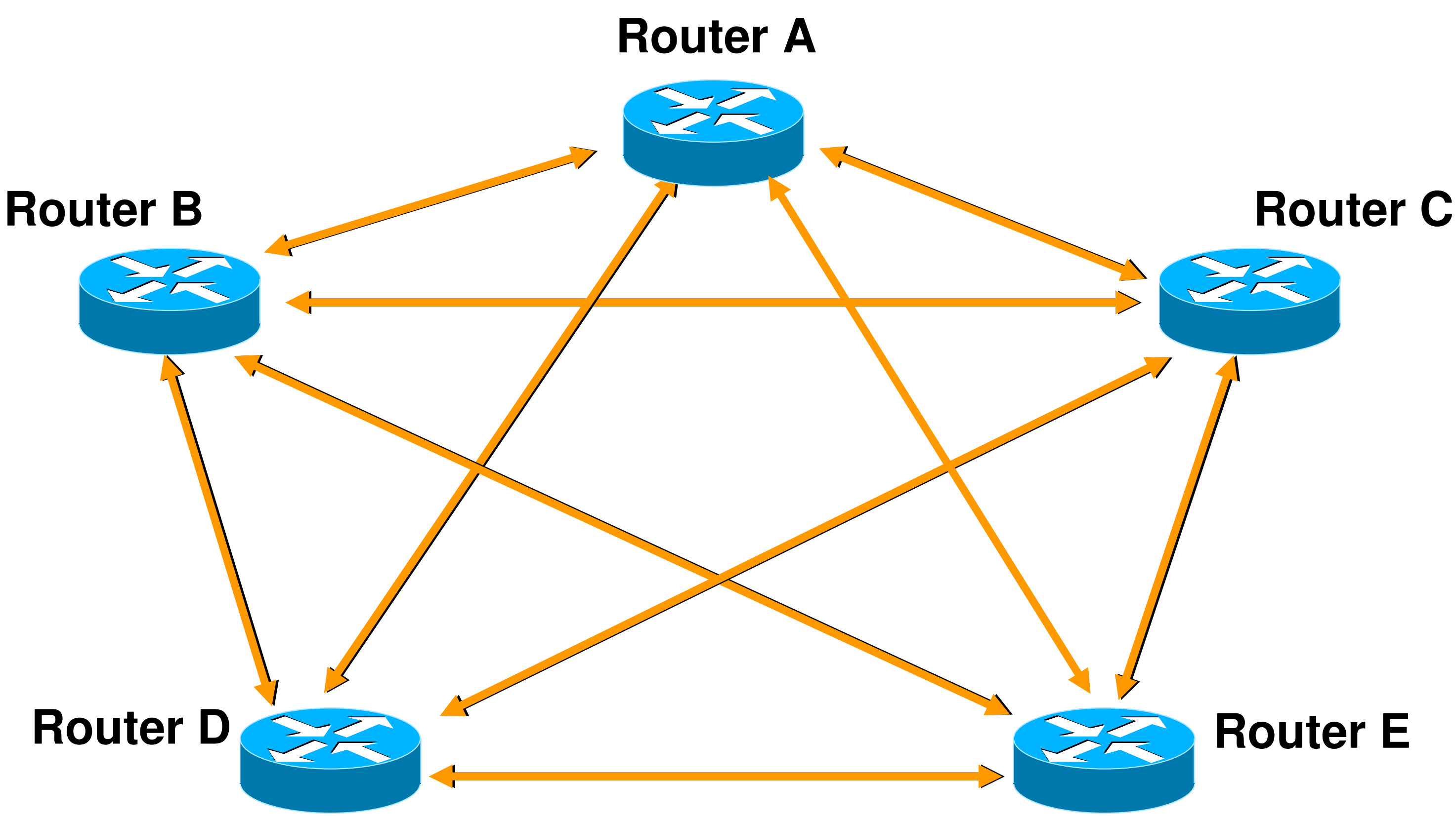
- **Physical topology is:**



Strategic

Cisco.com

- **Logical topology is***
  - ***Each link is actually 2 unidirectional tunnels**
- **Total of 20 tunnels in this network**

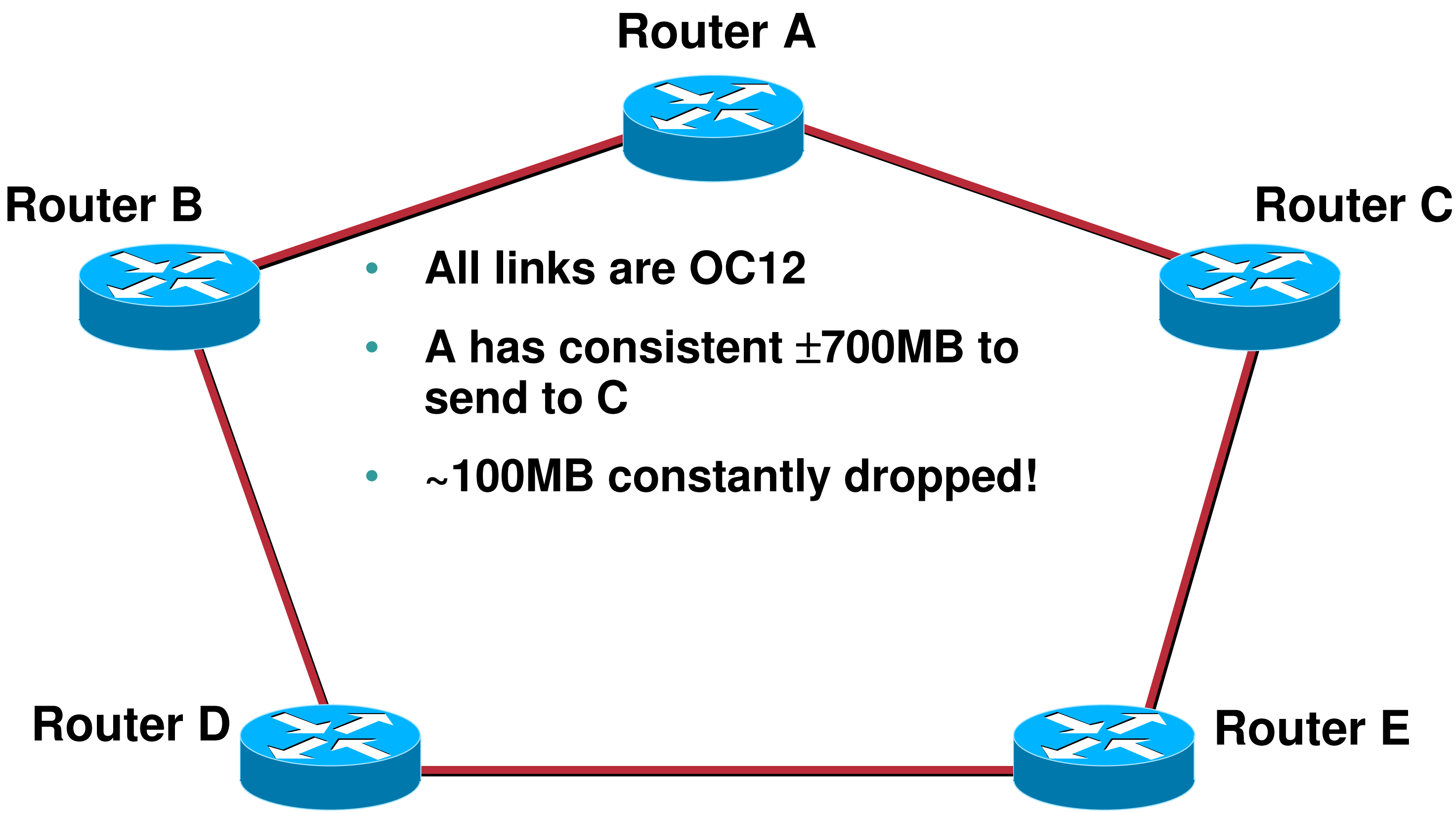


Strategic

Cisco.com

- **Things to remember with full mesh**
  - **N routers, N*(N-1) tunnels**
  - **Routing protocols not run over TE tunnels—Unlike an ATM/FR full mesh!**
  - **Tunnels are unidirectional—This is a good thing**
    - **…Can have different bandwidth reservations in two different directions**

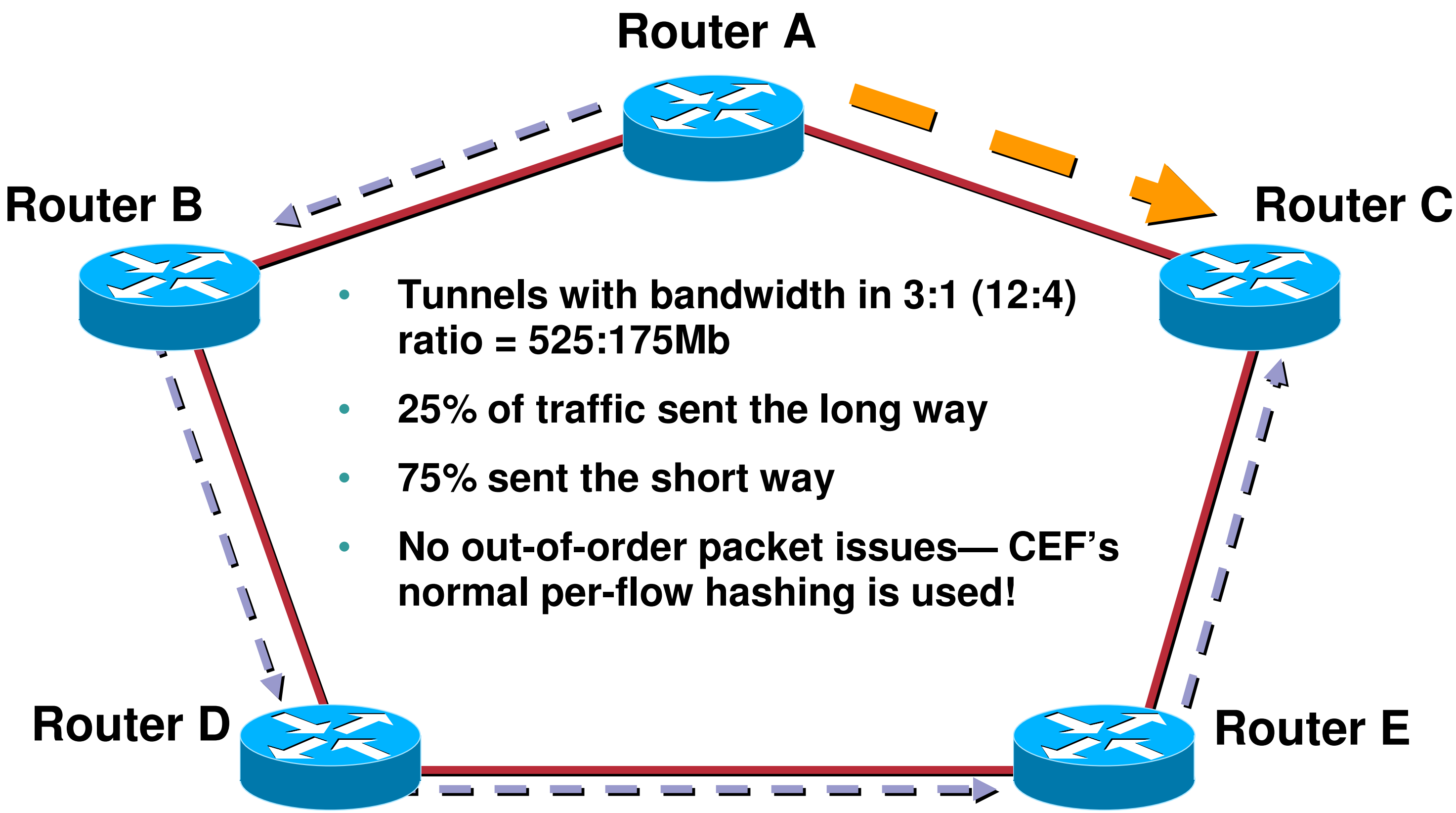
Case Study: A Large US ISP



- All links are OC12
- A has consistent ±700MB to send to C
- ~100MB constantly dropped!

Tactical

- **Solution: Multiple tunnels, unequal cost load sharing!**



- **Tunnels with bandwidth in 3:1 (12:4) ratio = 525:175Mb**
- **25% of traffic sent the long way**
- **75% sent the short way**
- **No out-of-order packet issues— CEF's normal per-flow hashing is used!**

Strategic vs. Tactical

- Both methods are in use today and have been for some years now
- Strategic means you always have a tunnel, and it means you have a lot of tunnels

  Consistent mode of operation, lots of interfaces to manage
- Tactical means you only have tunnels when you have problems

  …which means removing tunnels that are no longer necessary
- Which one you pick is up to you, both methods are valid

Designing with backup tunnels

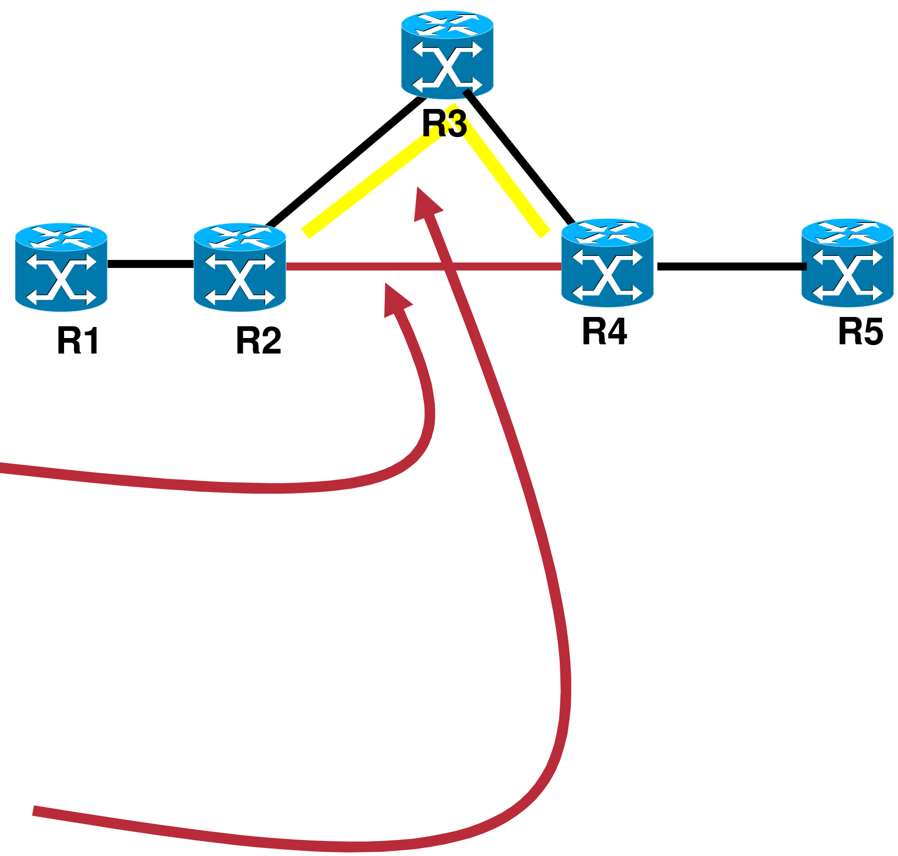
- Connectivity protection
  - Router calculates the path for its backup tunnel
  - Assume that any found path can carry any link’s traffic during failure
  - Don’t signal bandwidth for the backup tunnel!
  - Use DiffServ to solve any contention due to congestion while FRR is in use
- Bandwidth protection
  - Offline tool calculates paths for protection LSPs
  - Assurance that bandwidth is available during failure
  - More complex to maintain, may require additional network bandwidth
  - Allows you to always meet SLAs during failure

Reasonable combinations

primary -> | v backup	none	tactical	strategic
none	IP	TE to work around congestion	Bandwidth optimization (online or offline)
connectivity protection	1hop online	1hop online + sporadic tactical	Bandwidth optimization (online or offline) + online backup
bandwidth protection	1hop offline	1hop online + sporadic tactical	Bandwidth optimization (online or offline) + offline backup

1hop FRR

- Useful if you want to take advantage of FRR but don't need primary bandwidth optimization
- All primary tunnels go between two directly connected nodes (tunnels are 1 hop long)
- Backup tunnel protects only that primary
- Currently in production in a few large IP (VPN, VoIP) networks



Bandwidth override on path option

Cisco.com

- **Can specify a bandwidth on a path-option that overrides the tunnel BW:**

```
tunnel mpls traffic-eng bandwidth 1000
tunnel mpls traffic-eng path-option 1
  explicit name path1
tunnel mpls traffic-eng path-option 2
  explicit name path2 bandwidth 500
tunnel mpls traffic-eng path-option 3
  dynamic bandwidth 0
```

LSP Attribute Lists

- **Control full set of LSP attributes per path, not per tunnel**
- **More complex, more powerful**

AutoTunnel

- Obviates need to configure NHop and NNHop backup tunnels
- Further enhancements on the radar (mesh groups)

```
mpls traffic-eng auto-tunnel backup
```

- No configuring backup or 1-hop primary tunnels!
- Tradeoff between convenience and flexibility

Benefits of TE over Policy Routing

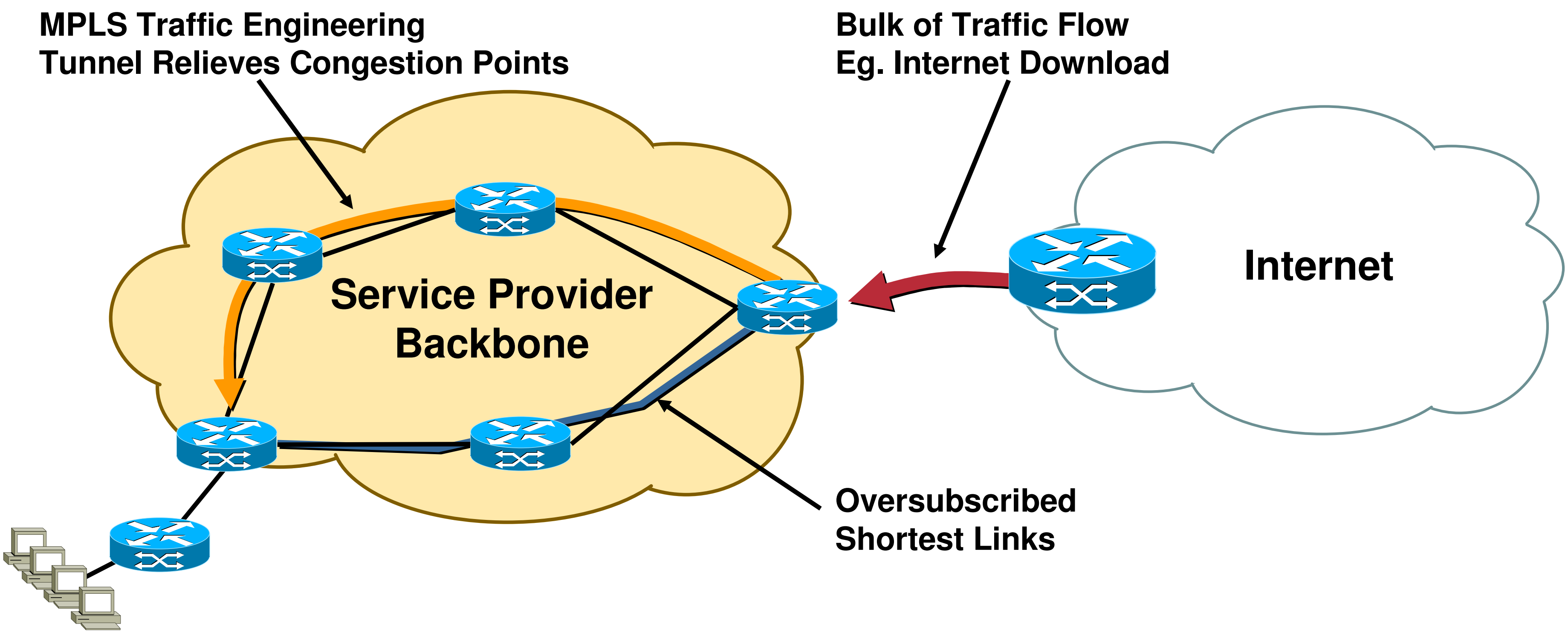
- **Policy Routing**
  - Hop-by-hop decision making
  - No accounting of bandwidth
- **Traffic Engineering**
  - Head end based
  - Accounts for available link bandwidth
  - Admission control

TE Deployment Scenarios

Tactical TE Deployment

Cisco.com

Requirement: Need to handle scattered congestion points in the Network
Solution: Deploy MPLS TE on only those nodes that face congestion

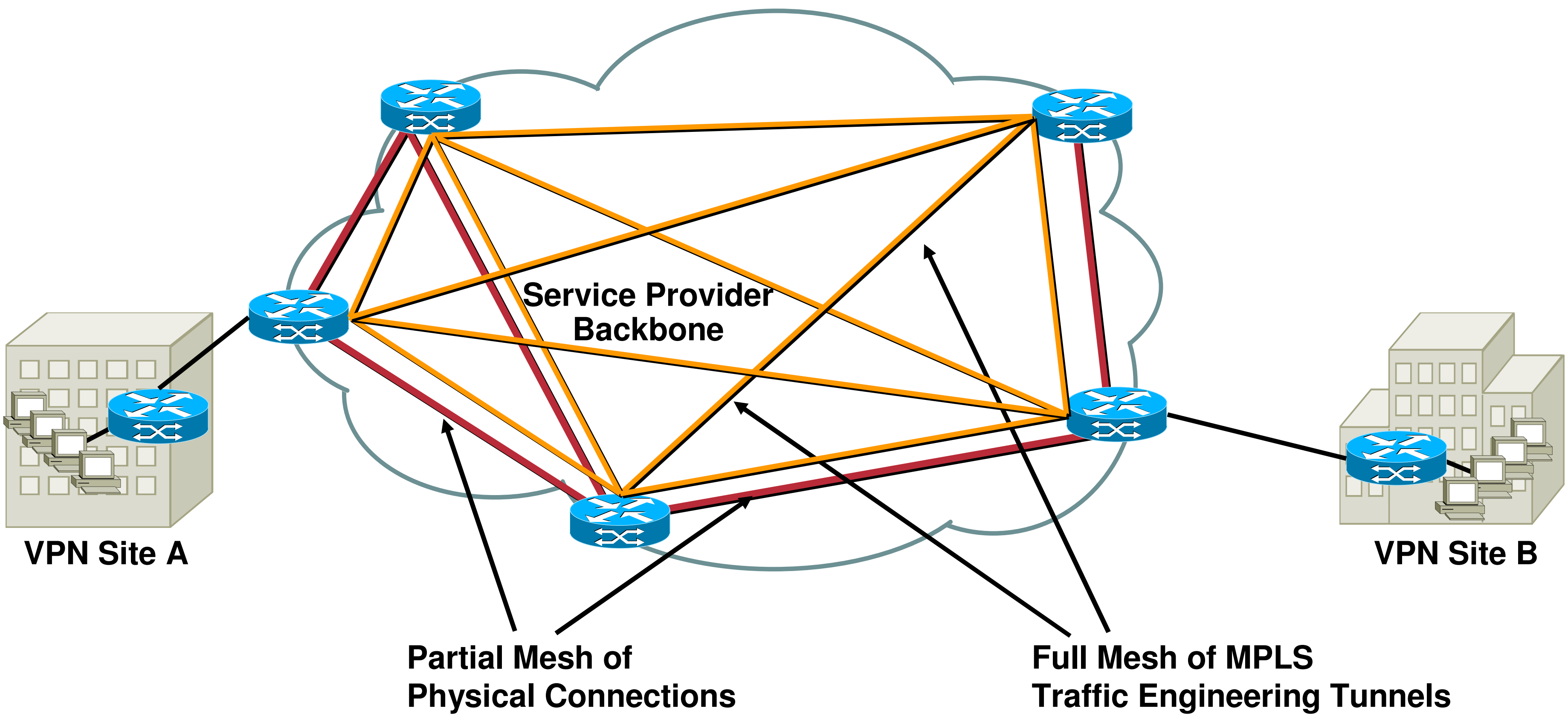


Full Mesh TE Deployment

Cisco.com

Requirement: Need to increase "bandwidth inventory" across the network

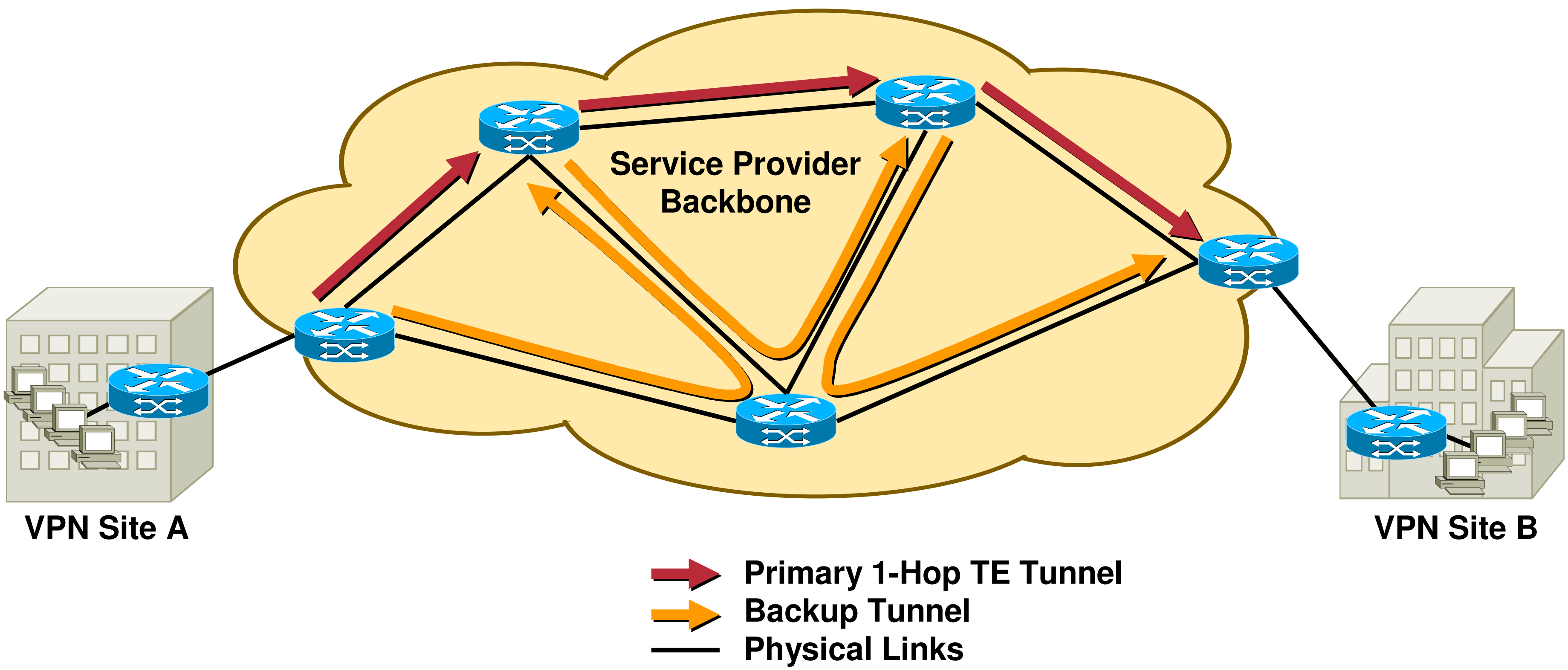
Solution: Deploy MPLS TE with a full logical mesh over a partial physical mesh and use Offline Capacity Planning Tool



1-Hop TE Deployment

Requirement: Need protection only—minimize packet loss
Lots of Bandwidth in the core

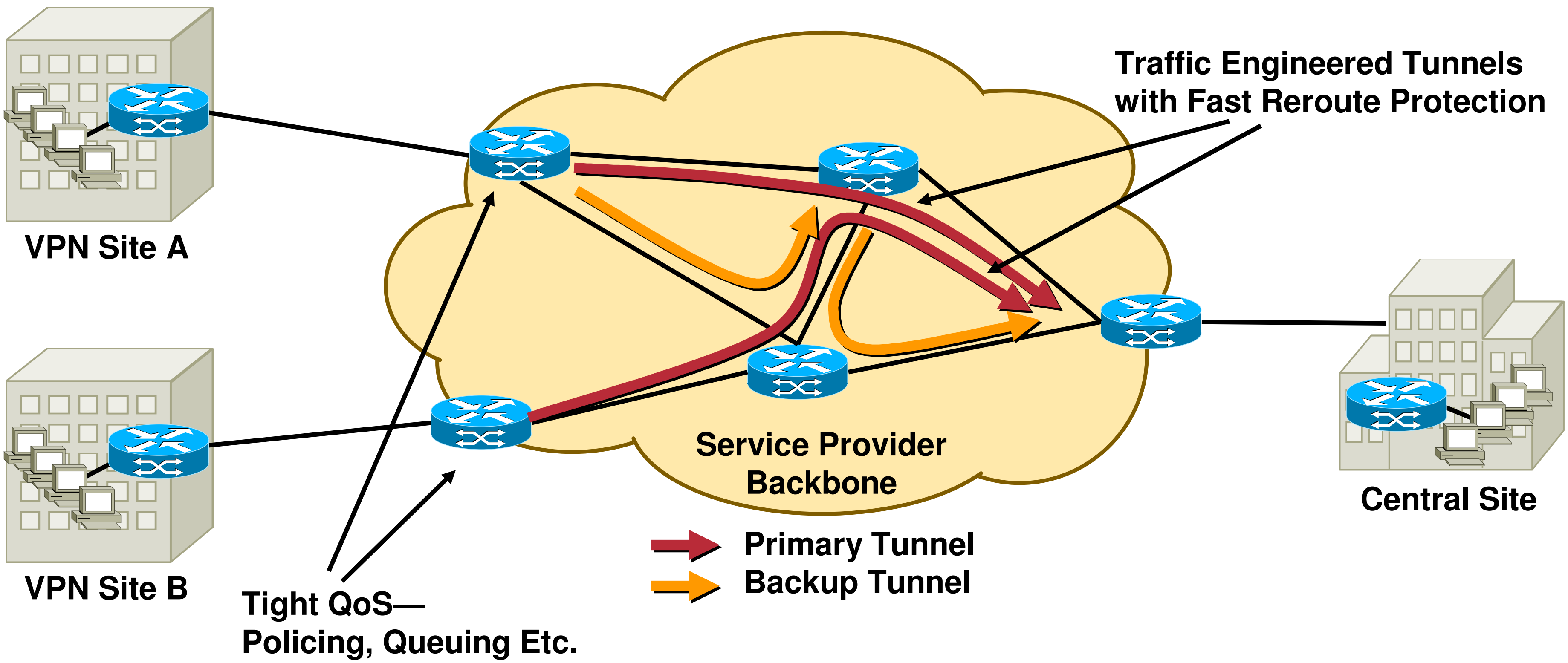
Solution: Deploy MPLS Fast Reroute for less than 50ms failover time with 1-Hop Primary TE Tunnels and Backup Tunnel for each



Virtual Leased Line Deployment

Requirement: Need to create dedicated point-to-point circuits with bandwidth guarantees—Virtual Leased Line (VLL)

Solution: Deploy MPLS TE (or DS-TE) with QoS; Forward traffic from L3 VPN or L2 VPN into a TE Tunnel; Unlike ATM PVCs, use 1 TE Tunnel for multiple VPNs creating a scalable architecture



MPLS TE Summary

Cisco.com

- Useful for re-routing traffic in congested environments
- Build innovative services like Virtual Leased line
- Build protection solutions using MPLS FRR

Management Considerations and MPLS OAM

Monique Morrow

What is MPLS Operations And Management?

- The tools and techniques required to successfully deploy an MPLS network

Fault-management
Configuration
Accounting
Performance
Security

Customer Requirements

Cisco.com

- Three categories of requirements from 1st tier PWE/MPLS Service Providers (and others).
  - ✓VC/LSP Path Verification and Tracing
  - ✓Built-in Protocol Operations
  - ✓Standard Management APIs/NMS Applications

    MIBs, CLI, XML, etc…

  - ➢Documented in: draft-ietf-mpls-oam-requirements-01.txt
  - ➢Must be addressed *before* many providers will deploy PWE3 services.

Summary Customer Requirements

- Management: Enabling service delivery
  - Fault management
  - Service Management
- ILEC view of network management very different than ISPs
  - Fault detection, isolation (details coming up)
- Customer visible OAM
  - OAM Emulation for ATM AAL5
  - OAM cell generation for ATM over MPLS upon change of VC status (eg – label withdrawal)
  - OAM Cell generation for LC_ATM

Fault Detection and Isolation

Control Plane Verification

- Consistency check
- Authentication

Data Plane Verification

- Ability to verify connectivity and trace
  - Paths from PE to PE – Global routing table as well as VPNs
  - Paths from CE to CE within a VPN
  - TE tunnels
  - Pseudo-wires

VC/LSP Connection Verification and Trace Requirements

- **Automated detection and diagnosis of broken transport LSPs and VCs:**
  - Point-to-point
  - Multipoint-to-point
  - Equal Cost Multi-Path (ECMP)
  - Using LSP ping/tunnel trace capability from both head-end and mid-points.
- **Data plane OAM packets must follow same path they are testing!**

VC/LSP Connection Verification and Trace Requirements (cont)

- Automatic lightweight IP-like ping to test end-to-end path connectivity (e.g.: CE-CE).
- Operator configurable parameters/actions:
  - –Frequency of VCCV.
  - –MPLS Fast-Reroute
  - –Automated VCCV
- Verification of VPN integrity by providing a mechanism to detect LSP mis-merging.
- Documented in:

www.ietf.org/internet-drafts/draft-ietf-pwe3-vccv-01.txt

LSP Ping

- Similar to ICMP (IP) Ping
  - Sequence Number
  - Timestamps
  - Sender Identification
- Full identification of FEC based the application
- Variable length for MTU discovery
- Support for tunnel/path tracing
- Multiple-reply modes
- Handles ECMP
- Reference
  - http://www.ietf.org/internet-drafts/draft-ietf-mpls-lsp-ping-03.txt

MPLS Ping: Operation

- Ping Mode: Connectivity check of an LSP
  - Test if a particular "FEC" ends at the right egress LSR
- Traceroute Mode: Hop by Hop fault localization
- Uses two messages
  - MPLS Echo Request
  - MPLS Echo Reply
- Packet need to follow data path

MPLS Ping Message Format

Cisco.com

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Version Number        |         Must Be Zero          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Message Type |   Reply mode  |  Return Code  | Return Subcode|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Sender's Handle                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Sequence Number                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    TimeStamp Sent (seconds)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  TimeStamp Sent (microseconds)                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  TimeStamp Received (seconds)                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                TimeStamp Received (microseconds)              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            TLVs ...                           |
:                                                               :
:                                                               :
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Message Type

1 Echo Request

2 Echo Reply

Reply Mode

No reply

IPv4 UDP packet

IPv4 UDP packet with Router alert

Control Plane

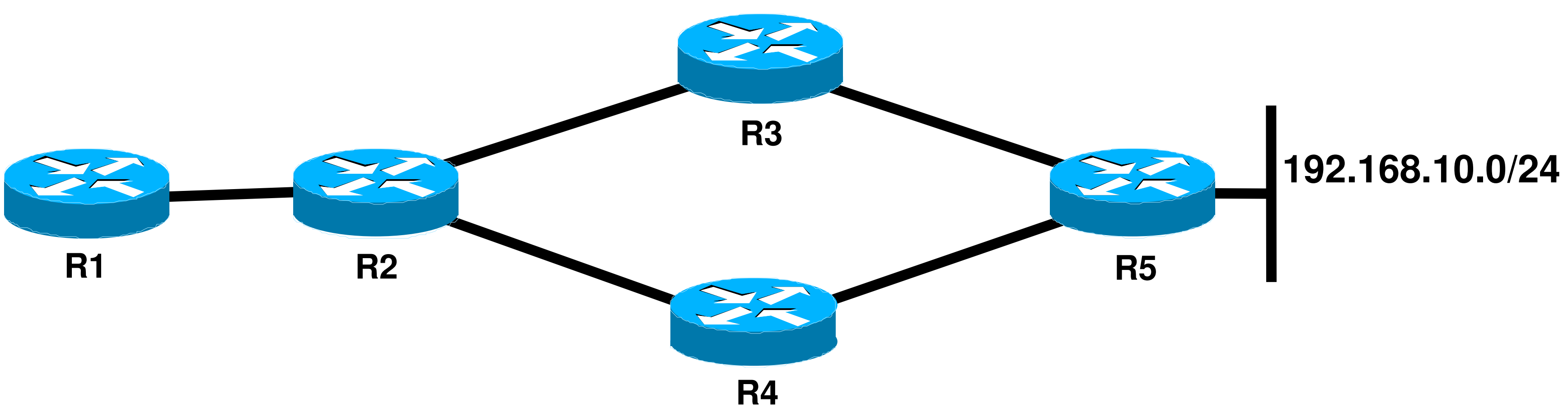
TLVs include

FEC to be checked

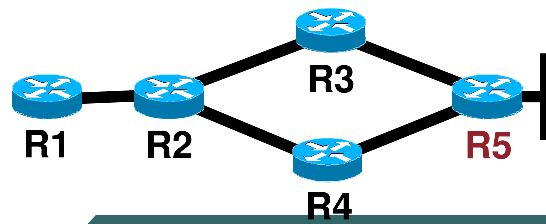
MPLS Ping: Packet Flow

Cisco.com

- Ping with label for FEC=192.169.10.0/24
- Label Switched at R2, R3
- R3 pops label off
- R4 processes packet



Packet Flow Ping Mode: Egress node

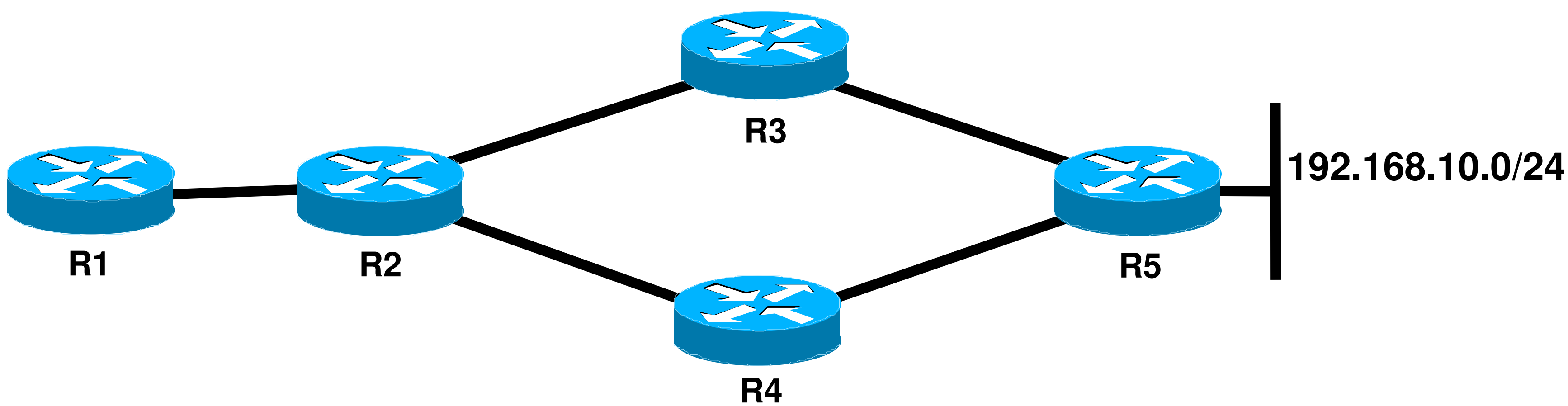


Cisco.com

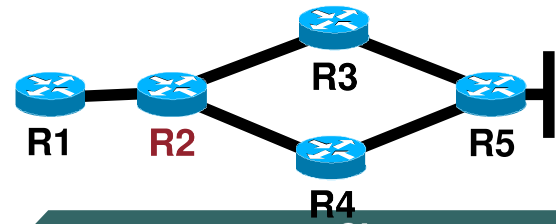
- Check Packet integrity
- Check if FEC distribution protocol is associated with incoming interface
- Check if valid egress node for the FEC
- Send echo Reply according to value of Reply Mode

MPLS Traceroute: Packet Flow

- MPLS Ping Packets are sent with TTL=1,2,3
- Label switched if TTL > 1
- Processed where TTL expires

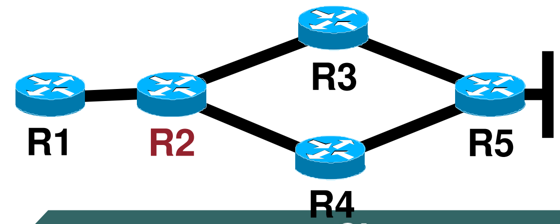


Packet Flow Trace Mode: Transit Node



- Reply processing same as Ping, then
- Check for Downstream Mapping TLV
  - Determine nexthop routers
- Add Downstream Mapping TLVs for each
  - Compute label stacks, address/label ranges
- Return received Label Stack if requested

Packet Flow Trace Mode: Transit Node



Cisco.com

- Reply processing same as Ping, then
- Check for Downstream Mapping TLV
  - Determine nexthop routers
- Add Downstream Mapping TLVs for each
  - Compute label stacks, address/label ranges
- Return received Label Stack if requested

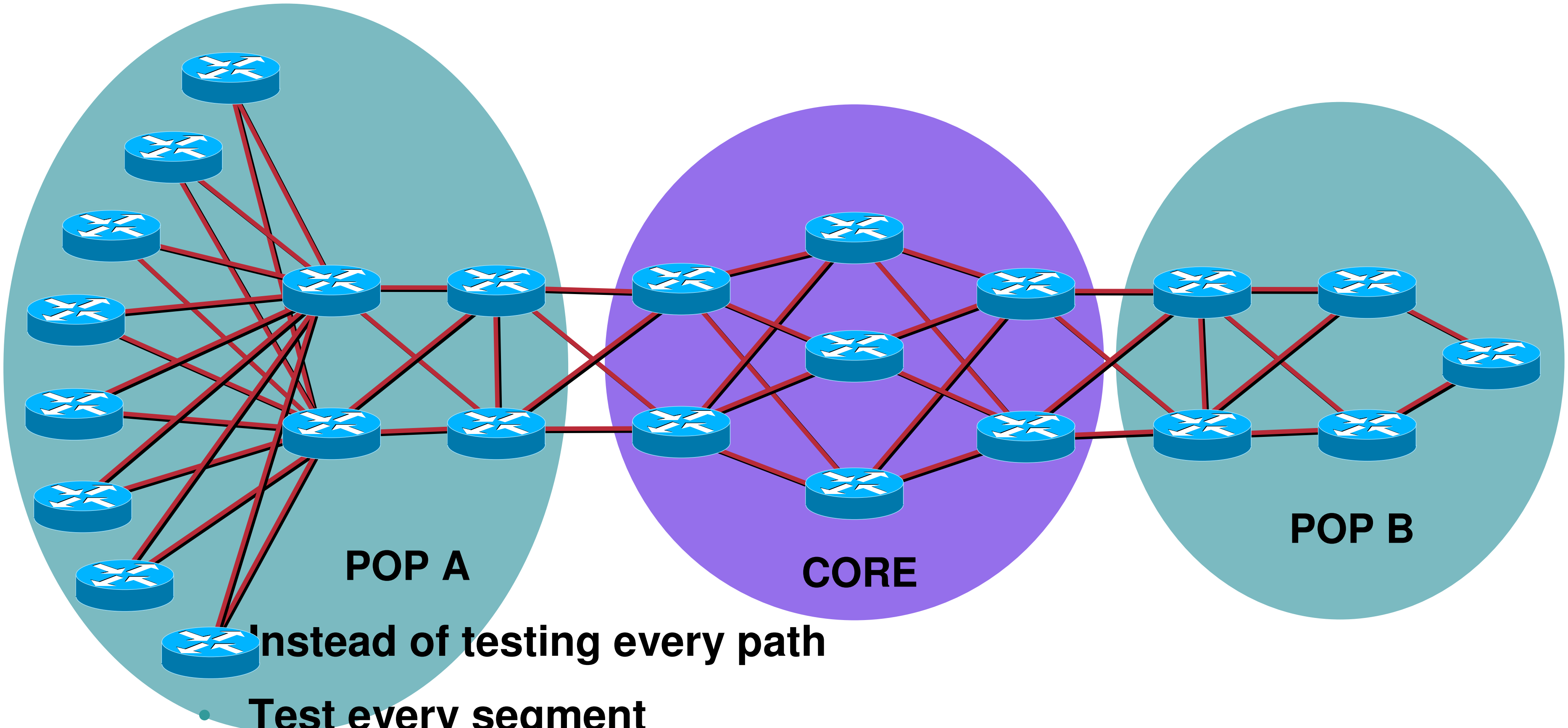
Trace Mode: TTL>1

- Copy one Downstream Mapping TLV from Echo Reply
- Pick one IP Address from address in DM TLV
- Send a new Echo Request with TTL+1
- Repeat (if appropriated) for each DM TLV
- Reply from Egress stops iteration

Motivation

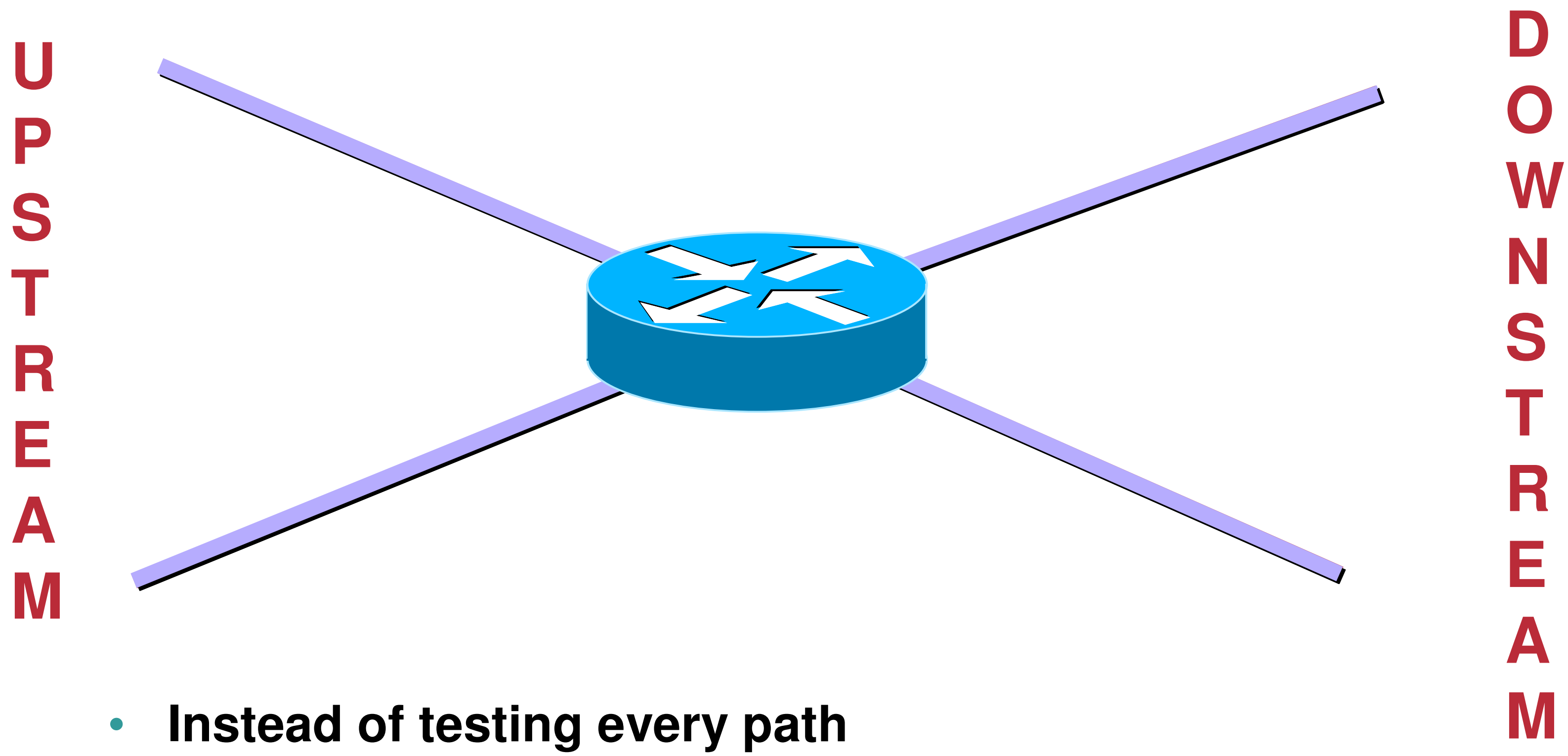
- Scalability
- Locality of alerts
- Exchange Link Local Identifiers if your IGP can't do it for you
- Test dormant paths

Self Test



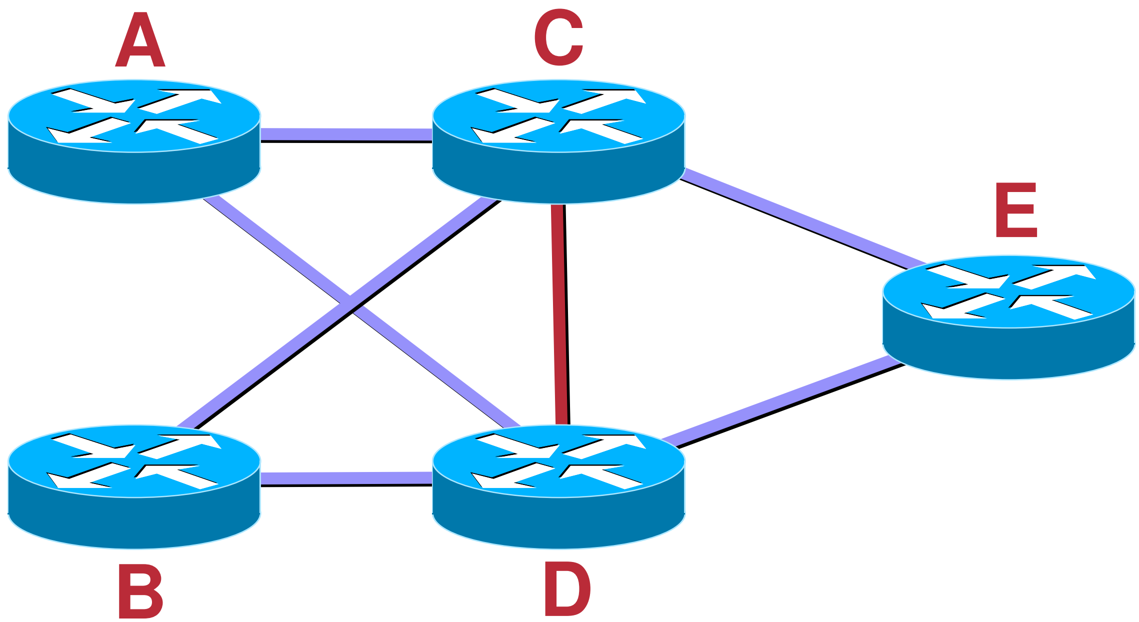
- Instead of testing every path
- Test every segment

Self Test



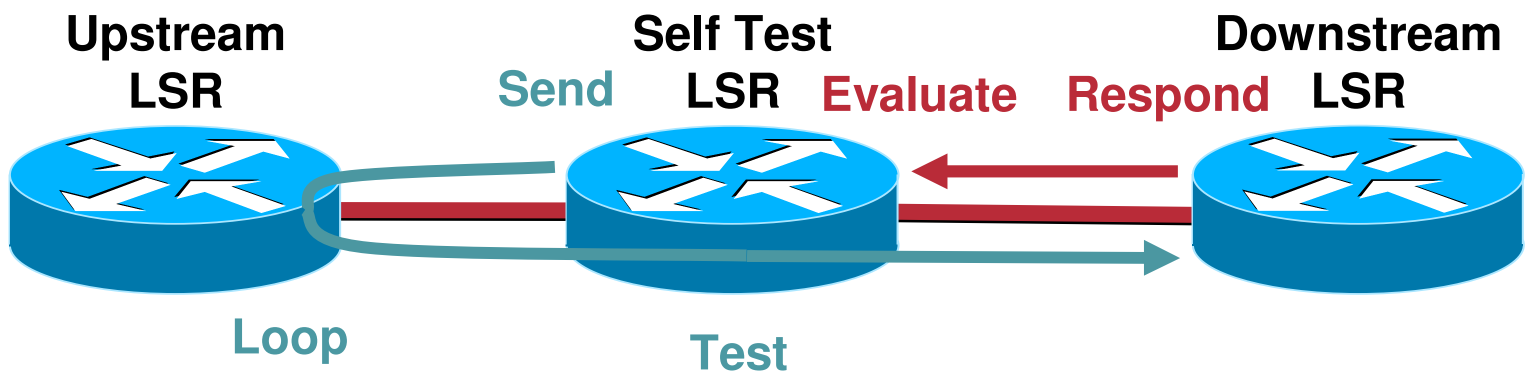
- **Instead of testing every path**
- **Test every segment**

Dormant Interfaces



- **Interface labels programmed ahead of time**
- **E2E OAM tests only active paths**
- **If link D-E fails link D will begin using link C-D C gets no notification of this event**

Overview of Operation



Two messages, five actions:

Echo Request
- Send (CP)
- Loop (DP)
- Test (DP)

Echo Reply
- Respond (CP*)
- Evaluate (CP)

CP – Control Plane
DP – Data Plane

*Handled on linecard

Loopback Label

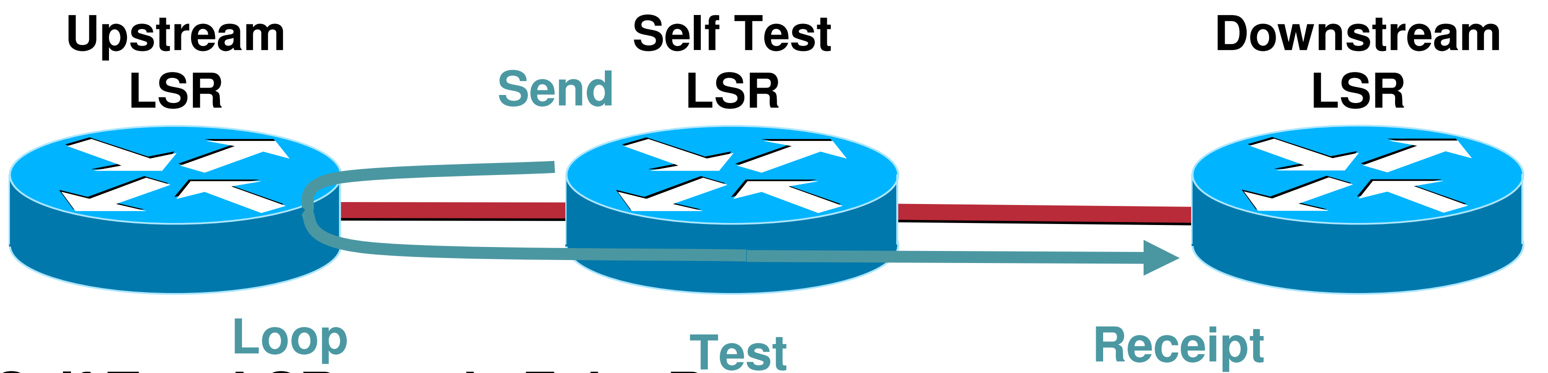
- Semantics are simple
- Label applies to a particular interface
- Pop label
- Forward out advertised interface

Initiation details

Cisco.com

- Pick an interface and label to be tested
- Pick addresses so that ECMP should forward to Downstream LSR
- Record Downstream LSR, outgoing interface and label stack
- Affix label, set TTL=2, affix loopback label

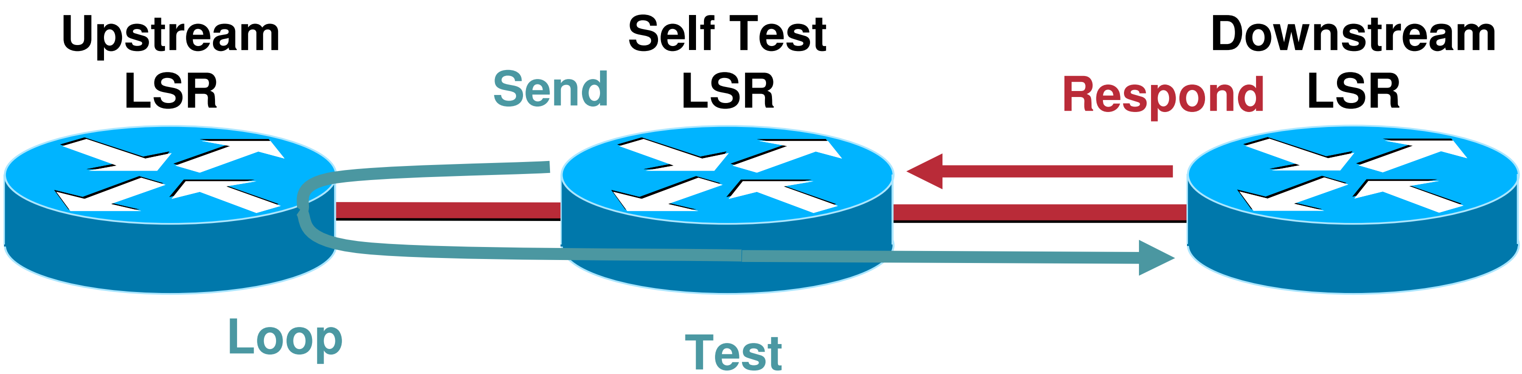
Echo Request



- Self Test LSR sends Echo Request
- Looped through dataplane of Upstream LSR
- TTL is not decremented
- Flows through dataplane of Self Test LST
- TTL-expired causes receipt at Downstream LSR

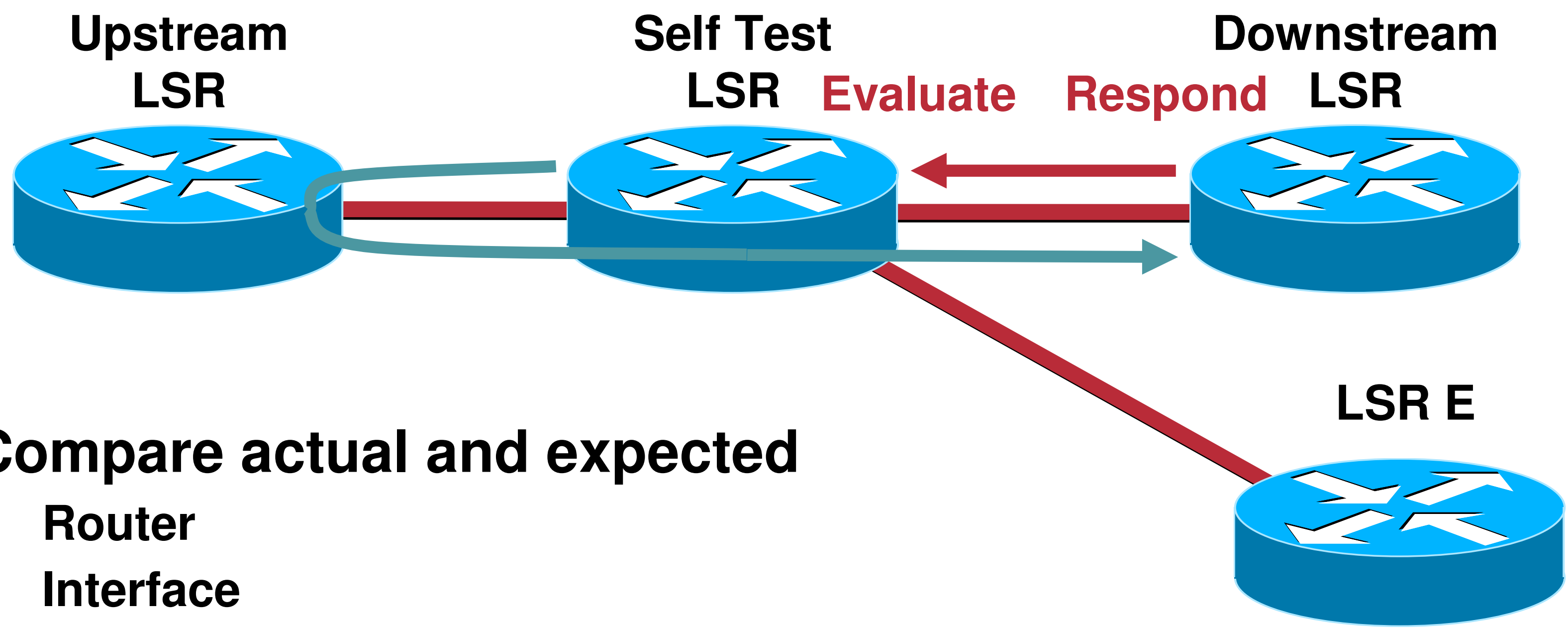
Downstream LSR Response

Cisco.com



- Format Echo Reply
- Include incoming interface & label stack
- Send

Self Test Evaluation



- **Compare actual and expected**
  - Router
  - Interface
  - Label stack
- **On error notify network management**
  - Other automated responses possible

Bidirectional Forwarding Detection

- Simple, fixed-field, hello protocol
- Nodes transmit BFD packets periodically over respective directions of a path
- If a node stops receiving BFD packets some component of the bidirectional path is assumed to have failed
- Several modes of operation

BFD Control Packet


```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Vers |  Diag   |H|D|P|F| Rsvd  |  Detect Mult  |    Length     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       My Discriminator                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Your Discriminator                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Desired Min TX Interval                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Required Min RX Interval                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Required Min Echo RX Interval                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Variable detection intervals

- Each node estimates how quickly it can send and receive BFD packets
- Nodes exchange the follow parameters in every control packet
  - Desired Min TX Interval
  - Required Min RX Interval
  - Detect Multiplier
- These estimates can be modified in real time in order to adapt to unusual situations

Determining Detection Time

TX – Transmission Interval

RX – Receive Interval

Note that TX(a->b) = RX(b->a)

TX(a->b) = max(Desired Min TX(a), Required Min RX(b))

TX(b->a) = max(Desired Min TX(a), Required Min RX(b))

Detection Time(b) = Detect Mult(a) x T(a->b)

TX is jittered by 25%

Diagnostics

0 -- No Diagnostic

1 -- Control Detection Time Expired (RDI)

2 -- Echo Function Failed (N/A to VCCV)

3 -- Neighbor Signaled Session Down (FDI)

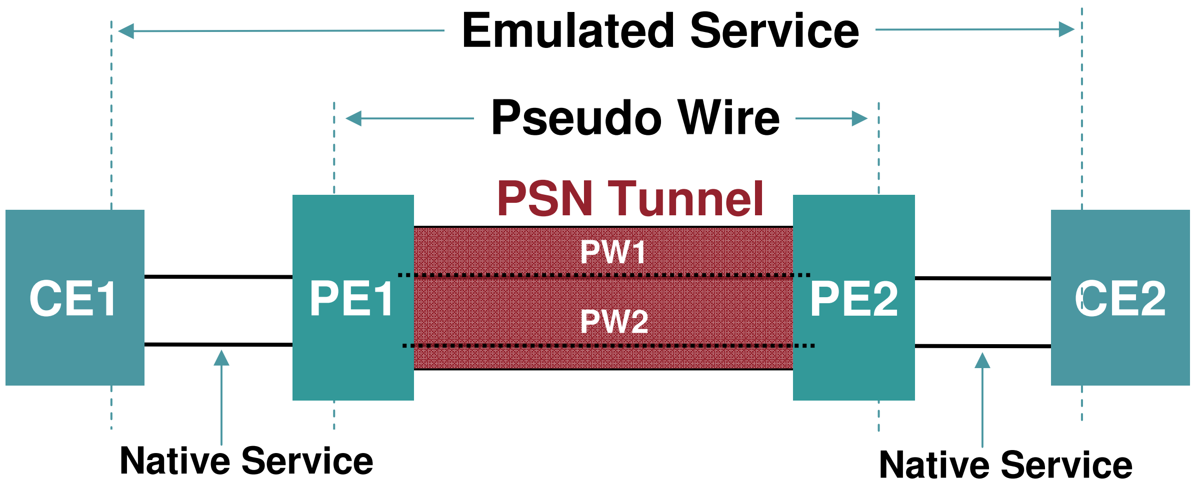
4 -- Forwarding Plane Reset (Indicates local equipment failure)

5 -- Path Down (Alarm Suppression)

6 -- Concatenated Path Down (used to propagate access link alarms)

7 -- Administratively Down

Virtual Circuit Connection Verification (VCCV)



- **Multiple PSN Tunnel Types**

  MPLS, IPSEC, L2TP, GRE,…

- **Motivation**

  One tunnel can serve many pseudo-wires.

  MPLS LSP ping is sufficient to monitor the PSN tunnel (PE-PE connectivity), but not VCs inside of tunnel.

VCCV Overview

- Mechanism for connectivity verification of PW
- Features
  - Works over MPLS or IP networks
  - In-band CV via control word flag or out-of-band option by inserting router alert label between tunnel and PW labels
  - Works with BFD, ICMP Ping and/or LSP ping
- VCCV results may drive OAM/LMI injection on corresponding AC(s)
- http://www.ietf.org/internet-drafts/draft-ietf-pwe3-vccv-02.txt

In Band VCCV Format

Cisco.com

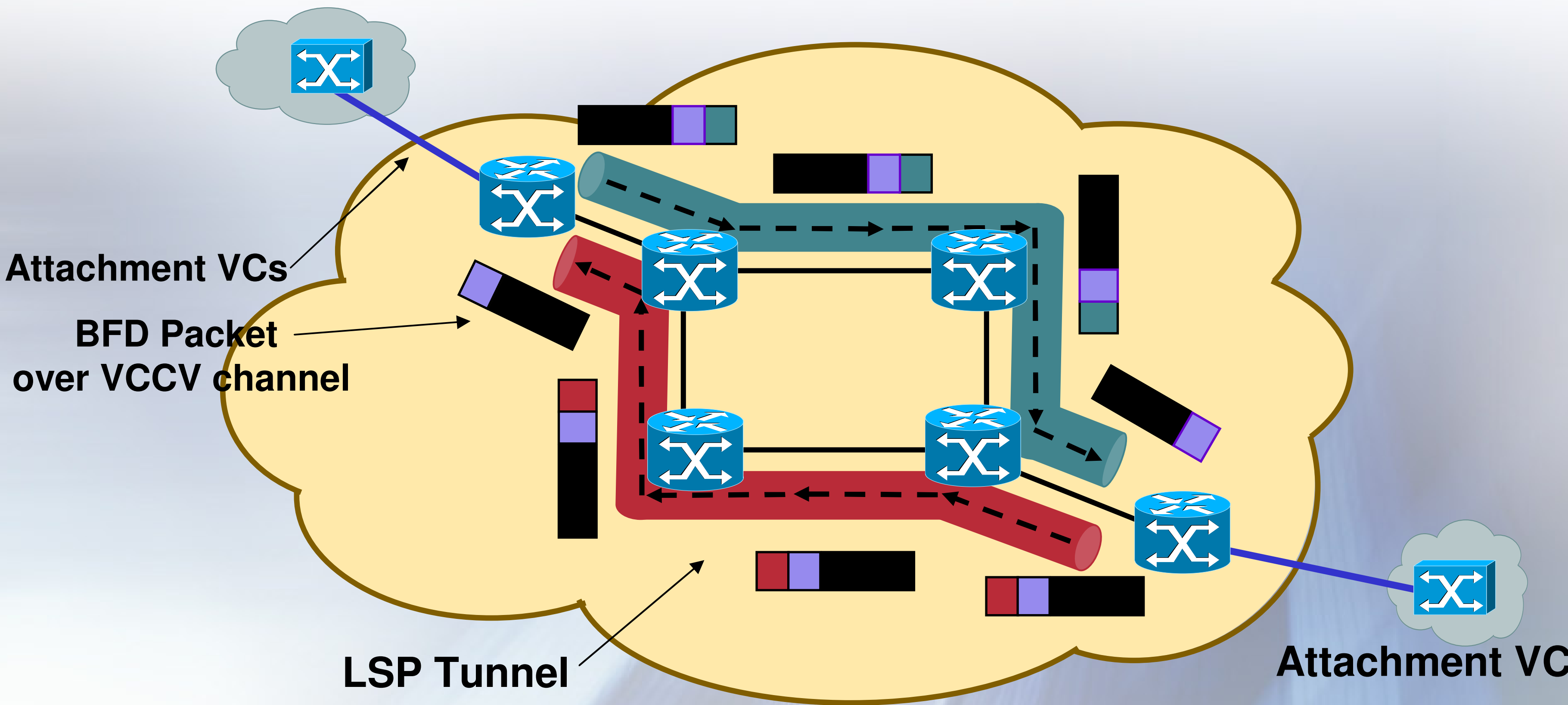
Control word use is signaled in LDP - Standard form:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0 0 0 0| Flags |FRG|  Length   | Sequence Number               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

OAM uses a different 1st nibble

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|0 0 0 1|  reserved             | PPP DLL Protocol Number=IPvX  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             IP OAM Packet: Ping / BFD / LSP Ping              |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

PWE3 OAM Example: Continuity Verification



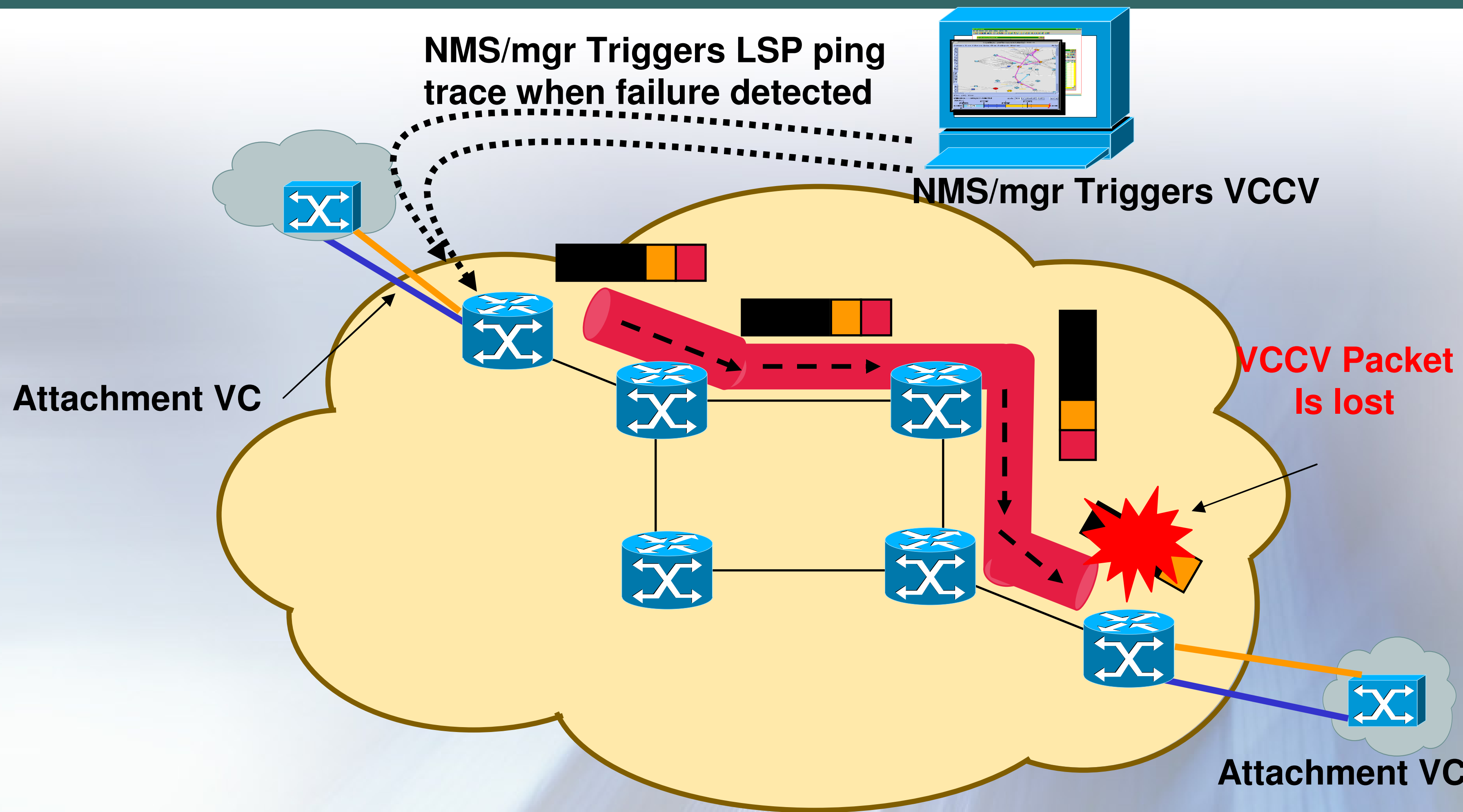
- BFD provides a lightweight means of regular periodic CV

SLA Monitoring / Verification

Cisco.com

- The OAM CV function can be extended for SLA measurement
- Measure quantity of OAM packets at each end of PW
- Timestamps in Ping, LSP Ping

Example of Operation
CV/Trace Using VCCV and LSP Ping

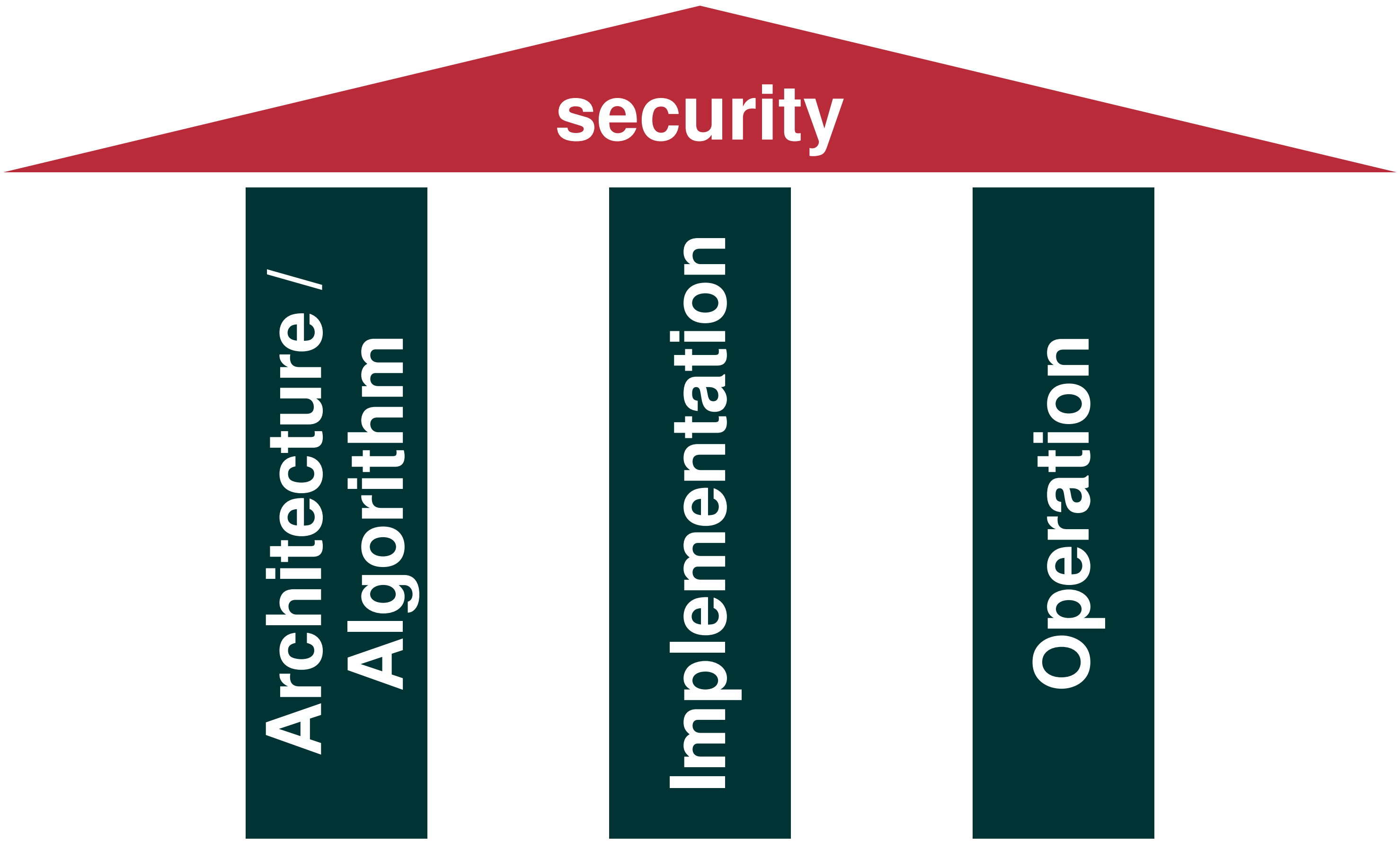


MPLS Security Considerations

Monique Morrow

Three Pillars of Security

Cisco.com



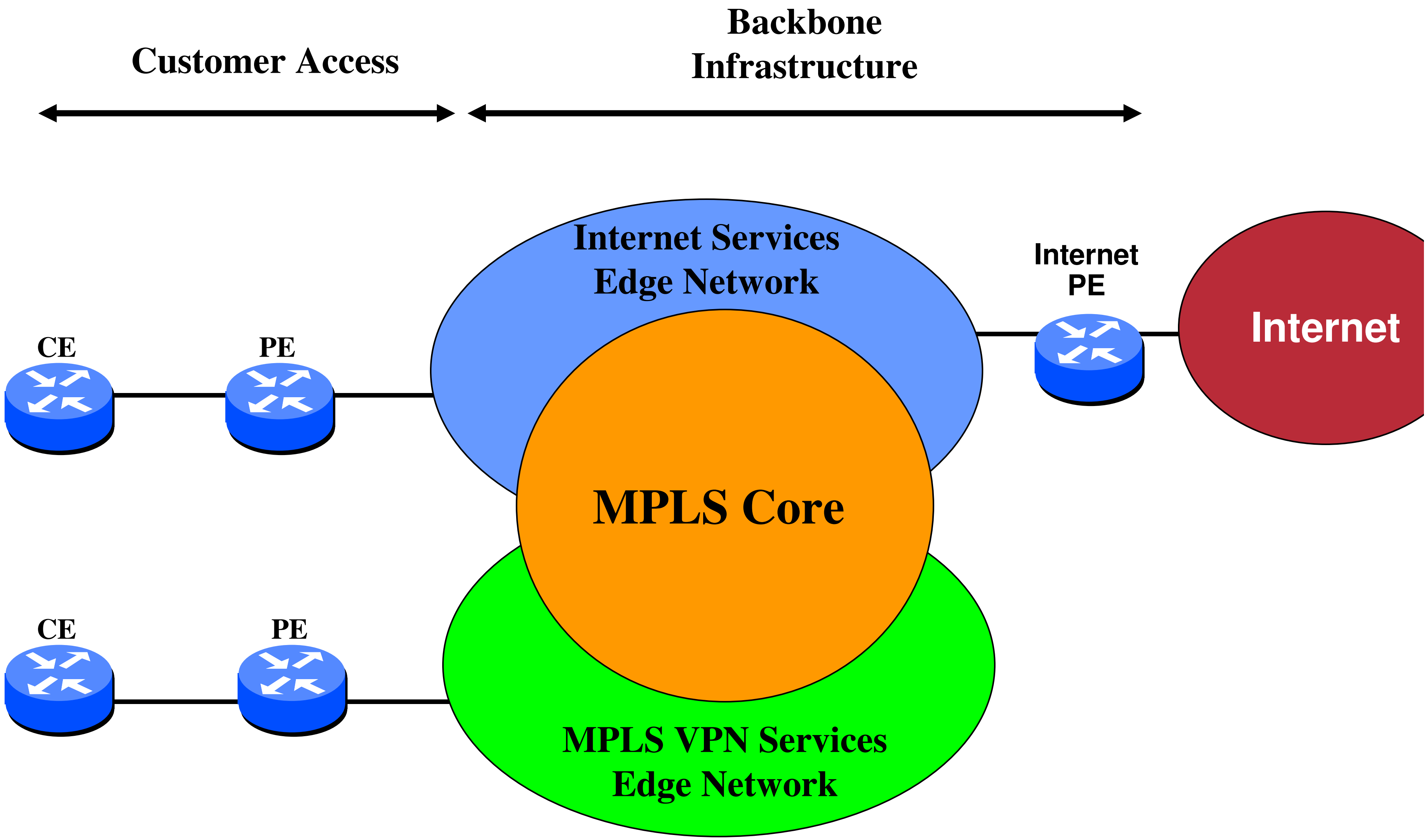
Break one, and all security is gone!

What Kind of Threats?

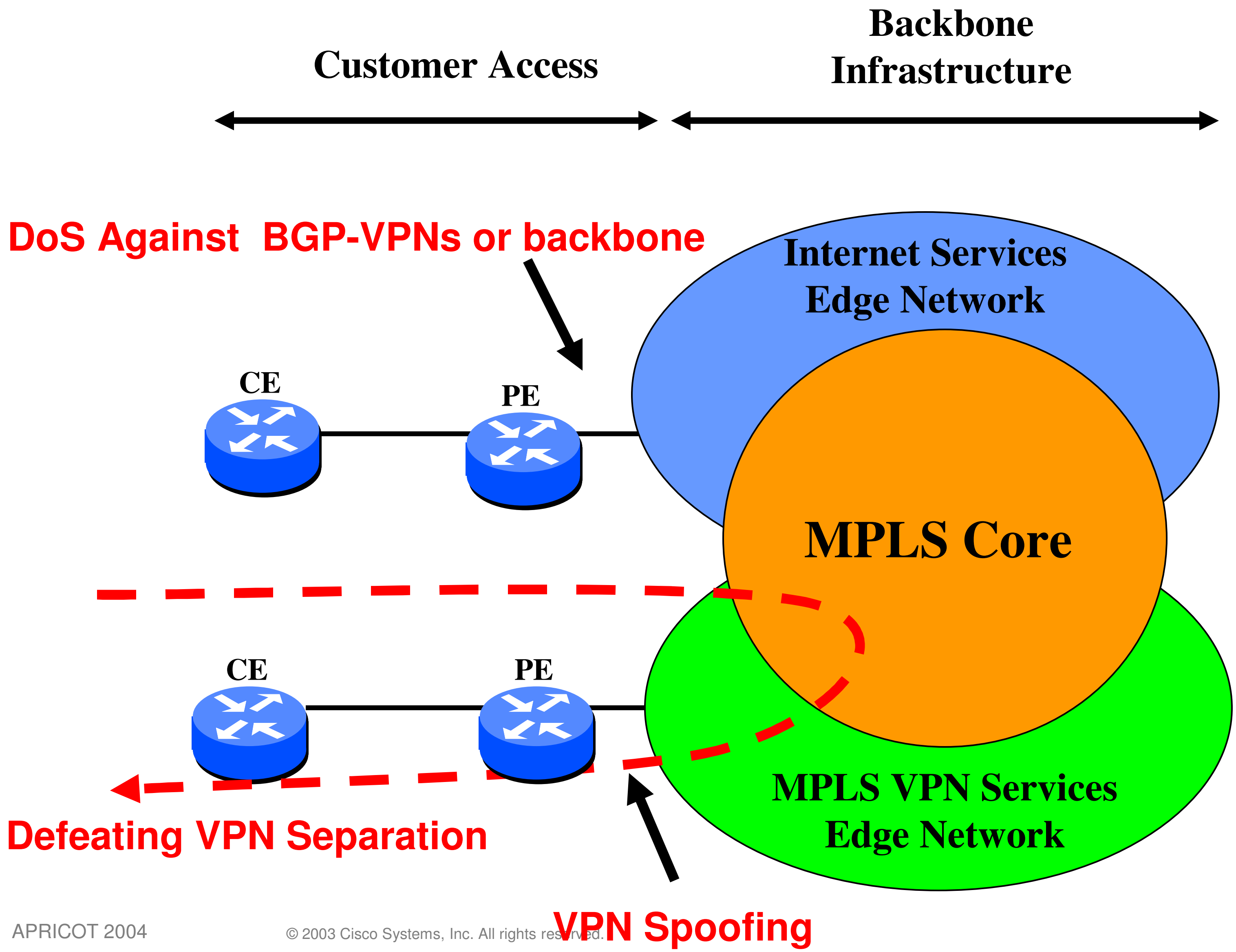
- **Threats from Outside the Backbone**
  - From VPN customers
  - From the Internet
- **Threats from Inside the Backbone**
  - SP misconfigurations (error or deliberate)
  - Hacker "on the line" in the core
- **Threats that are independent of MPLS**
  - Customer network security

Reference model for best practice deployments

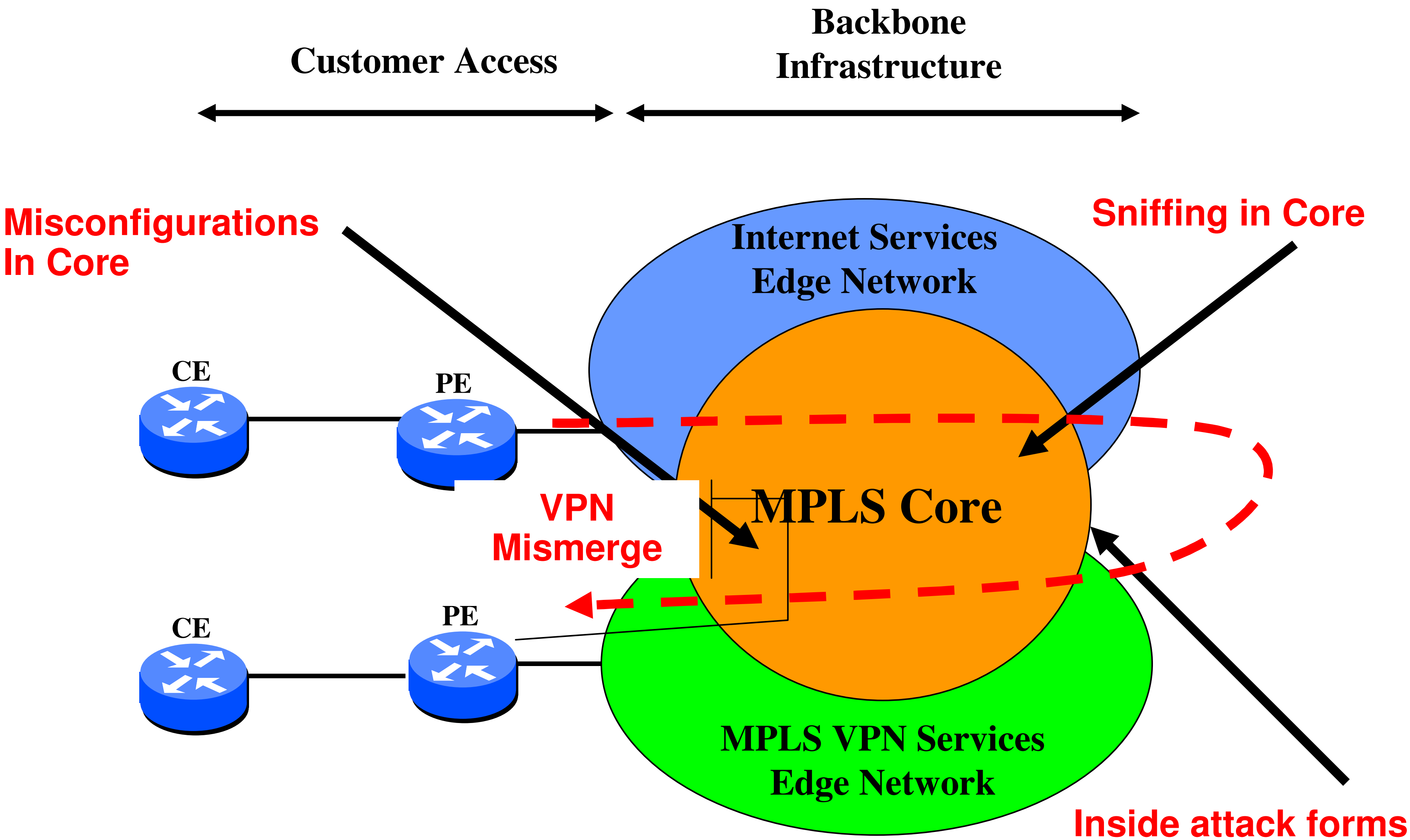
Threat Points of References



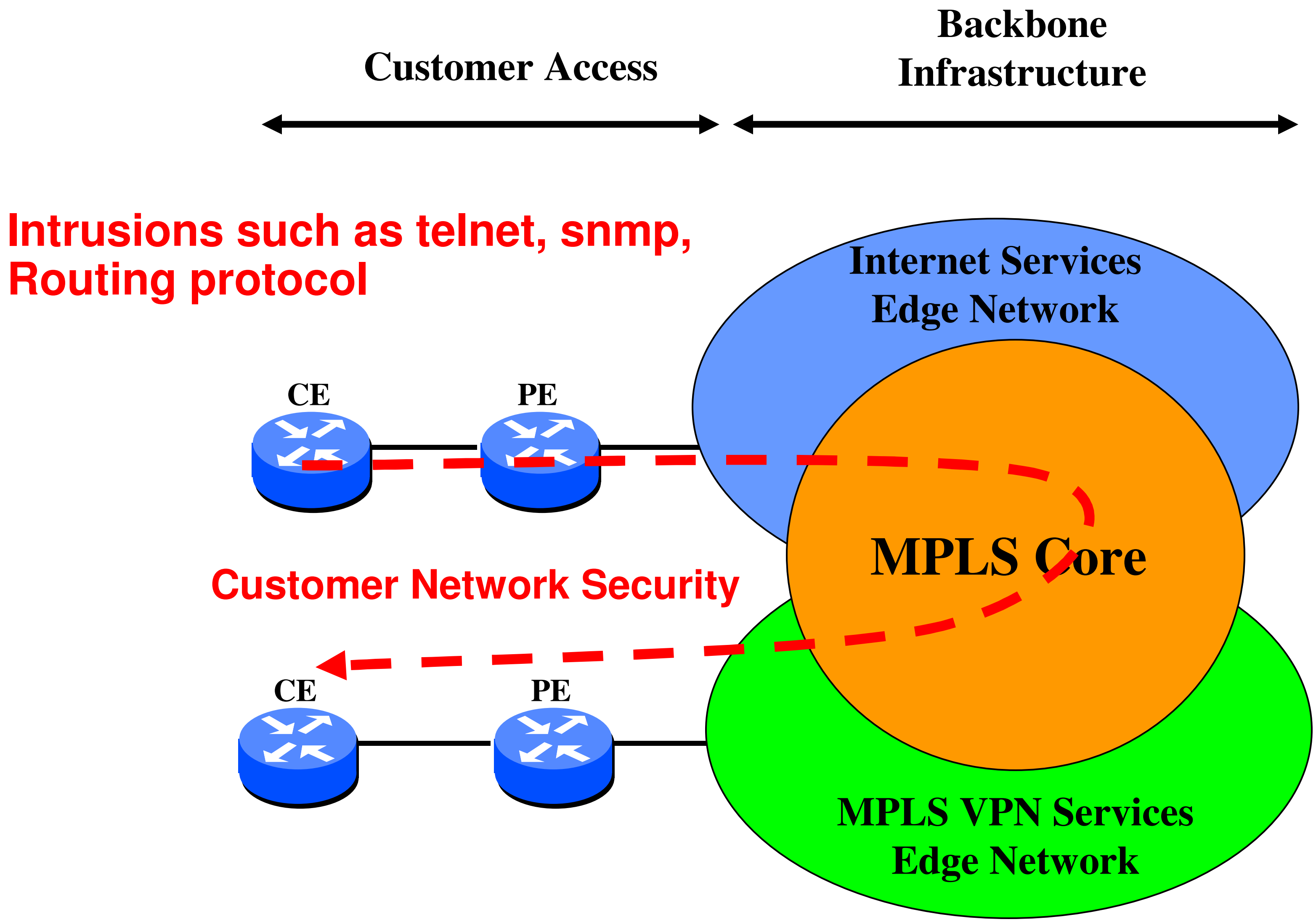
Outside Backbone



Inside the Backbone



Threats Independent of MPLS



Ways to Attack

- “Intrusion”: Get un-authorised access

  Theory: Not possible (as shown before)

  Practice: Depends on:

  - Vendor implementation

  - Correct config and management

No Trust?

Use IPsec between CEs!

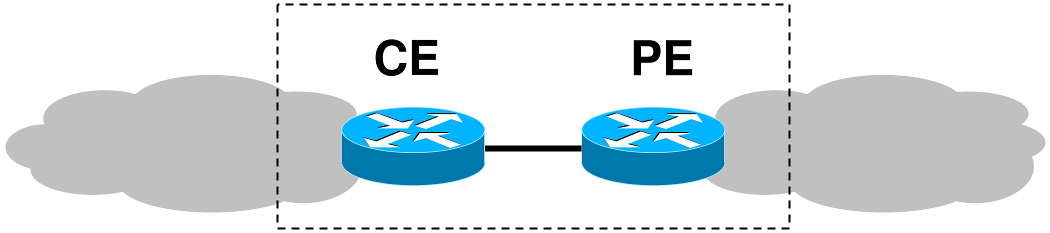
- “Denial-of-Service”: Deny access of others

  Much more interesting…

DoS against MPLS

Cisco.com

- **DoS is about Resource Starvation, one of:**
  - **Bandwidth**
  - **CPU**
  - **Memory (buffers, routing tables, …)**
- **In MPLS, we have to examine:**



- **Rest is the same as in other networks**

Attacking a CE from MPLS (other VPN)

Cisco.com

- **Is the CE reachable from the MPLS side?**

  -> only if this is an Internet CE, otherwise not!
  (CE-PE addressing is part of VPN!)

- **For Internet CEs:**

  Same security rules apply as for any other access router.

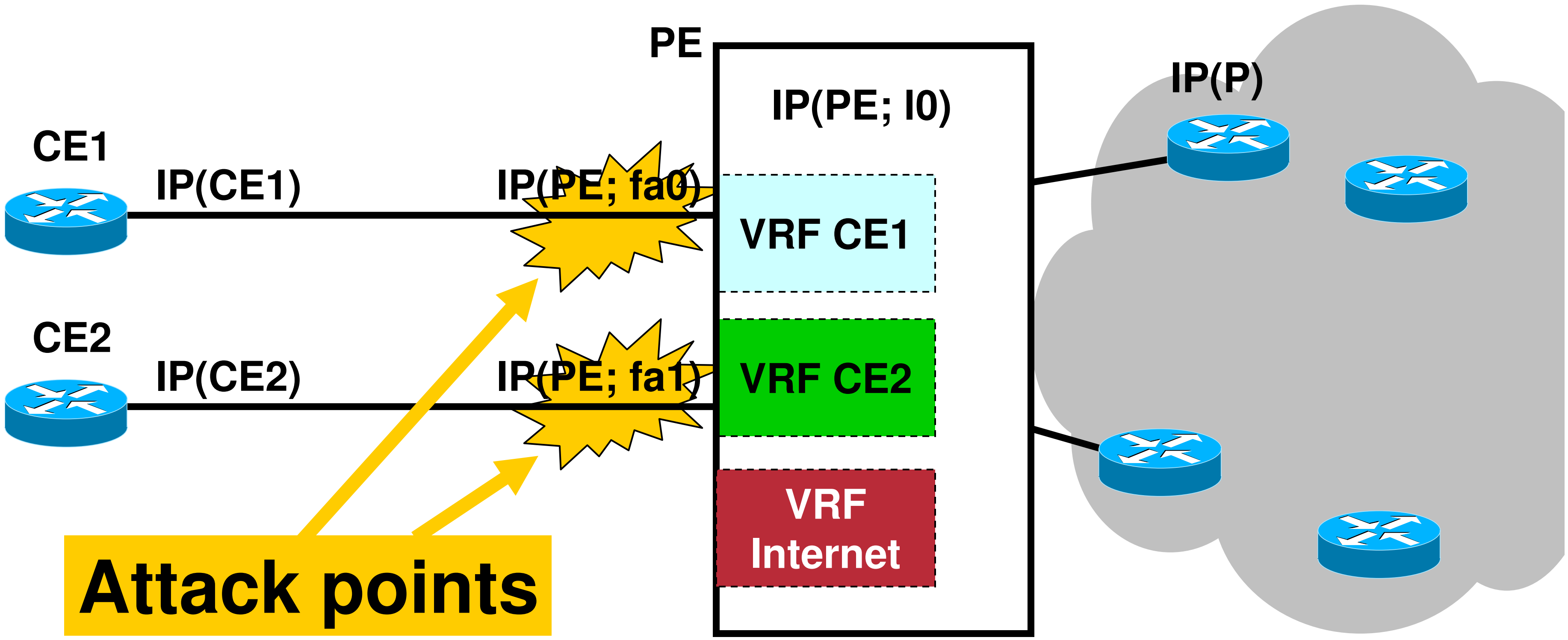
MPLS hides VPN-CEs: Secure!
Internet CEs: Same as in other networks

Attacking a CE-PE Line

- Also depends on reachability of CE or the VPN behind it
- Only an issue for Lines to Internet-CEs
  - Same considerations as in normal networks
- If CE-PE line shared (VPN and Internet):
  - DoS on Internet may influence VPN! Use CAR!

MPLS hides VPN-CEs: Secure!
Internet CEs: Same as in other networks

Attacking a PE Router



Only visible: "your" interface and interfaces of Internet CEs

DoS Attacks to PE can come from:

- Other VPN, connected to same PE
- Internet, if PE carries Internet VRF

Possible Attacks:

- Resource starvation on PE

Too many routing updates, too many SNMP requests, small servers, …

Has to be secured

Layer 2 Comparison Context

- VPNs delivered via Layer 2 point-to-point connections such as ATM, Frame Relay
- Address and routing separation in MPLS-VPN architecture is equivalent to Layer 2 models
- An MPLS-VPN network is resistant to DoS attacks as a Layer 2 network

Non-IP networks: Not 100% secure!! Example: Telephone Network

"I had access to most, if not all, of the switches in Las Vegas," testified Mitnick, at a hearing of Nevada's Public Utilities Commission (PUC). "I had the same privileges as a Northern Telecom technician."

Source:
http://online.securityfocus.com/news/497

Non-IP networks: Not 100% secure!! Example: ATM Switch

Cisco.com

> “a single 'land' packet sent to the telnet port (23) of either the inband or out-of-band interface will cause the device to stop responding to ip traffic. Over the course of 6-1/2 minutes, all CPU will be consumed and device reboots.”

Source: Bugtraq, 15 June 2002: “Fore/Marconi ATM Switch 'land' vulnerability”, by seeker_sojourn@hotmail.com;

Comparison with ATM / FR

	ATM/FR	MPLS
Address space separation	yes	yes
Routing separation	yes	yes
Resistance to attacks	yes	yes
Resistance to Label Spoofing	yes	yes
Direct CE-CE Authentication (layer 3)	yes	with IPsec

From RFC2547bis: Data Plane Protection

1. a backbone router does not accept labeled packets over a particular data link, unless it is known that that data link attaches only to trusted systems, or unless it is known that such packets will leave the backbone before the IP header or any labels lower in the stack will be inspected, and …

- Inter-AS should *only* be provisioned over secure, private peerings
- Specifically NOT: Internet Exchange Points (anyone could send labelled packets!! No filtering possible!!)

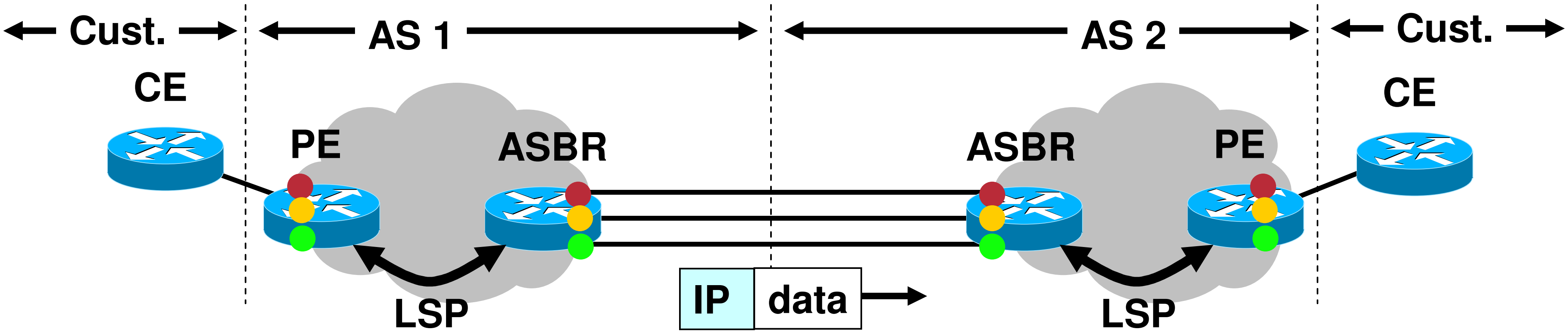
From RFC2547bis: Control Plane Protection

2. labeled VPN-IPv4 routes are not accepted from untrusted or unreliable routing peers,

- Accept routes with labels only from trusted peers
- Plus usual BGP filtering (see ISP Essentials*)

Inter-AS: Case 10.a)
VRF-VRF back-to-back

Cisco.com



- Control plane: No signalling, no labels
- Data plane: IPv4 only, no labels accepted
- Security: as in 2547
- Customer must trust both SPs

Security of Inter-AS 10.a)

- **Static mapping**

  SP1 does not “see” SP2’s network

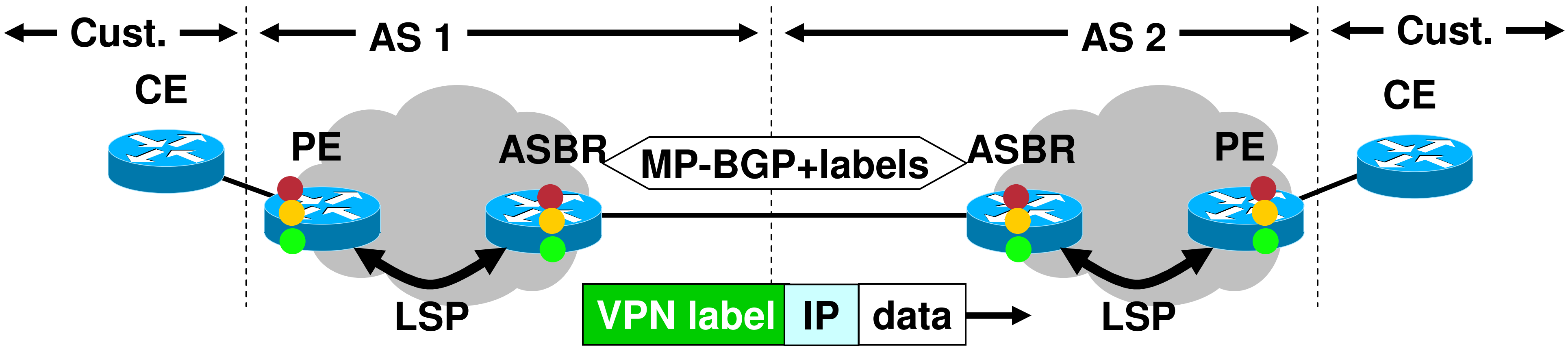
  And does not run routing with SP2, except within the VPNs.

  → Quite secure

- **Potential issues:**

  SP 1 can connect VPN connection wrongly
  (like in ATM/FR)

Inter-AS: Case 10.b)
ASBR exchange labelled VPNv4 routes



- Control plane: MP-BGP, labels
- Data plane: Packets with one label
- AS1 can insert traffic into any shared VPN of AS2
- Customer must trust both SPs

Security of Inter-AS 10.b)

- **ASBR1 does signalling with ASBR2**

  MP-BGP: has to be secured, dampening etc

  Otherwise no visibility of the other AS
  (ASBR1 – ASBR2 is the only interface between the SPs.)

- **Potential Issues:**

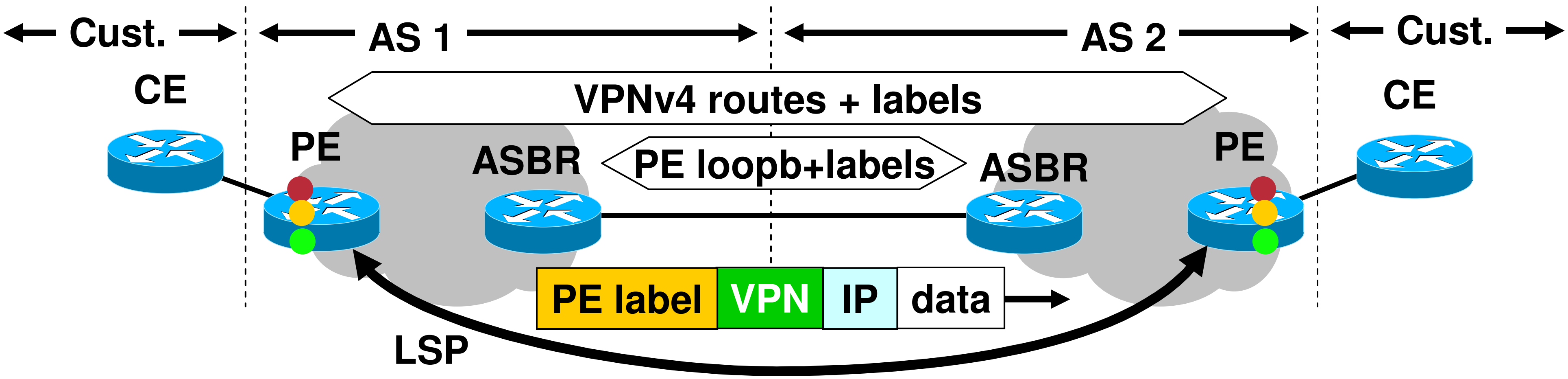
  SP1 can bring wrong CEs into any shared VPN

  SP1 can send packets into any shared VPN (not into VPNs that are not shared, since label is checked);

  → SP can make any shared VPN insecure

**Watch layer-2 security!!
(more later)**

Inter-AS: Case 10.c)
ASBRs exchange PE loopbacks



- Control plane: ASBR: just PE loopback + labels; PE/RR: VPNv4 routes + labels
- Data plane: PE label + VPN label
- AS1 can insert traffic into VPNs in AS2
- Customer must trust both SPs

Security of Inter-AS 10.c)

- ASBR-ASBR signalling (BGP)
  RR-RR signalling (MP-BGP)
  - Much more “open” than 10.a) and 10.b)
  - LSPs between PEs, BGP between RR, ASBR
- Potential Issues:
  - SP1 can bring a CE into any VPN on “shared” PEs
  - SP1 can intrude into any VPN on “shared” PEs
- Very open architecture
  - probably only applicable for ASes controlled by the same SP.

Watch layer-2 security!!
(more later)

Inter-AS Summary and Recommendation

Cisco.com

- Three different models for Inter-AS
  - Different security properties
  - Most secure: Static VRF connections (10.a), but least scalable
- Basically the SPs have to trust each other
  - Hard / impossible to secure against other SP in this model
- Okay if all ASes in control of one SP
- Current Recommendation: Use 10.a)

Inter-AS Recommendation

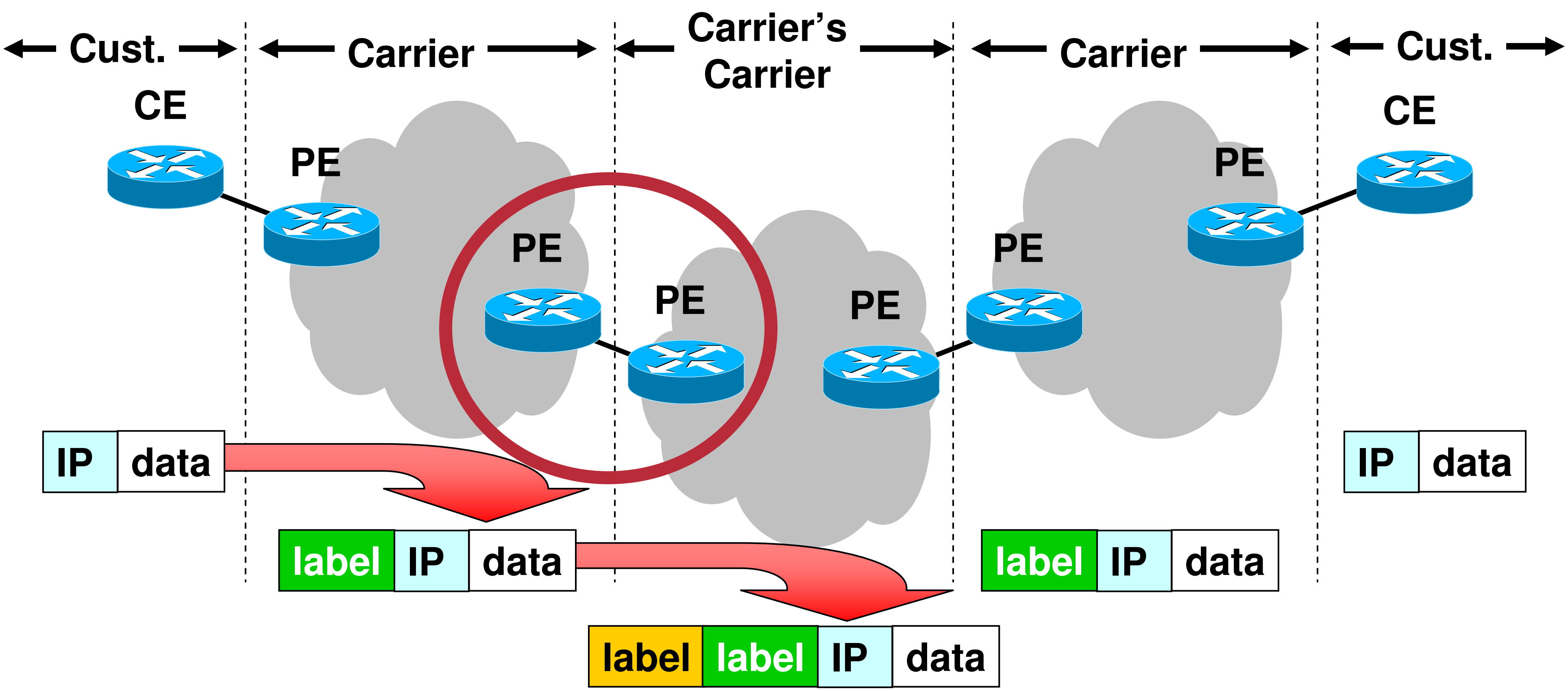
Cisco.com

- Start with 10.a) (static VPN connections)

  Not many Inter-AS customers yet anyway → Easy start

- Maybe at some point (when many Inter-AS customers), move to 10.b) (ease of provisioning)
- 10.c) felt by most SPs as too open. Current recommendation: Only when both ASes under one common control

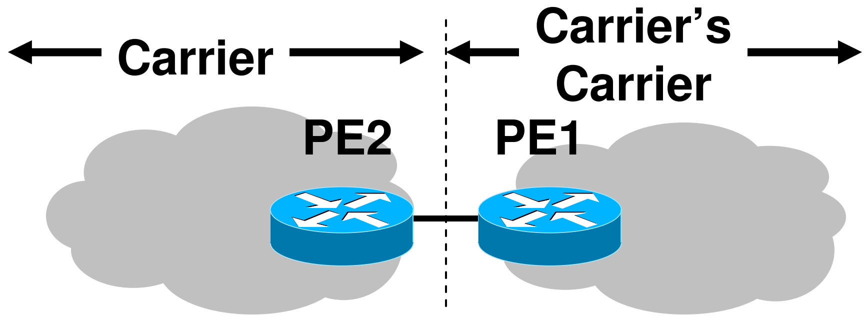
Carrier's Carrier



- Same principles as in normal MPLS
- Customer trusts carrier who trusts carrier

Carrier's Carrier: The Interface

Cisco.com



- **Control Plane:**

  PE1 assigns label to PE2

- **Data Plane:**

  PE1 only accepts packets with this label on this i/f

  →PE1 controls data plane

  →No label spoofing possible

Watch layer-2 security!! (more later)

Carrier's Carrier: Security

- Carrier is a VPN on core Carrier's network
- Cannot spoof other VPN/carrier:
  - PE verifies top label in data path
  - Top label determines egress PE
- Can mess up his own VPN!
- Basically like normal 2547

Carrier’s Carrier: Summary

- **Can be secured well**

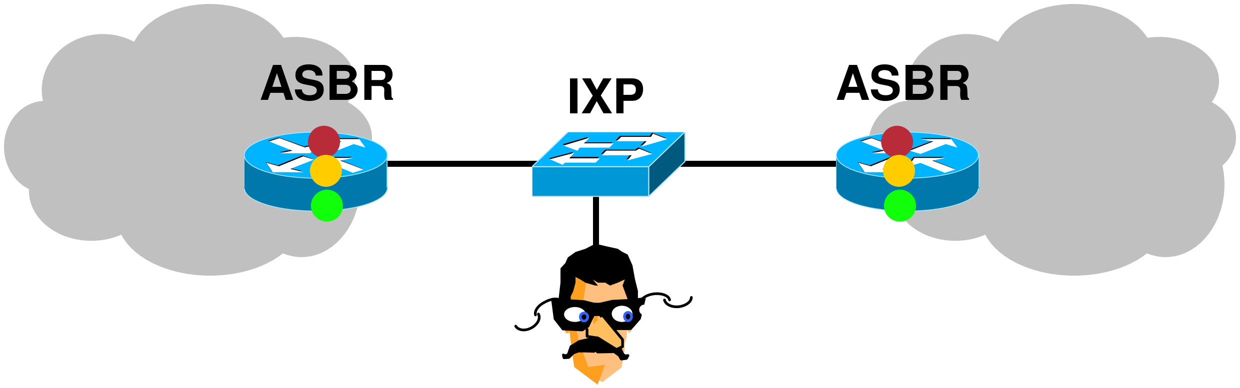
  Carrier has VPN on Carrier’s Carrier MPLS cloud

  Carrier cannot intrude into other VPNs.

  Carrier *can* mess up his own VPN (VPNs he offers to his customers)

- **End customer must trust both SPs.**

Watch out for Layer 2 Security!!



- 3rd party in same VLAN (e.g. IXP) can:

  insert spoofed packets into VPNs
  (cannot be prevented today technically!!)

  Do layer 2 attacks to do man-in-the-middle
  (could be mostly prevented, but is often not done)

Recommendation: Inter-AS and CsC connections only on private peerings!!

VLAN Separation

Cisco.com

- **VLANs can be assumed to be separate, if…**

  … The switch is not low end, very old or has bugs

  … VTP (VLAN trunking protocol) is *disabled* on all ports (this is the default these days)

  … Router ports are not trunk ports

  … No ISL or 802.1q signalling to router port

All this can be done, so assuming correct config, VLANs are separate

But….

Within (!) a VLAN, Attacks are Easy!!

Cisco.com

1. ARP spoofing (hacking tool *hunt, arpspoof)*
2. CAM overflow (hacking tool *macof*)
3. DoS against spanning tree
4. DoS storms (hacking tool exists)

Solutions:

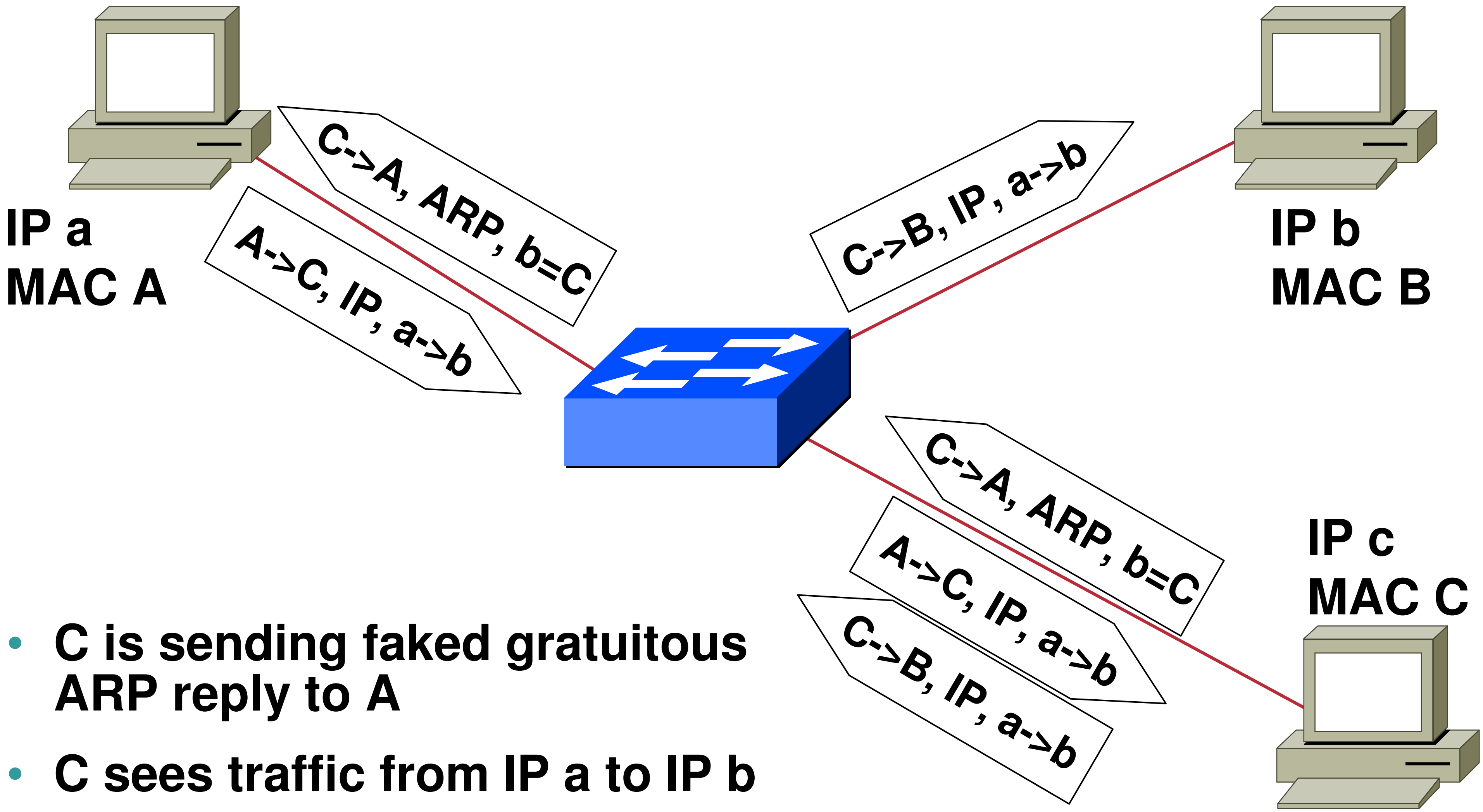
- For 1 and 2: port security (hard to maintain…)

  Few SPs do this normally, so this attack is easy

- Disable Spanning Tree on router port, hard code Root Bridge

ARP Spoofing

Cisco.com



- C is sending faked gratuitous ARP reply to A
- C sees traffic from IP a to IP b

Arpspoof in Action

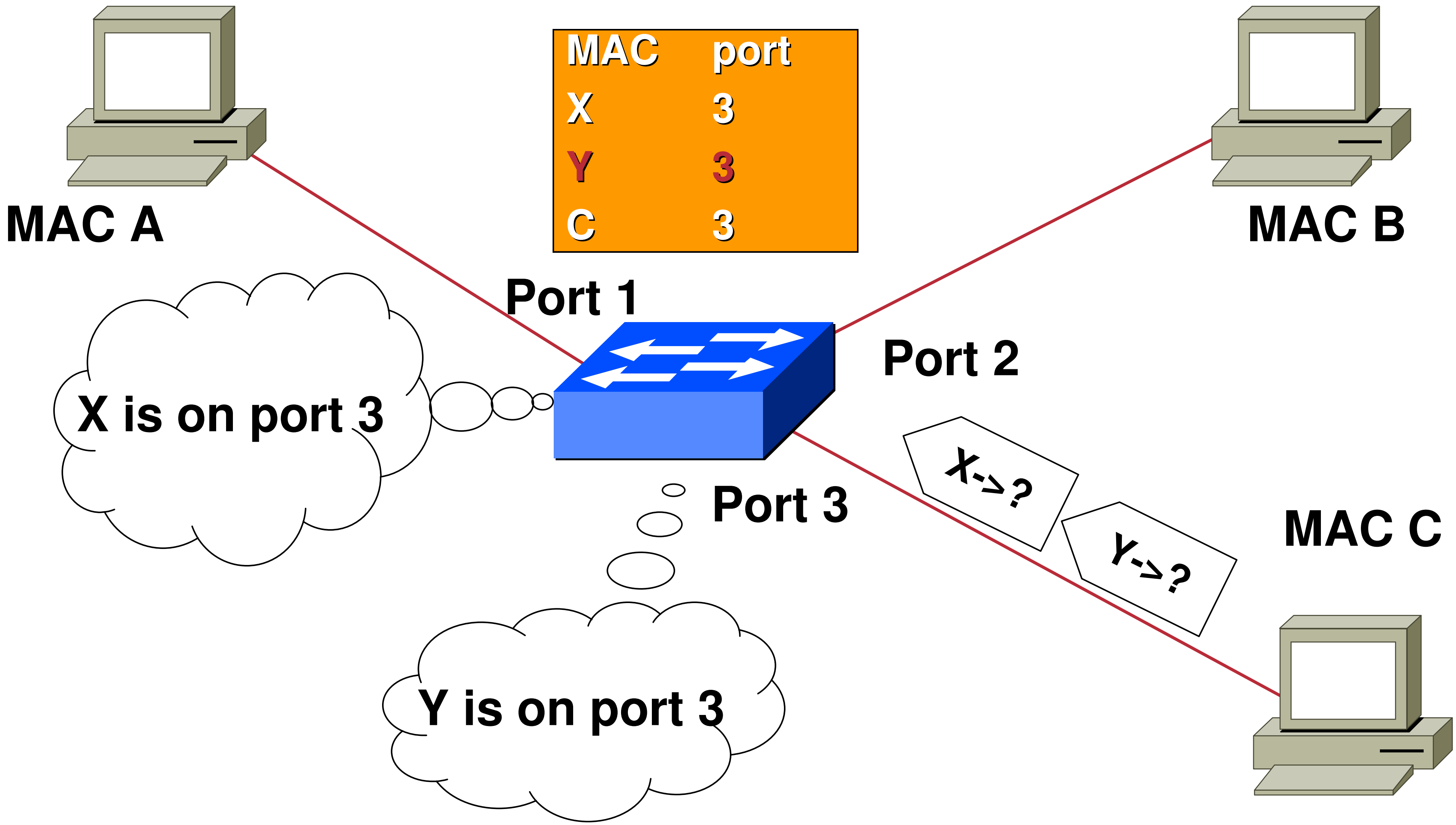

```
[root@hacker-lnx dsniff-2.3]# ./arpspoof 15.1.1.1
0:10:83:34:29:72 ff:ff:ff:ff:ff:ff 0806 42: arp
reply 15.1.1.1 is-at 0:10:83:34:29:72
```

```
C:\>test

C:\>arp -d 15.1.1.1

C:\>ping -n 1 15.1.1.1

Pinging 15.1.1.1 with 32 bytes of data:

Reply from 15.1.1.1: bytes=32 time<10ms TTL=255

C:\>arp -a

Interface: 15.1.1.26 on Interface 2
  Internet Address      Physical Address      Type
  15.1.1.1              00-04-4e-f2-d8-01     dynamic
  15.1.1.25             00-10-83-34-29-72     dynamic
C:\>arp -a

Interface: 15.1.1.26 on Interface 2
  Internet Address      Physical Address      Type
  15.1.1.1              00-10-83-34-29-72     dynamic
  15.1.1.25             00-10-83-34-29-72     dynamic
```

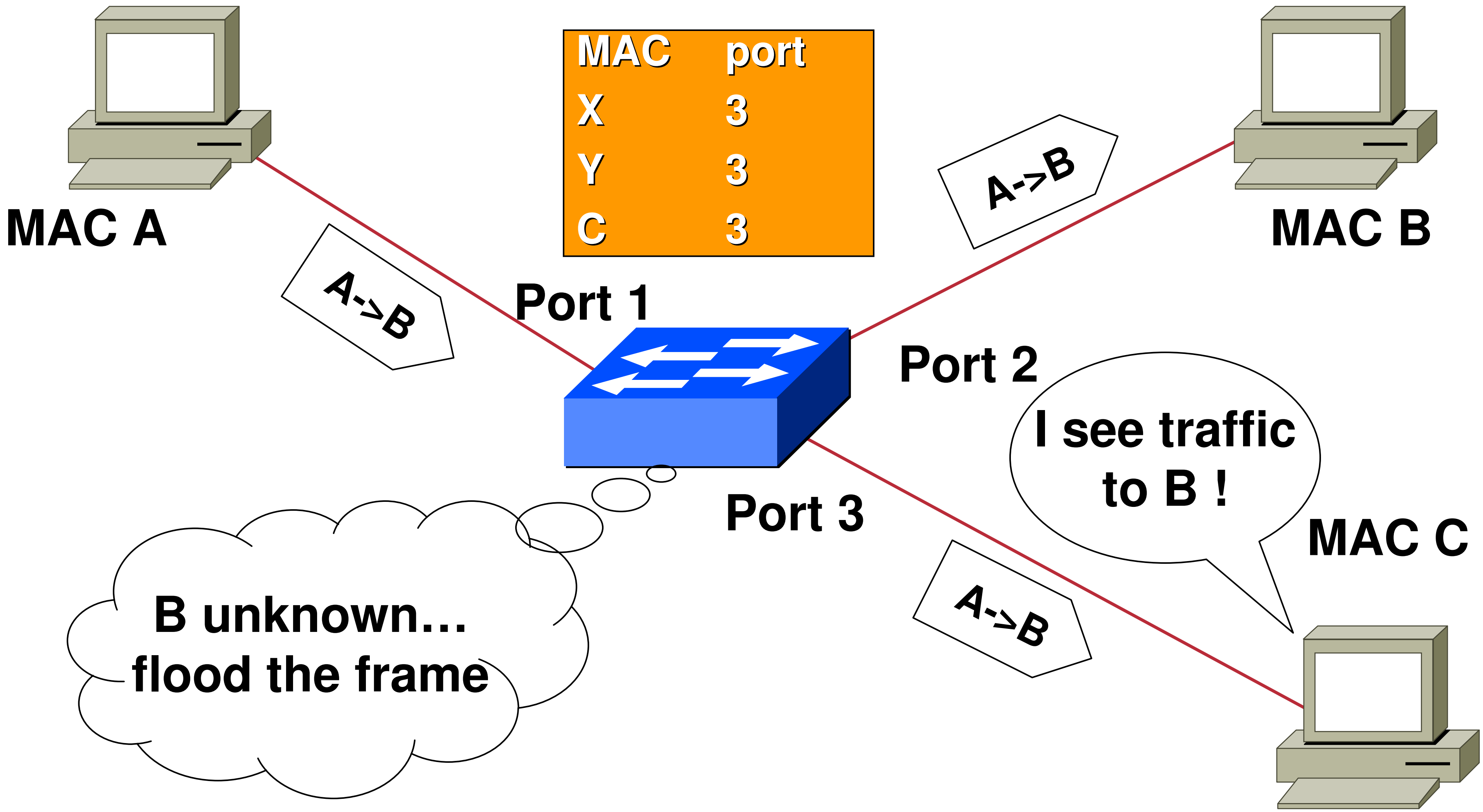
CAM Overflow 1/3

- **theoretical attack until May 1999**
- ***macof* cracker tool since May 1999** *(about 100 lines of perl)*
- **based on the limited size of CAM**

CAM Overflow 2/3



CAM Overflow 3/3



Within (!) a VLAN, Attacks are Easy!!

1. ARP spoofing (hacking tool *hunt, arpspoof)*
2. CAM overflow (hacking tool *macof*)
3. DoS against spanning tree
4. DoS storms (hacking tool exists)

Solutions:

- For 1 and 2: port security (hard to maintain…)

  Few SPs do this normally, so this attack is easy

- For 3 and 4: Disable Spanning Tree on router port, hard code Root Bridge

Labelled packets on a VLAN

Data plane:

- **Any label combination can be sent, by any station in the VLAN**
- **For CsC, top label (LSP) is checked by PE, VPN label cannot be checked, but affects only VPNs from the Carrier (not other carriers).**
- **For Inter-AS, neither LSP label nor VPN label is checked.**

Recommendation for Advanced MPLS Networks

For Inter-AS and CsC (when labeled packets are exchanged) do NOT use a shared VLAN.

Best: Dedicated connection
Second best: Dedicated VLAN

RFC 2547bis states this explicitly!

Best Practice Security Overview (1)

- Secure devices (PE, P): They are trusted!
- Core (PE+P): Secure with ACLs on all interfaces

    Ideal: deny ip any <core-networks>

- Static PE-CE routing where possible
- If routing: Use authentication (MD5)
- Separation of CE-PE links where possible (Internet / VPN)
- LDP authentication (MD5)
- VRF: Define maximum number of routes

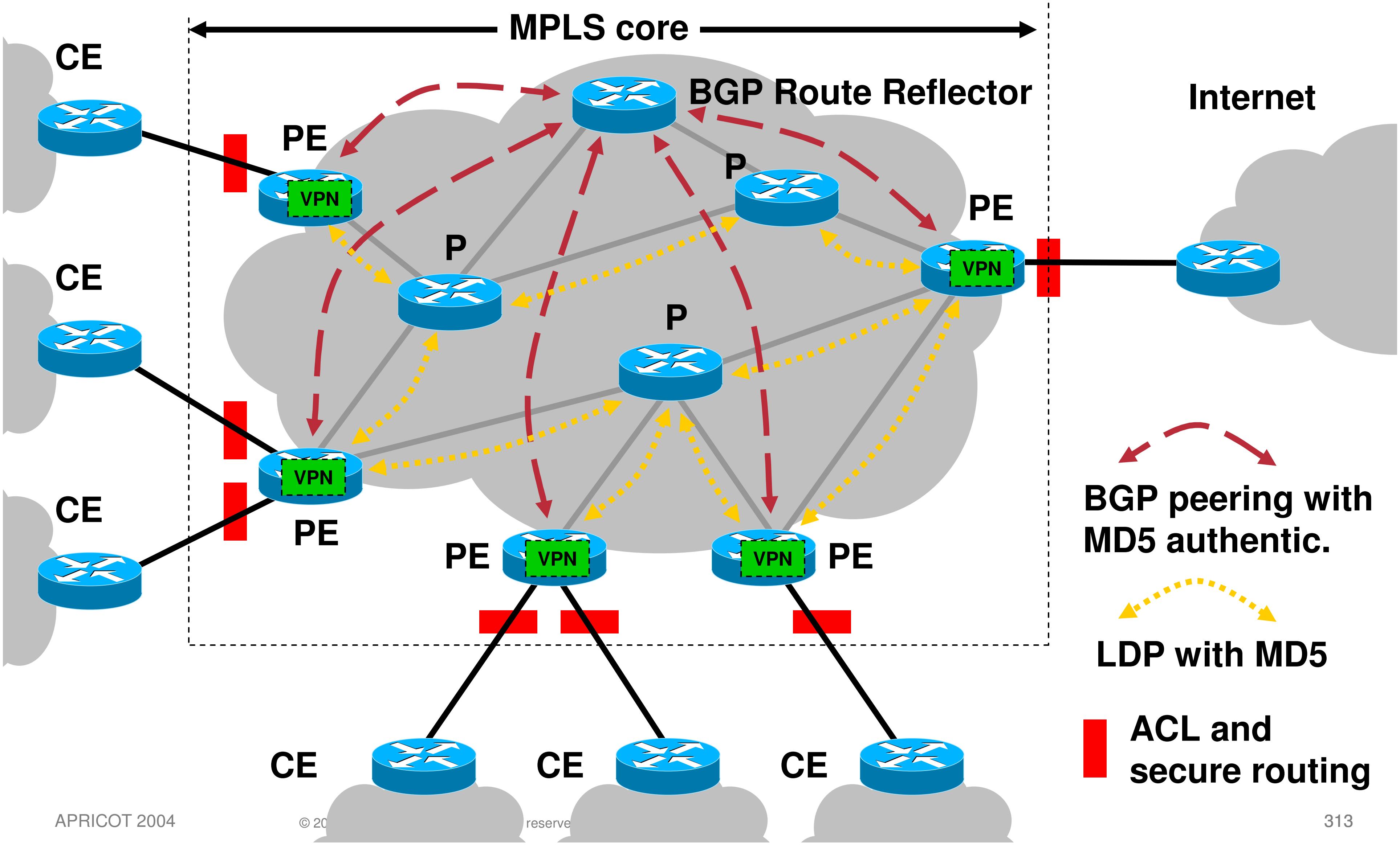
Note: Overall security depends on weakest link!

PE-CE Routing Security

In order of security preference:

1. **Static: If no dynamic routing required (no security implications)**
2. **BGP: For redundancy and dynamic updates (many security features)**
3. **IGPs: If BGP not supported (limited security features)**

Securing the MPLS Core



Neighbour Authentication (1)

- Prevents a router from receiving fraudulent updates from a routing neighbour
- Verifies updates it receives from a label distribution peer
- Support for BGP, ISIS, OSPF, EIGRP, RIPv2 and Label Distribution Protocol (LDP)

Neighbour Authentication (2)

- **PE-CE:** Selected PE-CE routing protocol plus LDP if CsC is enabled. If BGP+labels is being used on CsC, then authentication only on BGP session (no LDP required)

- **PE-PE:** BGP authentication for the secure exchange of VPNv4 routes

- **PE to P and P to P:** Authentication for the backbone routing protocol (IGP) plus LDP

Neighbour Authentication (3)

- Receiving router authenticates source of routing updates
- Two types: Plain text or message digest algorithm 5 (MD5)
- MD5 does not send key; creates message digest by using key and message as hash to MD5
- Resulting message digest exchanged among neighbours

Use IPsec if you need:

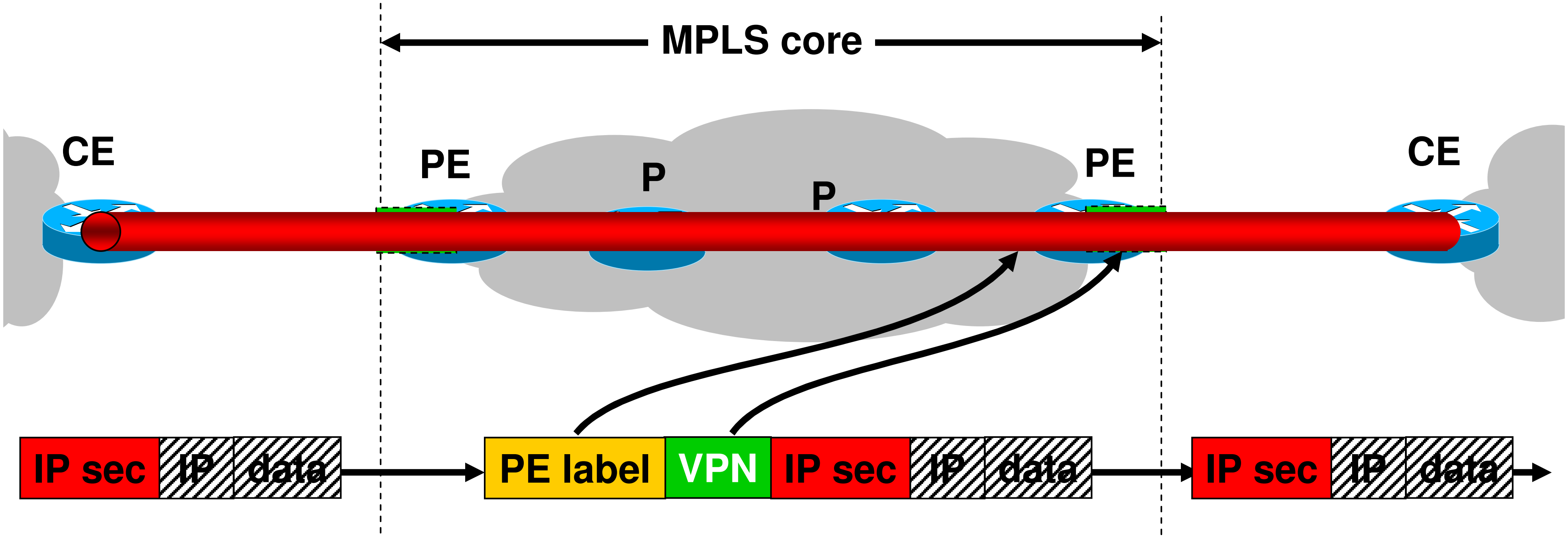
Cisco.com

- Encryption of traffic
- Direct authentication of CEs
- Integrity of traffic
- Replay detection

Maybe more important than encryption?

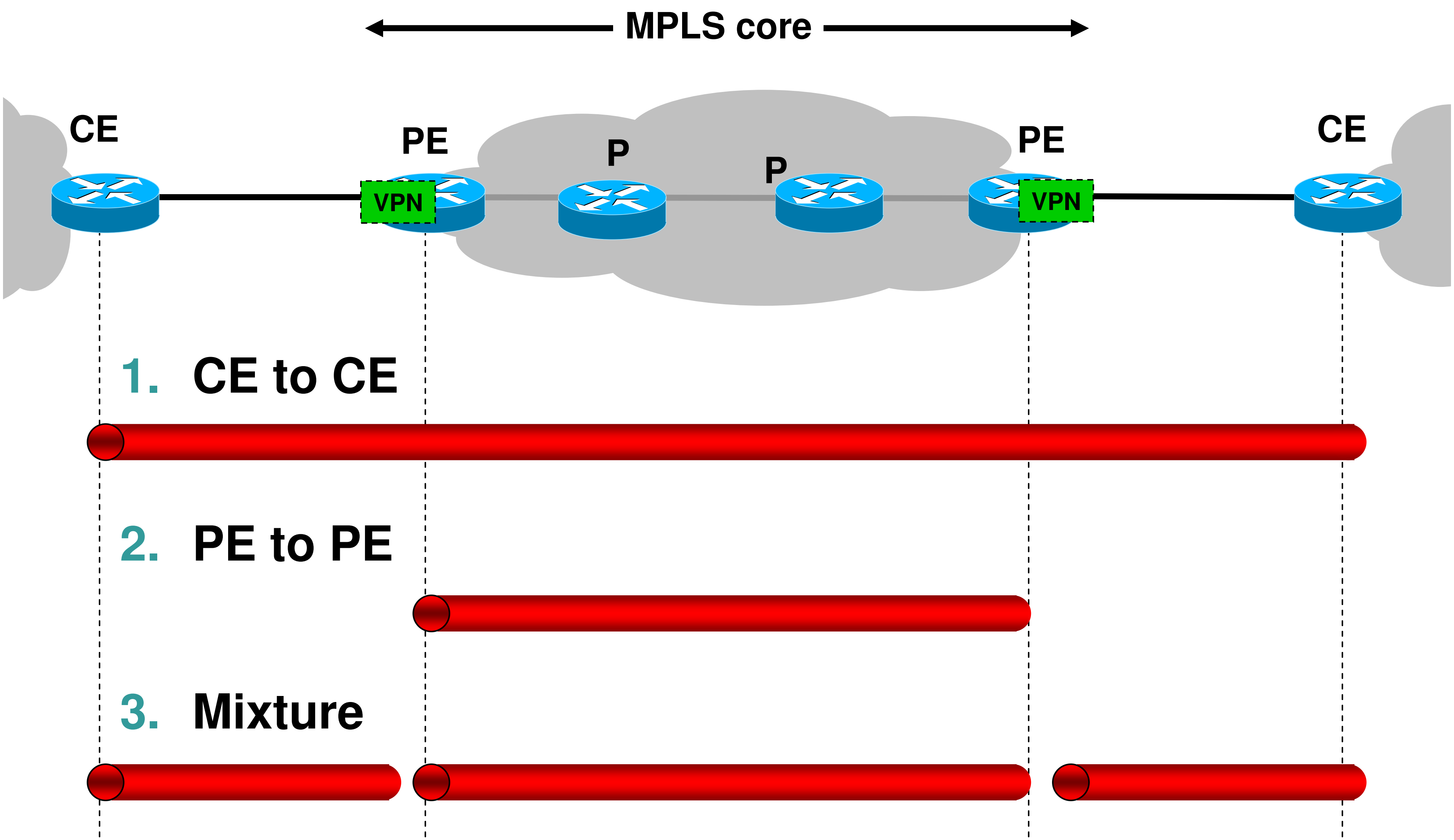
- Or: If you don't want to trust your ISP for traffic separation!

End-to-End Security with IPsec



- Encryption: Data invisible on core
- Authentication: Only known CEs
- Integrity: Data not changed in transit

Where to do IPsec



Where to do IPsec

1. **CE to CE**

   SP not involved (unless manages CEs)

   MPLS network only sees IPsec traffic → Very secure

2. **PE to PE**

   Does not prevent sniffing access line

   → Not very secure for the customer

   There are some specific applications for this (US ILECs)

3. **Mixtures**

   Need to trust SP

   Mostly for access into VPN

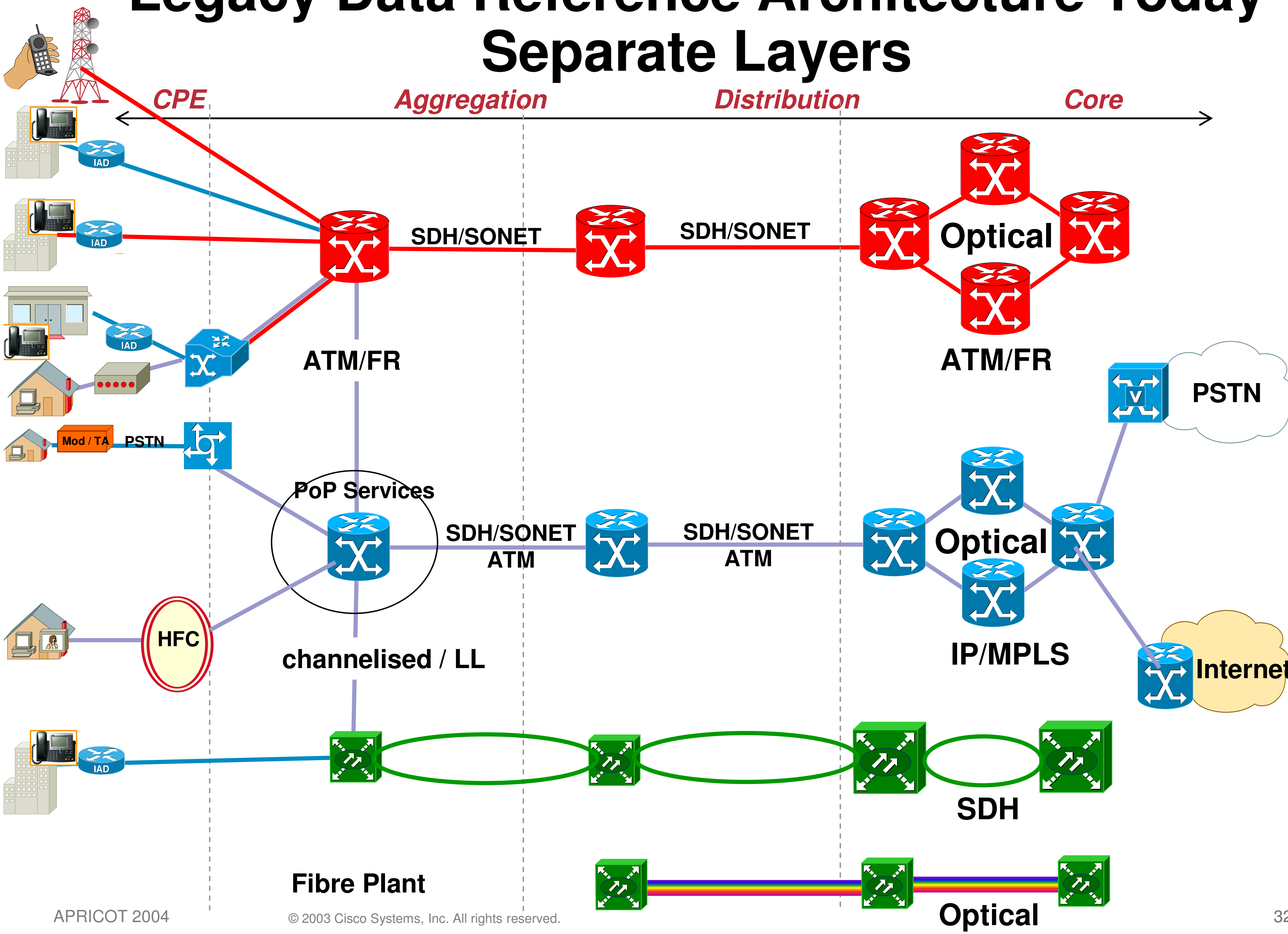
MPLS doesn't provide:

- Protection against
  mis-configurations in the core
- Protection against
  attacks from within the core
- Confidentiality, authentication, integrity, anti-replay
  → Use IPsec if required
- Customer network security

A Word About G-MPLS

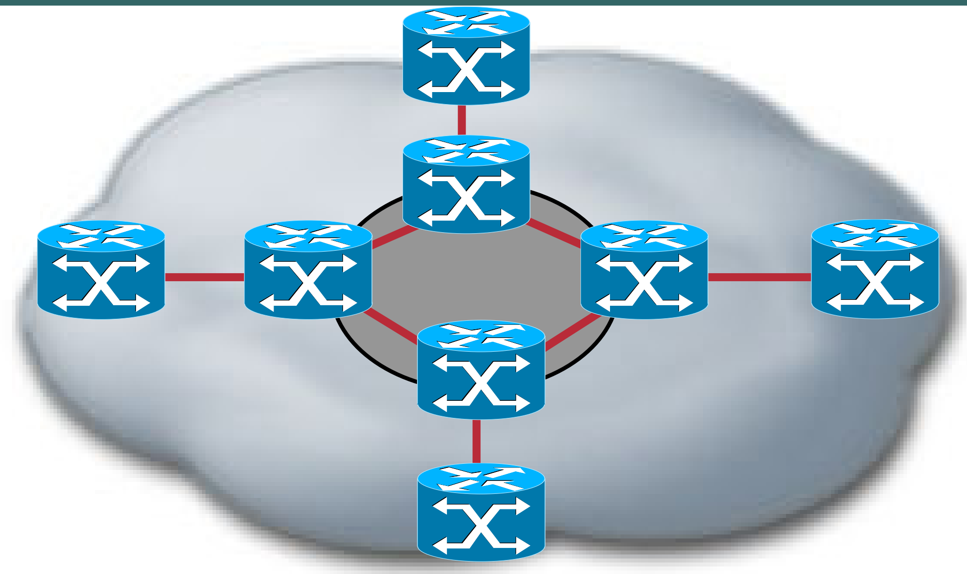
Monique Morrow

Legacy Data Reference Architecture Today
Separate Layers



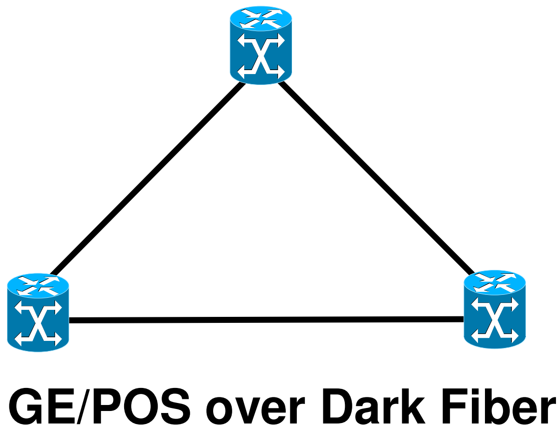
What is Happening in Core ?

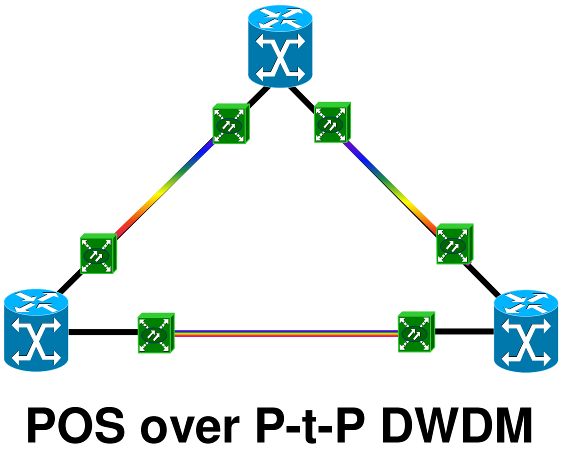
Cisco.com

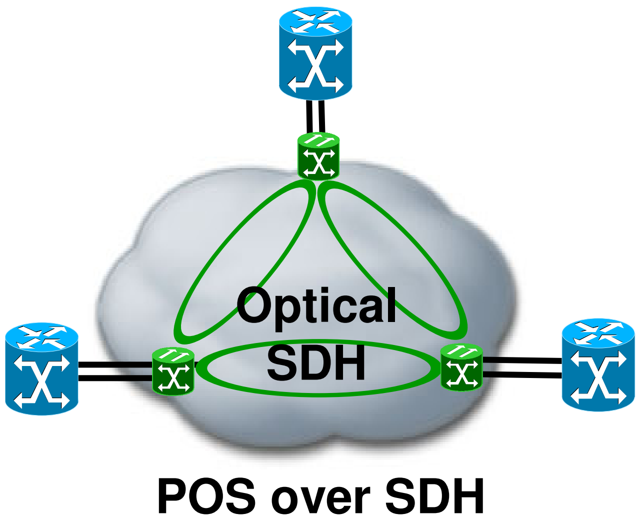


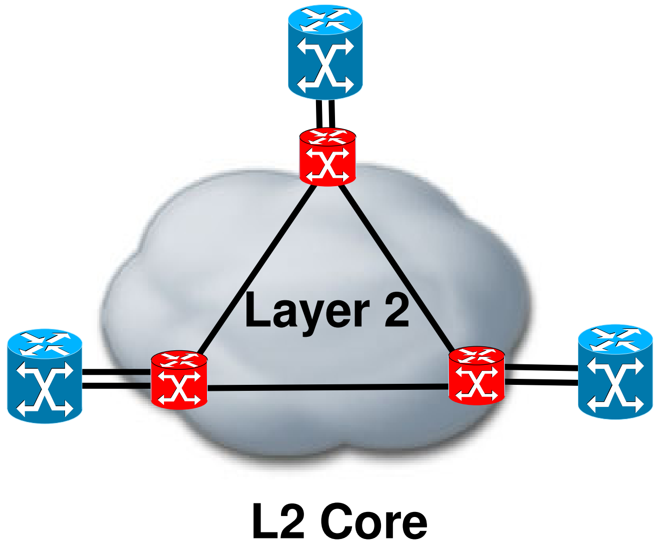
- Core bandwidth is increasing
  - Broadband based
  - New Business services
- Slot count pressure
- 10 Gbps in production in larger PTT networks
- 40 Gbps requirement appearing
- 100 Gbps under discussion !

IP Infrastructures Today

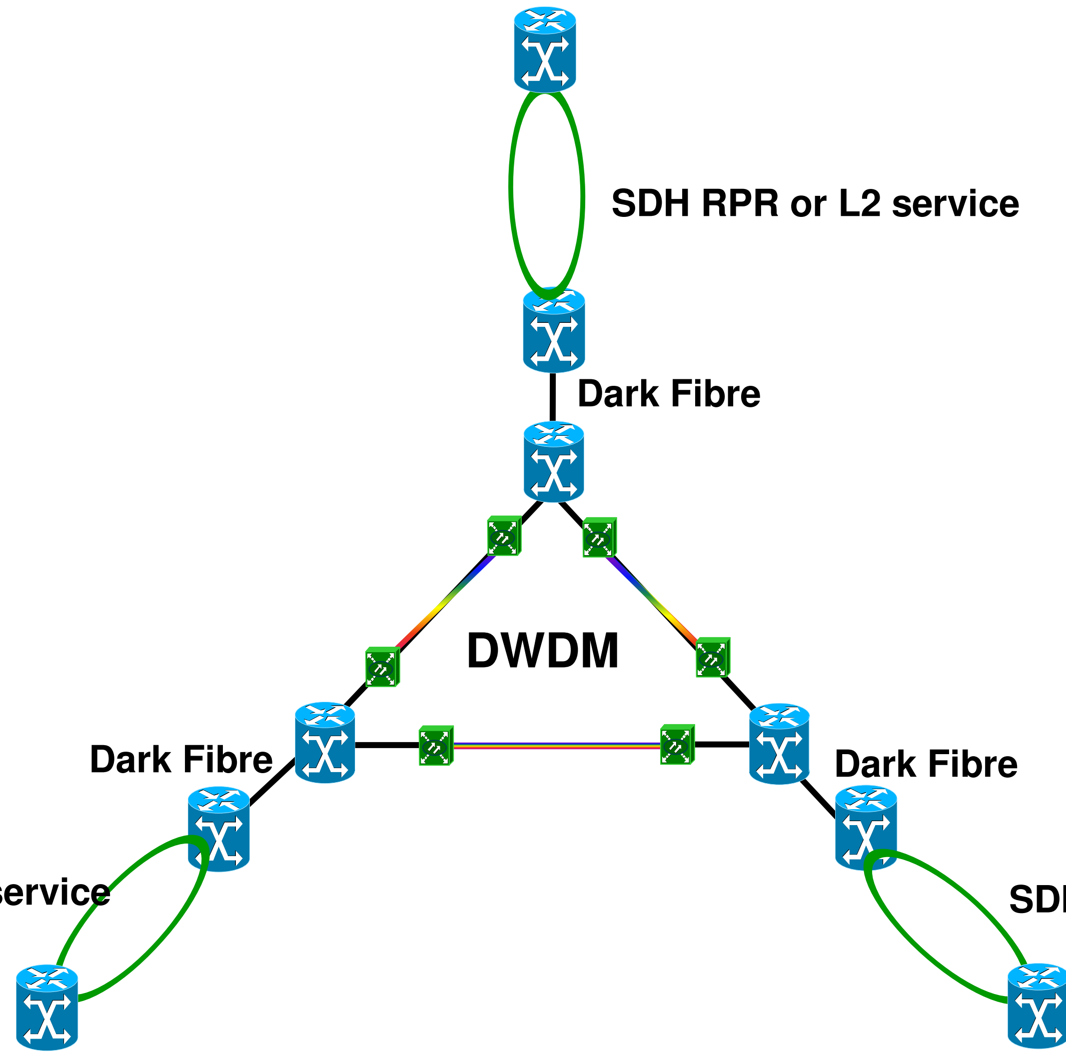




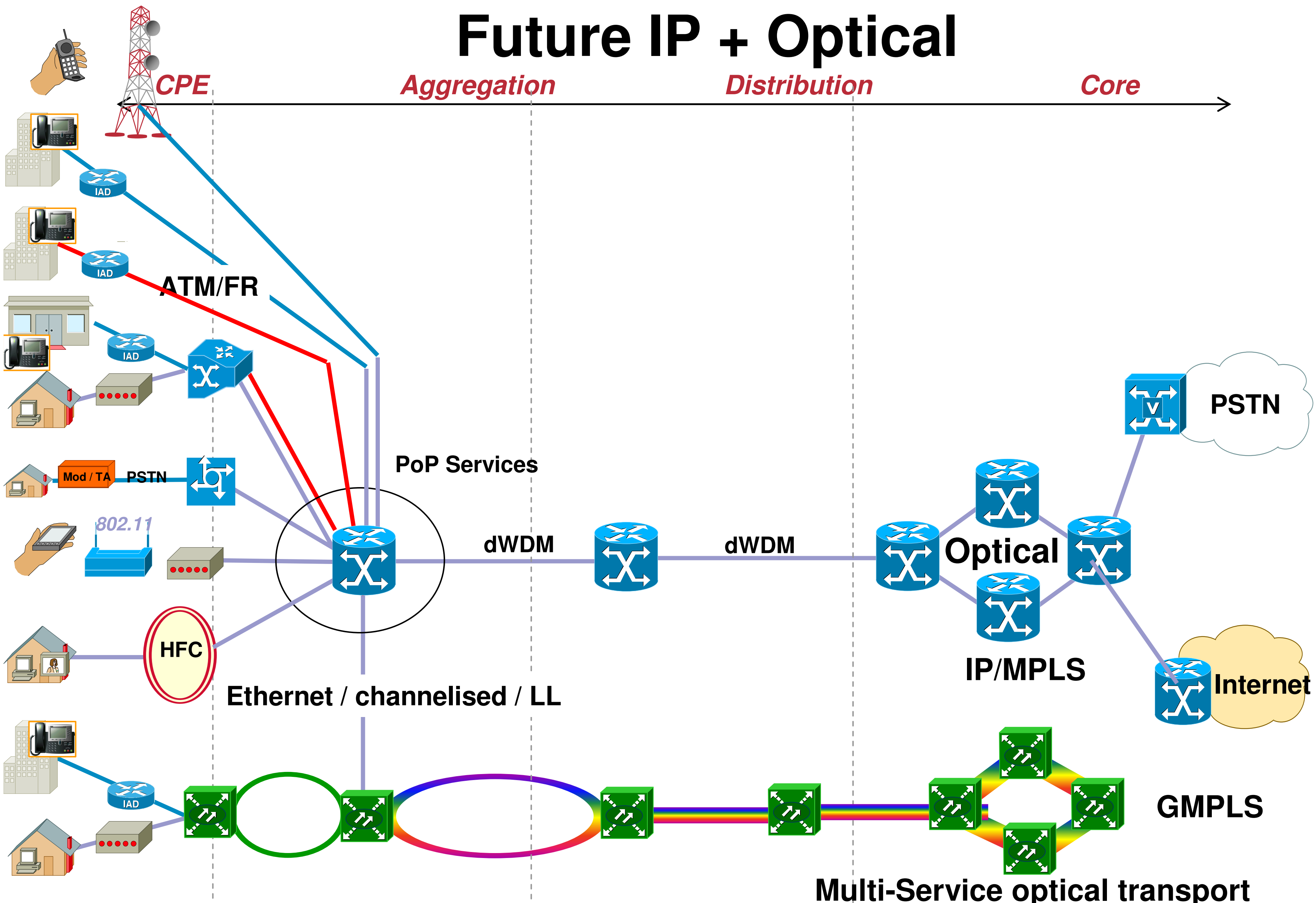




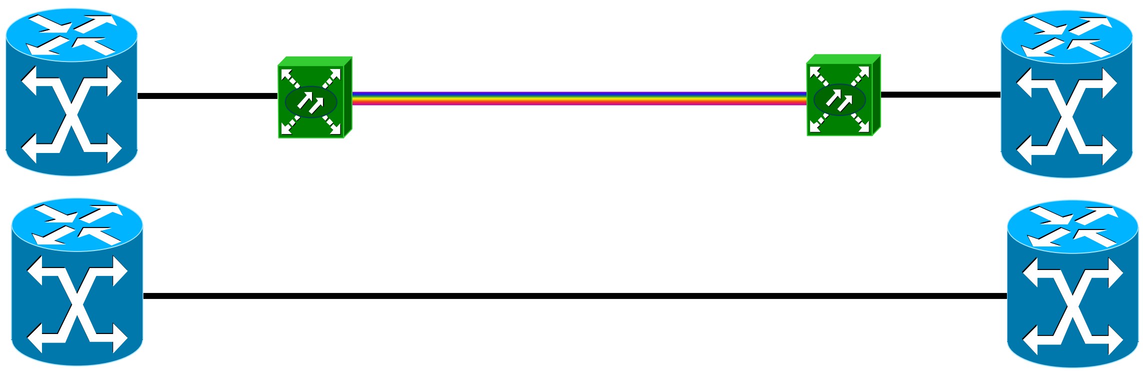
E2e IP Infrastructures Today



Data Reference Architecture
Future IP + Optical

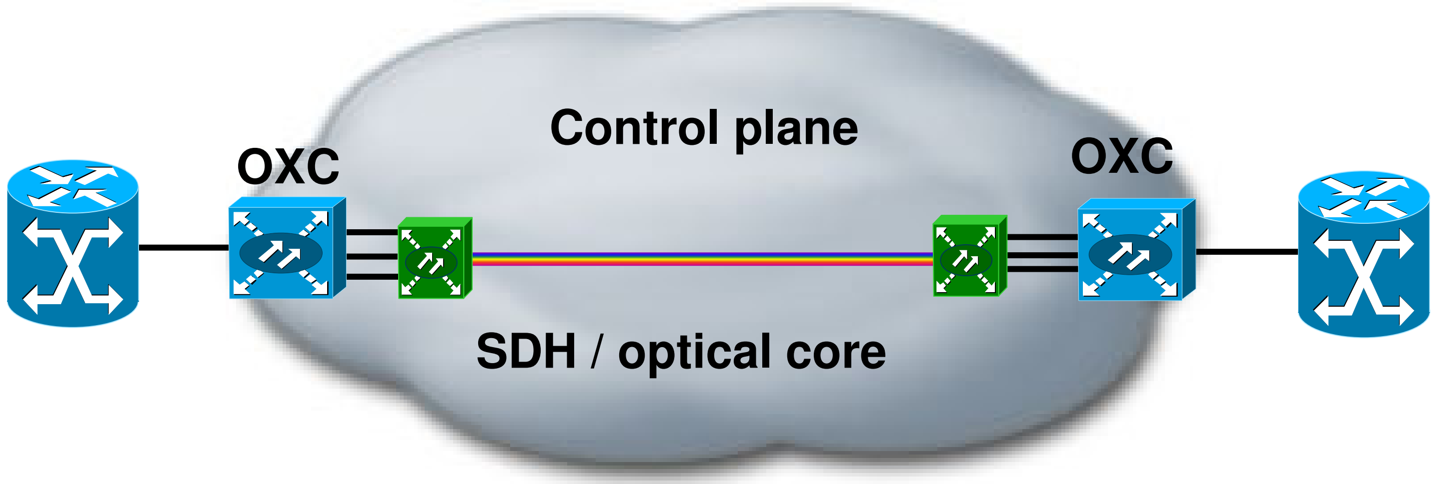


Core Infrastructures Option 1
P-to-P DWDM / Dark Fibre / GE Switches



- Simplest model
- Very high BW connections
    - STM-16c – STM-256c, RPR, GE, 10GE
    - WAN PHY & LAN PHY Long Distance
- Static - Does it matter ?
- No layer 1 recovery
    - L3 or FRR
- Cheap and efficient solution

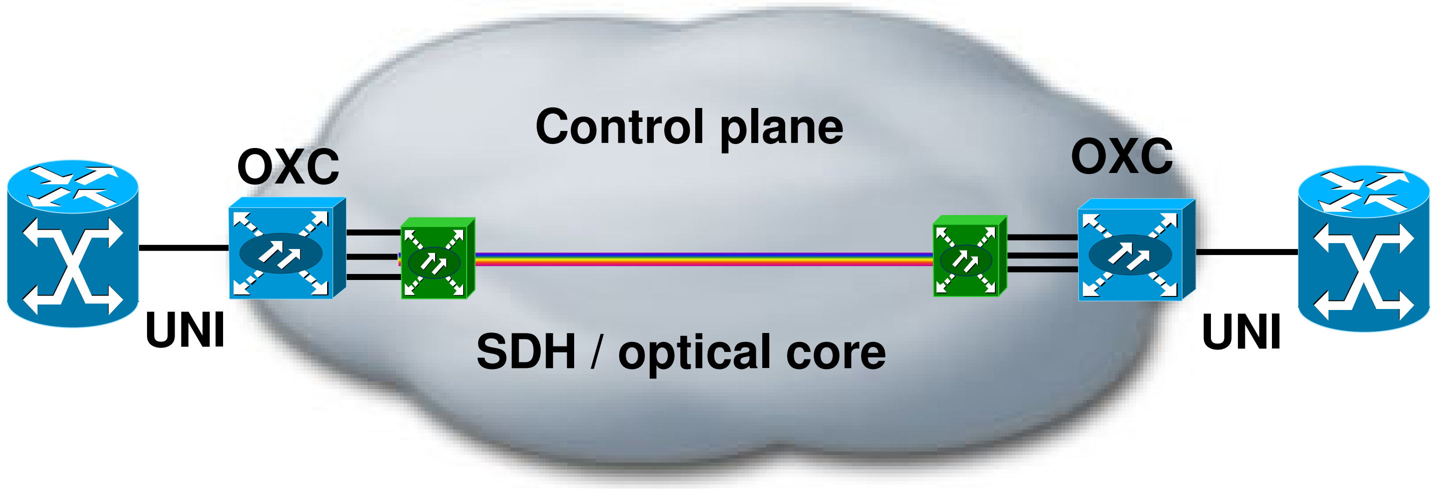
Core Infrastructures Option 2
Overlay without Signalling



- Router connected to optical network
- No signalling interaction
- Limited interaction between Router and optical layer
- Backup at either L1 or L3
- More dynamic / more cost
- Bandwidth capabilities determined by SDH / Optical layer

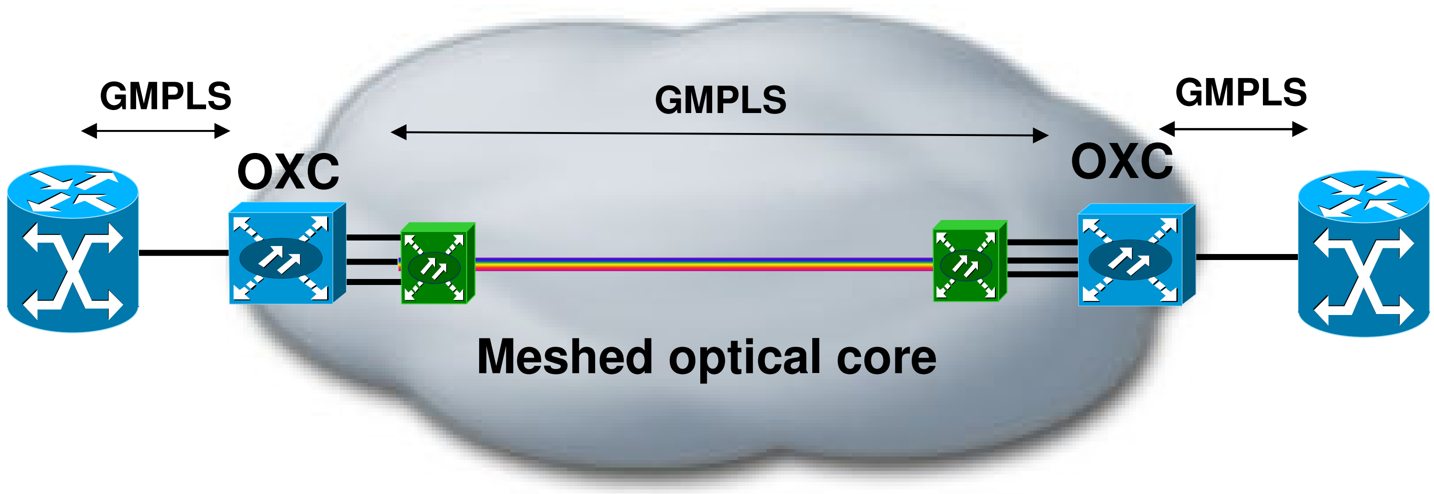
Core Infrastructures Option 3
Overlay with UNI

Cisco.com



- Optical UNI interface between Router and Optical Layer
- Overlay model
- Dynamic bandwidth / BW on demand
    - Initiated from the edge
- Bandwidth capabilities determined by Optical Layer

Core Infrastructures Option 4
Peer Model – GMPLS / G.ASON / …

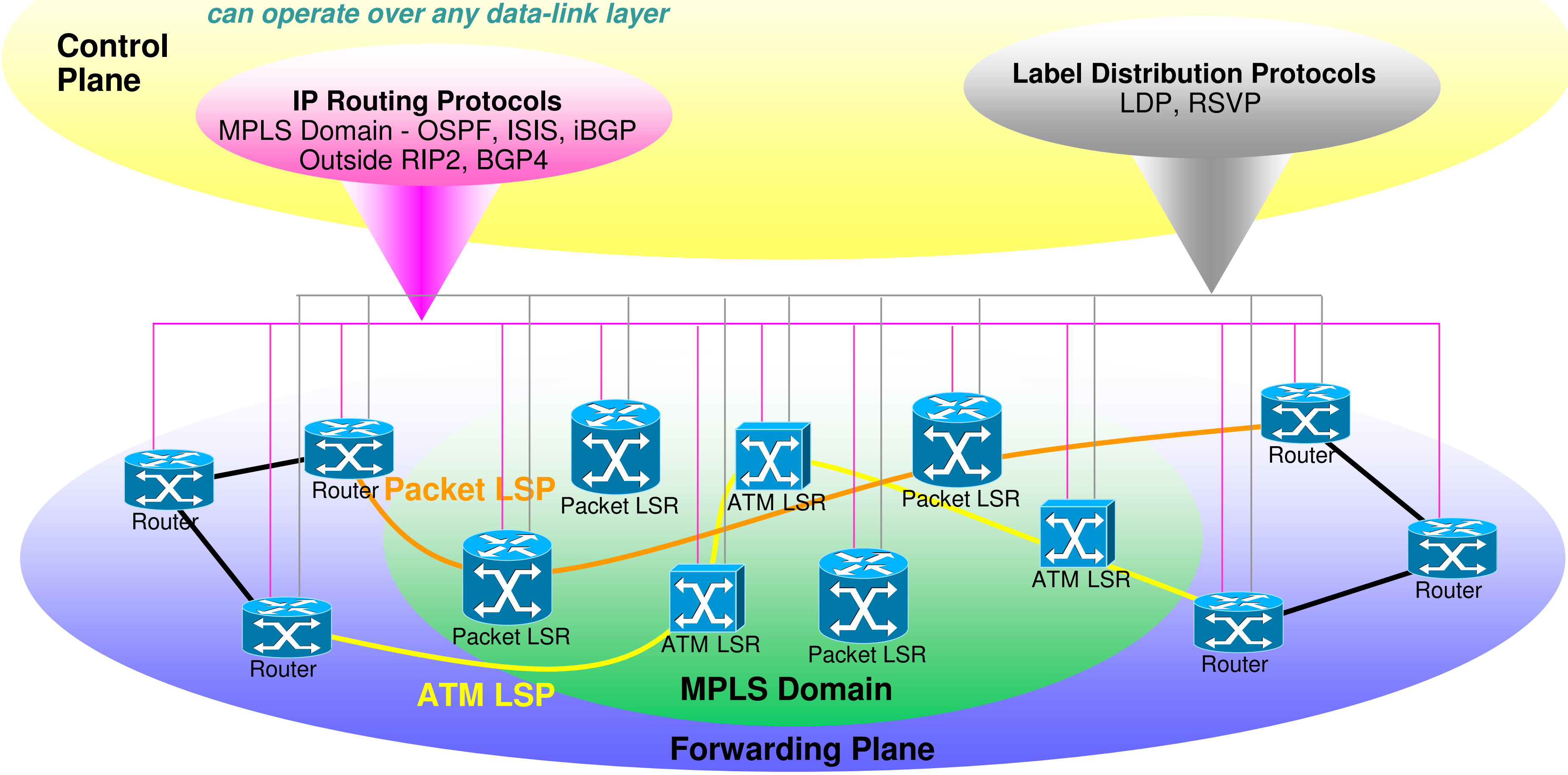


Standards Bodies

Standards	Focus	Applicability to Cisco
OIF OPTICAL·INTERNETWORKING·FORUM	Optical control plane requirements and signaling agreements for UNI and NNI	OIF UNI 1.0
IETF	GMPLS based on extension to IP-based routing and signaling protocols specification to support optical control plane	GMPLS as framework
ITU	Recommendations for ASON/ASTN covering architecture, technical concepts and functional components for control plane based optical paths setup. Leveraging OIF and IETF protocols	Compliance required
MEF	Developing Ethernet services support by OIF control plane	Monitor
Telcordia	Proposing OSS strategy coupled with control plane to set up optical paths	Monitor

…. when MPLS started …

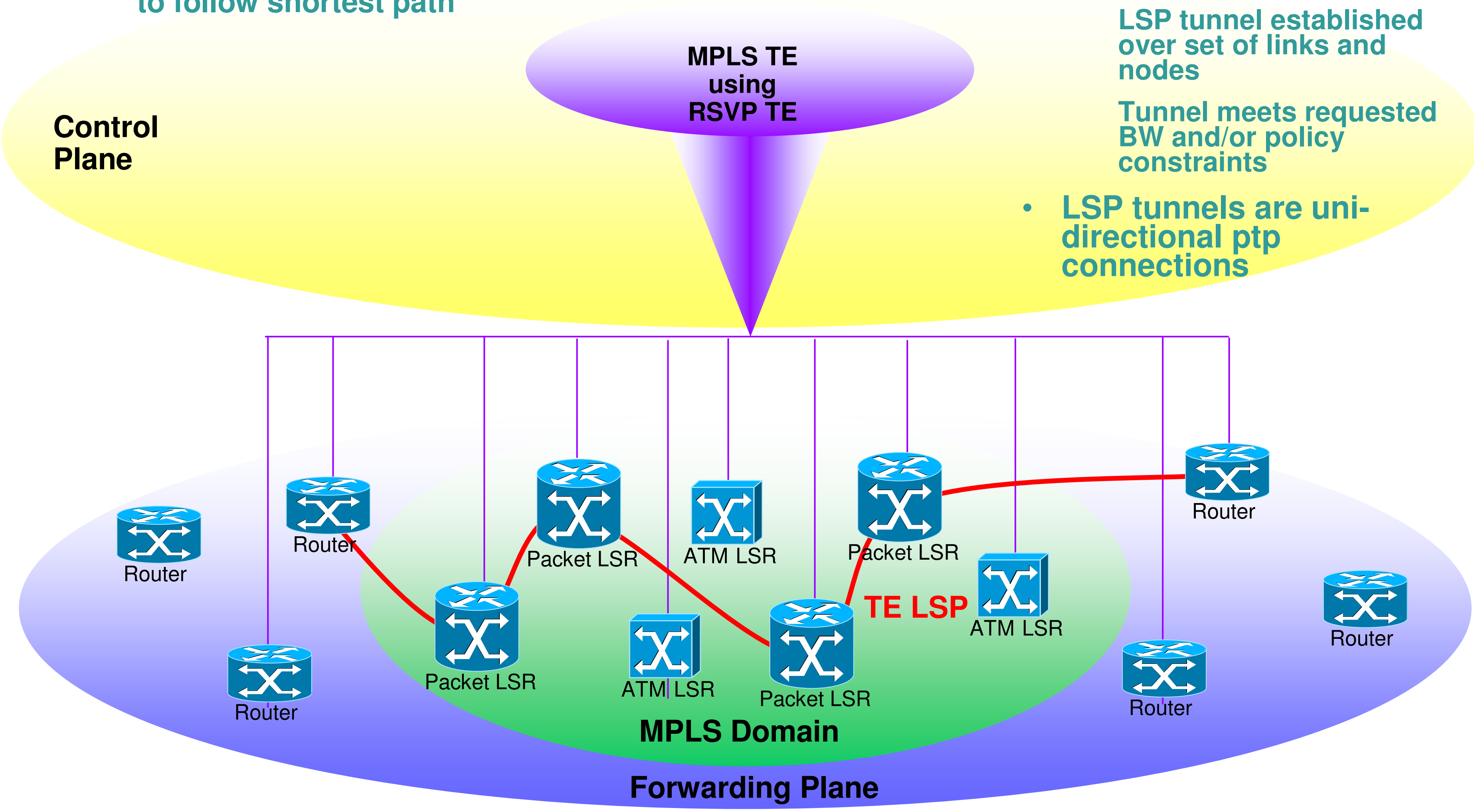
- *General-purpose tunneling mechanism*
    - *carry IP and non-IP payloads*
    - *uses label switching to forward packets/cells through the network*
    - *can operate over any data-link layer*
- *Separate Control Plane from Forwarding Plane*
- *Effort began 1996 ….. RFCs out 2001*
- *RFC 3031 MPLS Architecture*



…. MPLS TE emerged …

Cisco.com

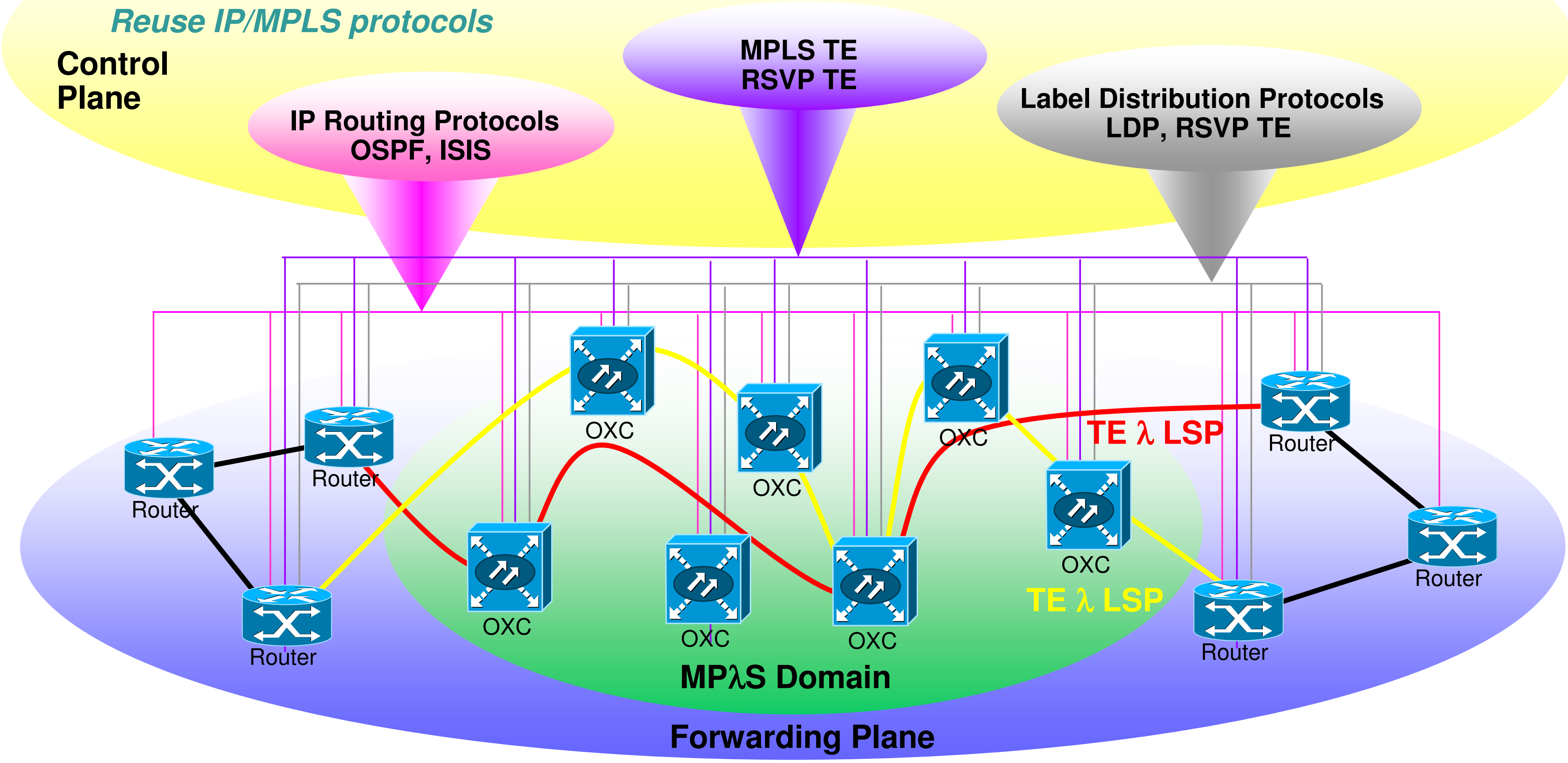
- Packets no longer need to follow shortest path
- Constraint-based routing
  - LSP tunnel established over set of links and nodes
  - Tunnel meets requested BW and/or policy constraints
- LSP tunnels are uni-directional ptp connections



.... then came MPλS ...

Cisco.com

- *Extend MPLS TE protocols to control optical cross-connect (OXC)*
  - *LSRs are like OXC*
  - *LSPs are like optical connections*
  - *Reuse IP/MPLS protocols*
- *Advantages*
  - *fast provisioning of optical connections*
  - *Unified IP/Optical Control Plane*
- *draft-awduche-mpls-te-optical-03.txt Q2 2001*



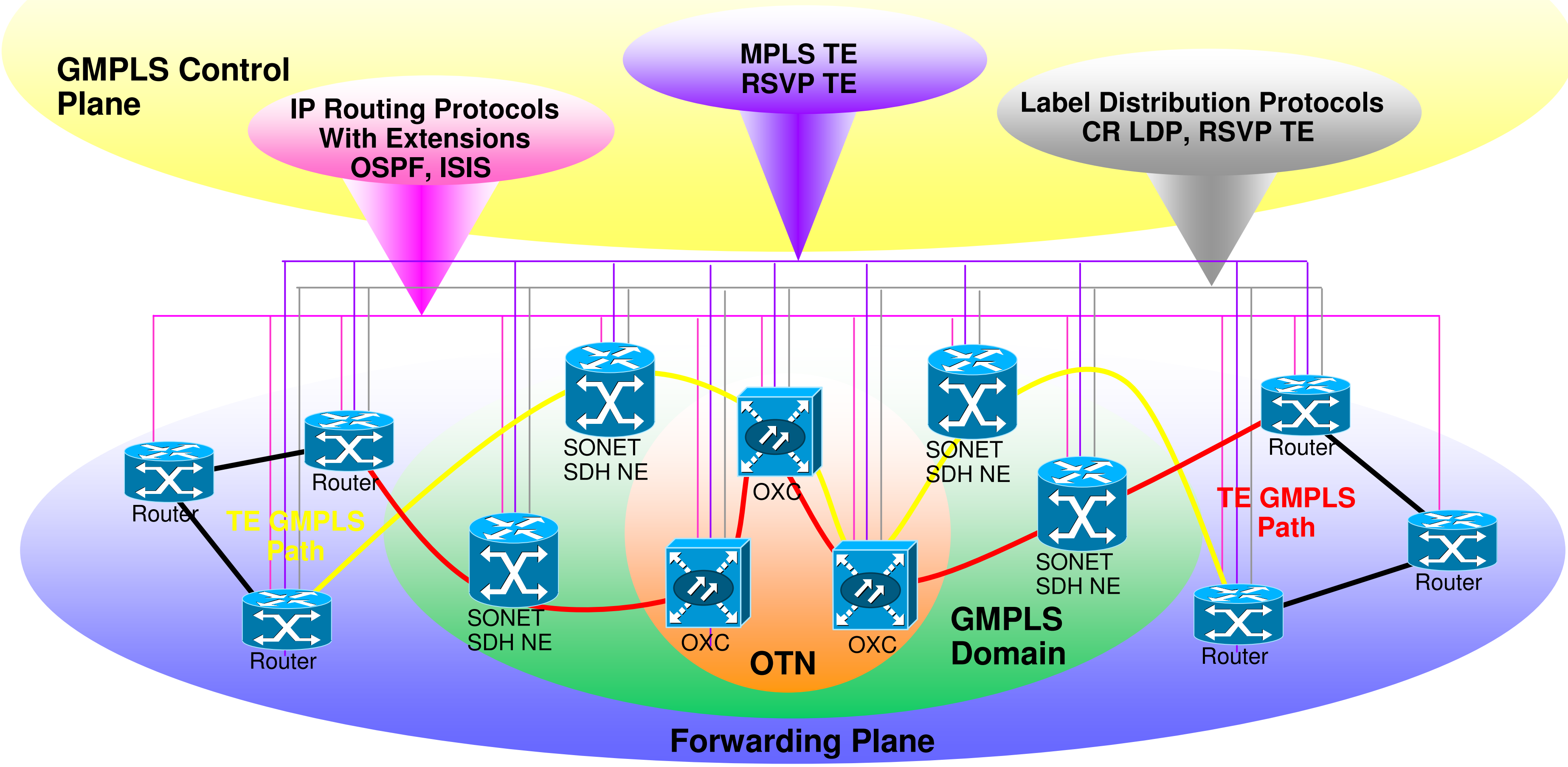
.... finally Generalized MPLS - GMPLS ...

Cisco.com

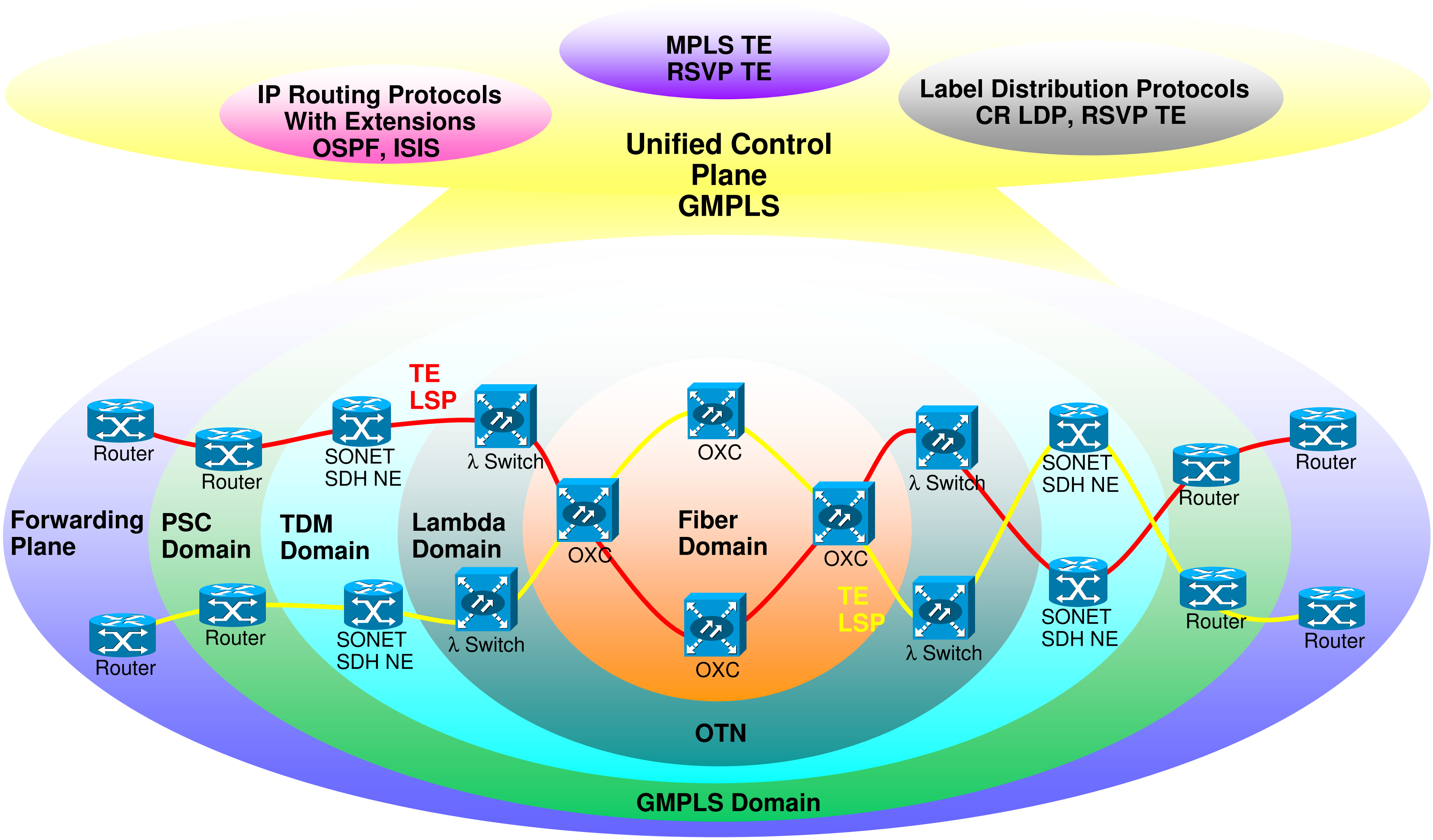
- *GMPLS control plane supports multiple switching and forwarding planes*
- *Introduces new functions to accommodate circuit-oriented optical network regimes*

GMPLS = MPLS + MPλS + N

- where N is MPLS control of new switching planes
- draft-ietf-ccamp-gmpls-architecture-07.txt

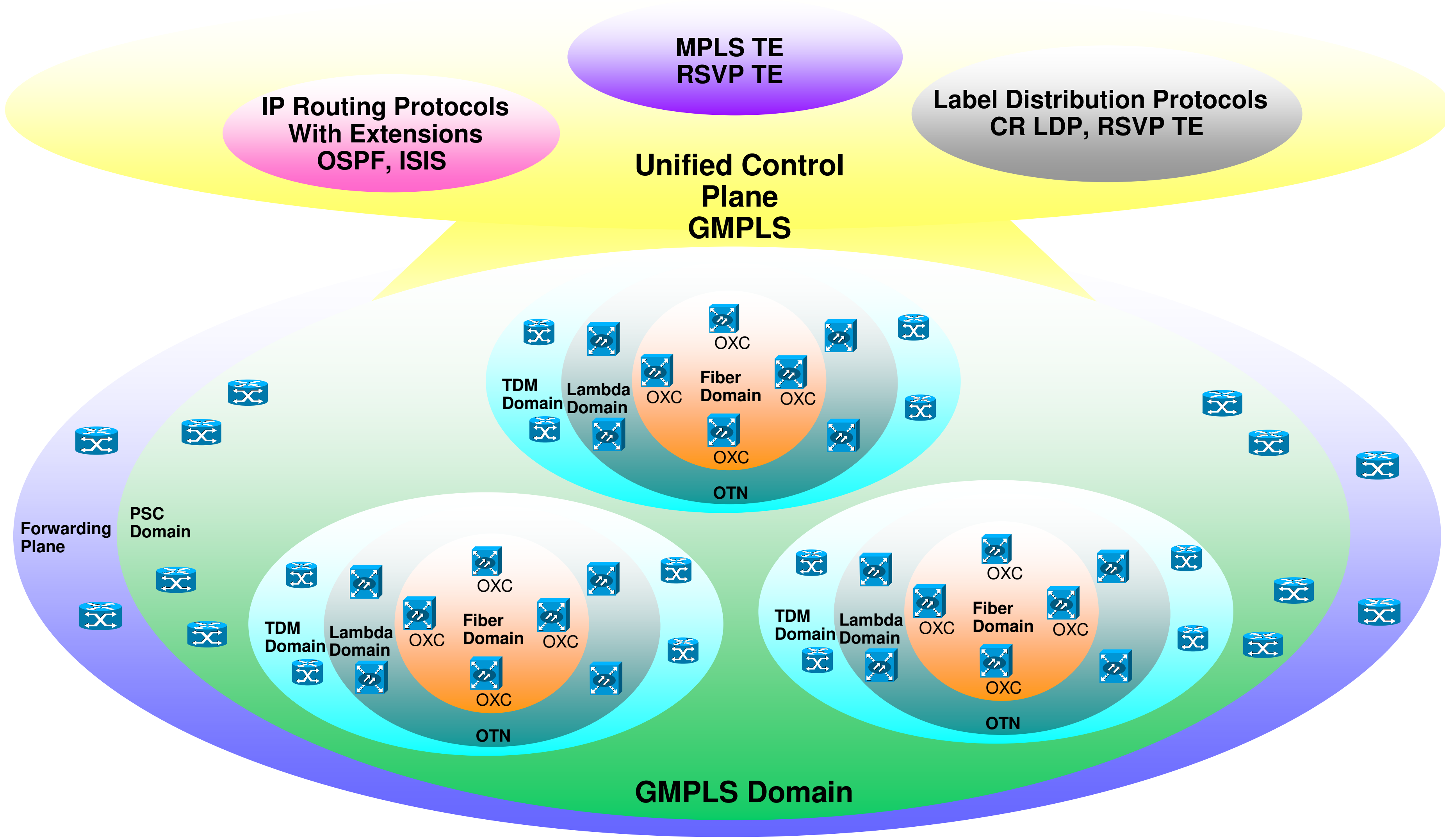


…. N-dimensional GMPLS …



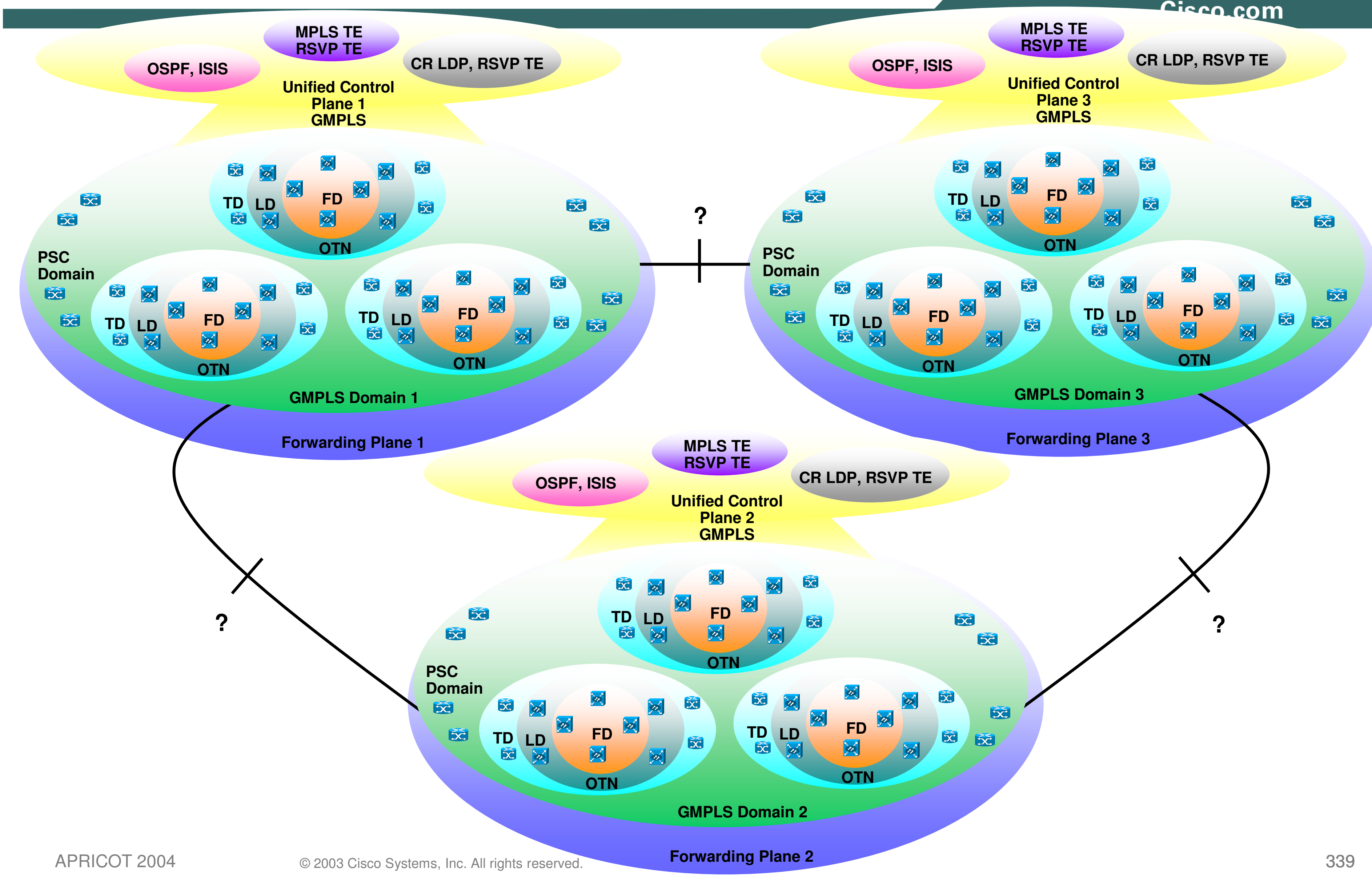
Multiple Sub-Domains in GMPLS Domain

Cisco.com



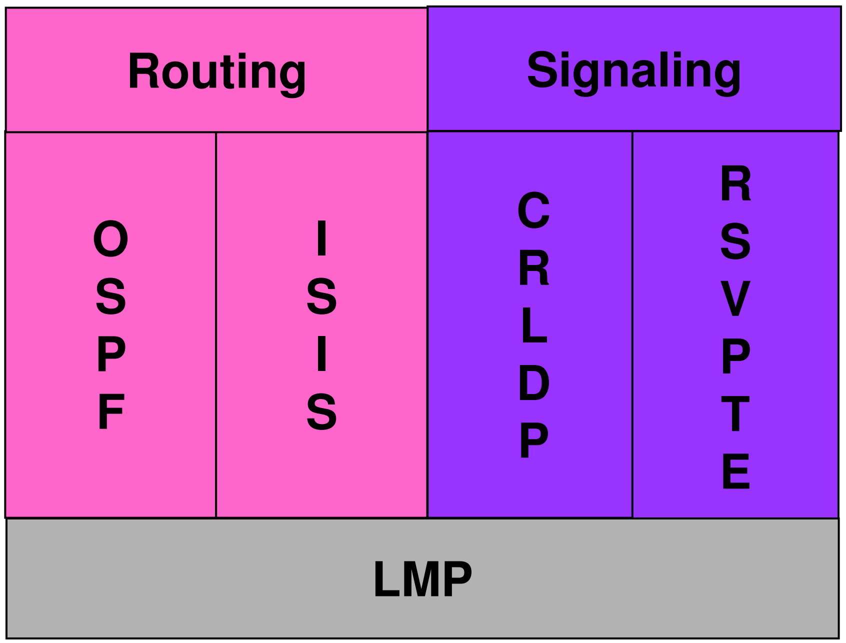
Multiple GMPLS Domains …

Cisco.com



Basic Concepts & Components

Cisco.com



- **Topology Discovery**

  running an IGP (OSPF or IS-IS) with extensions

- **Route Computation**

  Route computation done by NEs

  Link state aggregation and lack of lightpath related information affects efficiency

- **Neighbor Discovery**

  Link Management Protocol like LMP/NDP run in distributed way

- **Lightpath Setup**

  Done by ingress NE using signaling protocol like RSVP-TE

RFC 3472 GMPLS Signaling CR-LDP Extensions

RFC 3473 GMPLS Signaling RSVP-TE Extensions

Forwarding Planes

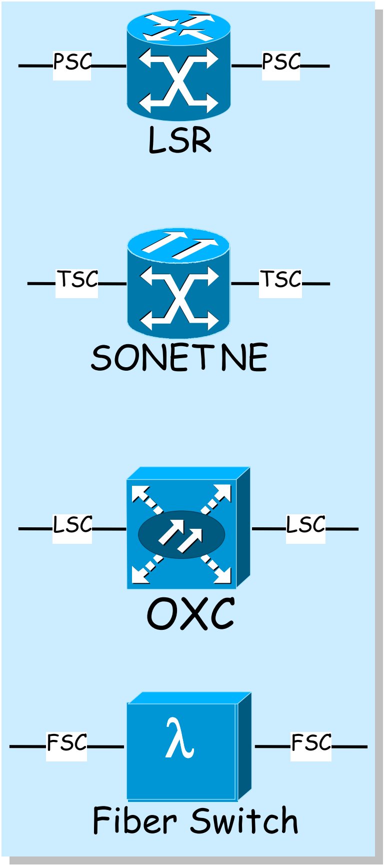
- **MPLS only supports LSRs which recognize packet/cell boundaries**
- **Support for devices making forwarding decision on other than packet/cell boundaries**
- **Forwarding plane switching decision based on interface type of LSR**

  Packet Switch Capable (PSC)

  TDM Switch Capable (TSC)

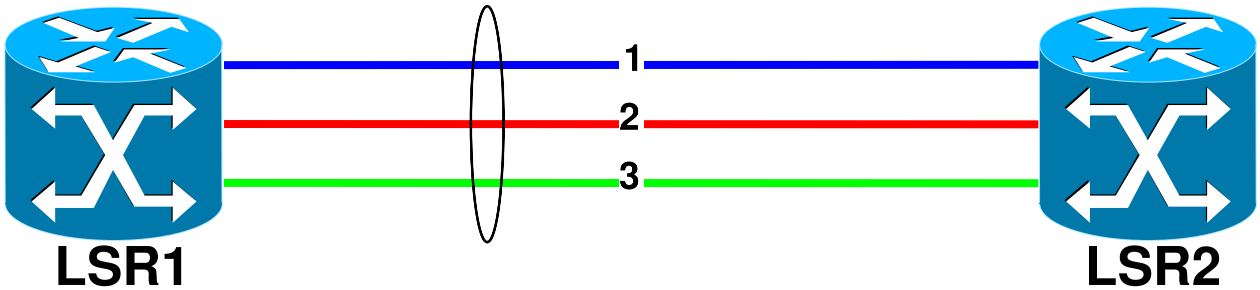
  Lambda Switch Capable (LSC)

  Fiber Switch Capable (FSC)



RFC 3471 GMPLS Signaling Functional Description

Link Bundling & Unnumbered Links



- **Issue**

  Neighboring LSRs connected by multiple parallel links

  Each link is addressed at each end and advertised into routing database … lots of links !!!

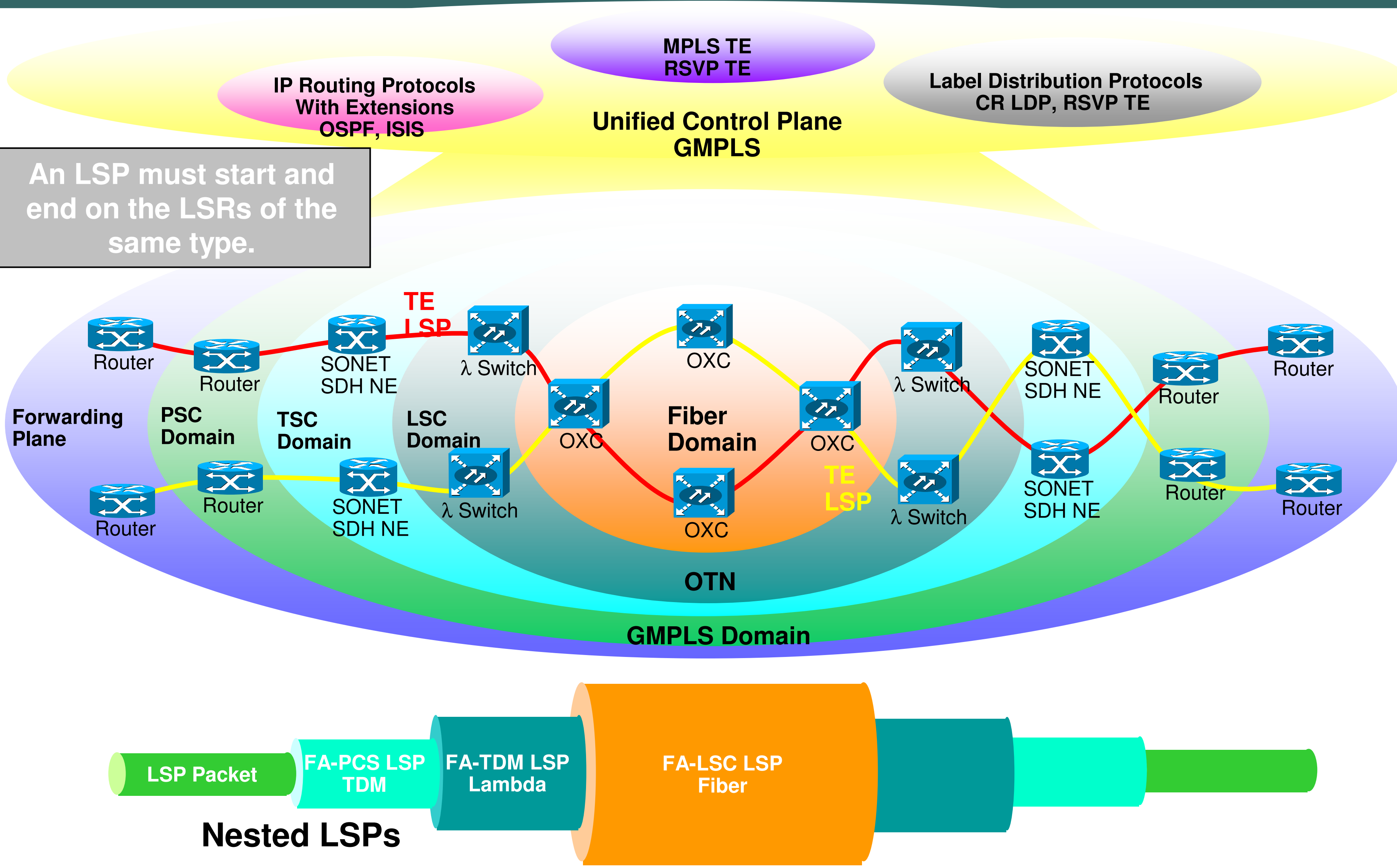
- **Solution**

  Aggregate multiple Components Links into a single Abstract Link

  Use (Router ID, Interface #) for link identifiers

- **Reduces number of links in routing database and amount of per-link configuration**
- **draft-kompella-mpls-bundle-05.txt**
- **draft-kompella-mpls-unnum-02.txt**

Hierarchical LSPs



LSP Hierarchy

Cisco.com

FA-LSP…Forwarding Adjacency LSP

Nested LSPs

- **Enables aggregation of GMPLS LSP tunnels**
- **Accomplished by**

  Inter-LSR LSP tunnel (FA-LSP) link is created

  Ingress LSR injects link (FA-LSP) into IGP database

  Other routers use the link in path calculation/setup

  Other LSP tunnels are nested inside FA-LSP
- **Advantages**

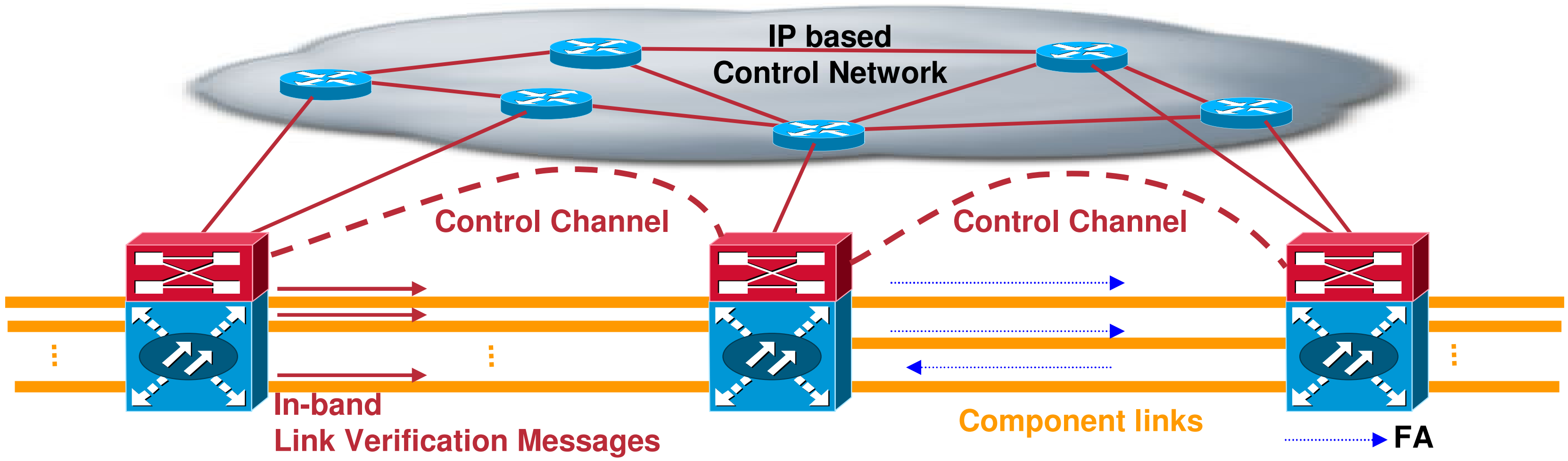
  Fewer high-order labels (e.g.lambdas) consumed

  Nested LSPs can be of non-discrete bandwidth

  FA-LSP can "hide" topology

- **draft-ietf-mpls-lsp-hierarchy-08.txt**

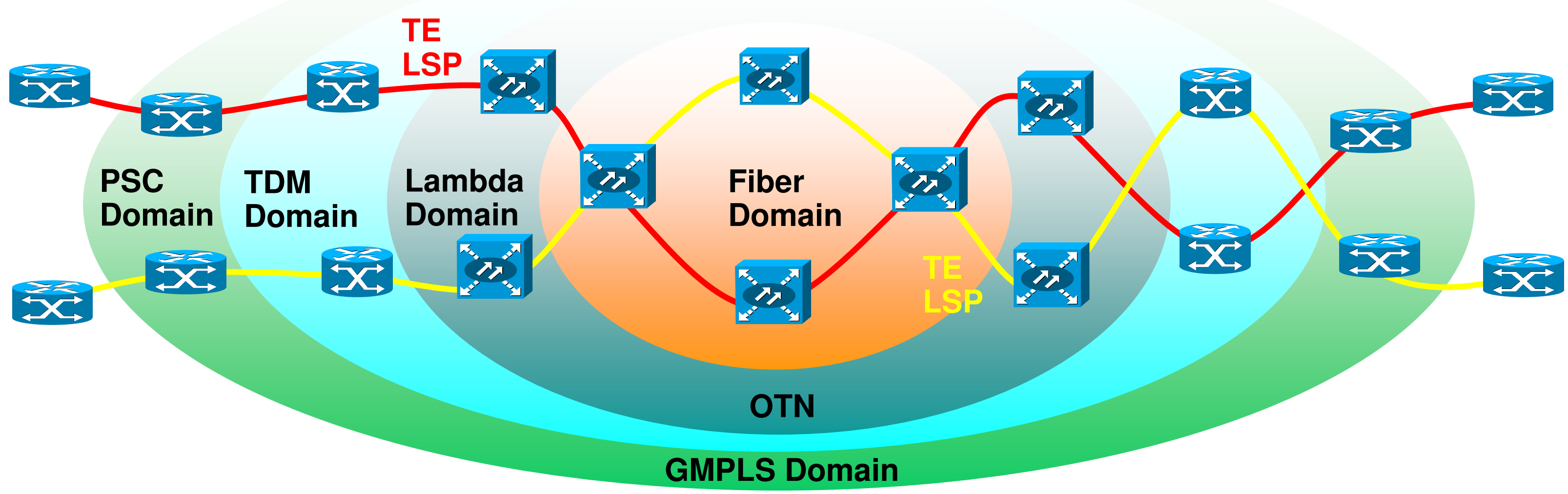
LMP & Link Management



- **LMP Functionality**
  - Most LMP messages sent out-of-band through CC
  - In-band messages sent for Component Link Verification
  - Once allocated, Component Link is not assumed to be opaque
  - Port ID mapping
  - One CC per one or more Component Link Bundles
  - Fault isolation
  - End-system and service discovery (UNI related)
- **Flooding Adjacencies are maintained over CC (via control network)**
- **Forwarding Adjacencies (FA) are maintained over Component Links and announced as links into the IGP**

- draft-ietf-mpls-lmp-02.txt
- draft-ietf-ccamp-lmp-10.txt
- draft-ietf-ccamp-lmp-wdm-02.txt

GMPLS Signaling



- **Extended label semantics for Fiber, Waveband, Lambda, TDM and PSC LSP setup**
- **Extend RSVP-TE/CR-LDP for opaquely carrying new label objects over explicit path**
- **Suggested Label - conveyed by upstream LSR to downstream LSR to speed up configuration (on upstream)**
- **Label Set - limits choice of labels that downstream LSR can choose from**
  - **If no wavelength conversion available then same lambdas must be used ete**
- **Bidirectional LSP setup**

draft-ietf-mpls-generalized-signaling-09.txt

GMPLS Routing Extensions

Cisco.com

- Extensions needed to deal with the polymorphic nature of GMPLS links
  - links that are not capable of forwarding packets nor can they support router adjacencies
  - links that are aggregates of many component links (e.g. link bundles)
  - links that are FAs between non-adjacent routers
- Define new sub-TLVs for
  - OSPF Link TLV
  - IS-IS Reachability TLV
- Flooded over bi-directional control channels (CC) connecting GMPLS nodes
  - CC may not necessarily follow topology of data bearing (component) links

- draft-ietf-ccamp-gmpls-routing-09.txt
- draft-ietf-ccamp-ospf-gmpls-extensions-12.txt
- draft-ietf-isis-gmpls-extensions-19.txt
- draft-ietf-ccamp-rsvp-te-exclude-route-00.txt

GMPLS Routing sub-TLVs

- **Link Mux Capability**

  defines the receiving nodes ability to demultiplex data based on packets, TDM timeslots, lambdas or fiber

- **Link Descriptor**

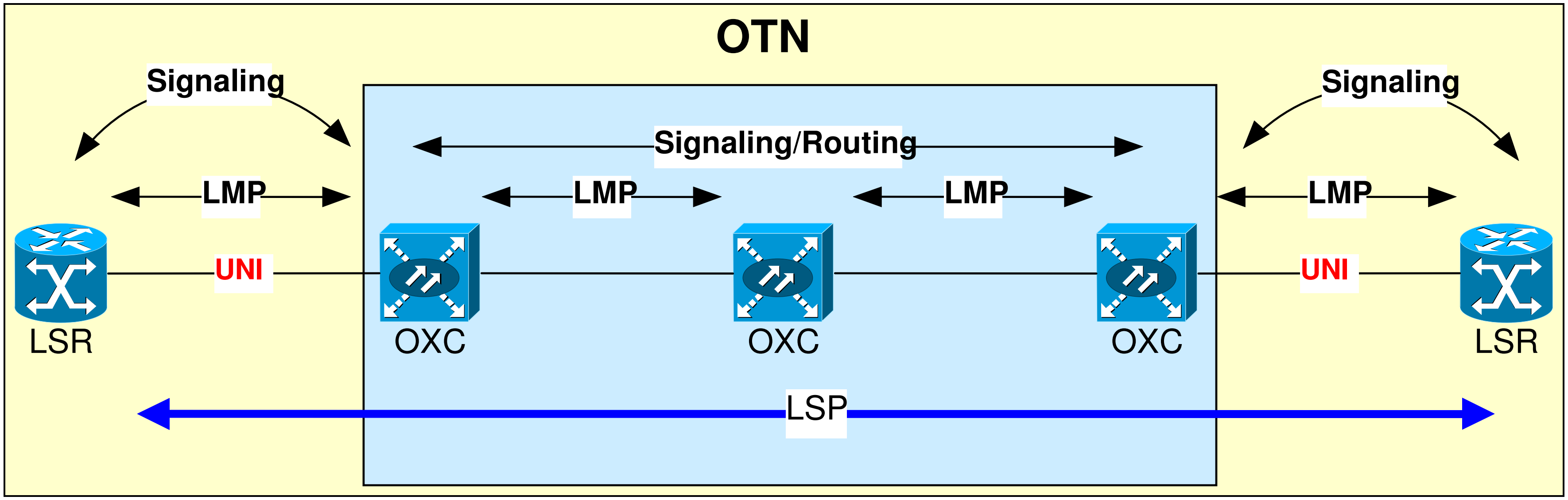
  link encoding type and bandwidth granularity

- **Shared Risk Link Group (SRLG)**

  physical fiber diversity - e.g. two fibers with same SRLG are in the same conduit

- **Link Protection Type**

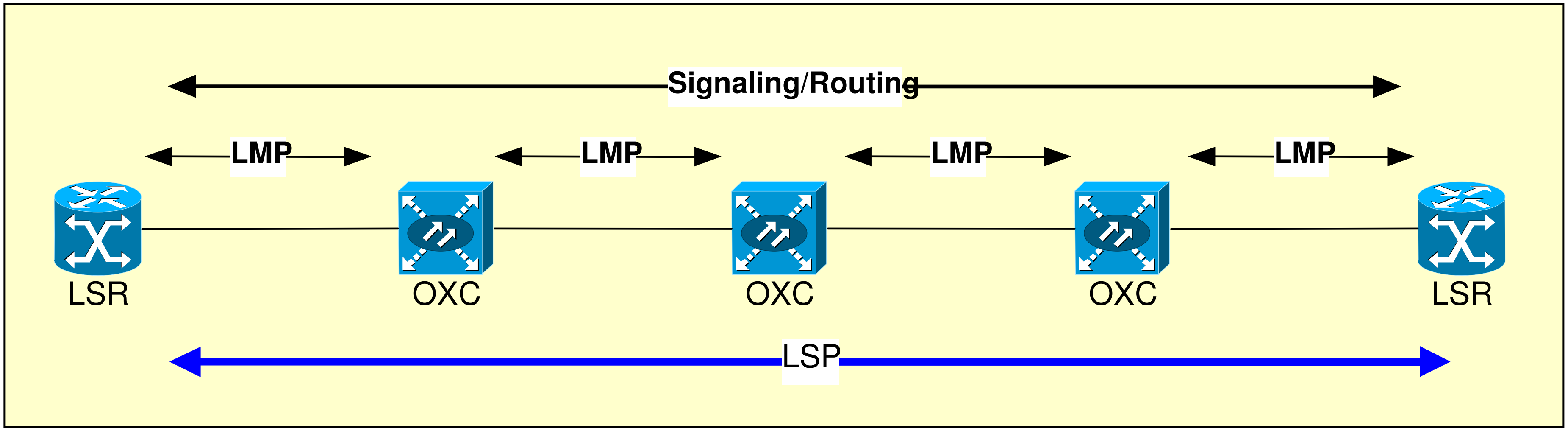
GMPLS Overlay Routing Model



- **UNI interactions - GMPLS signaling, LMP**
- **OTN interactions - GMPLS signaling, routing and LMP**
- **draft-ietf-ccamp-gmpls-overlay-02.txt**

  (RSVP Support for Overlay Model)

GMPLS Peer Routing Model



- **OTN interactions - GMPLS signaling, routing and LMP**
- **GMPLS protocol machinery can support overlay or peer routing models**
- **RFC 3473 GMPLS Signaling RSVP-TE Extensions**

Protection & Restoration

Many different Restoration & Protection Schemes (Co) exist today !

SDH

IP

Optical Protection

MPLS TE FRR

draft-ietf-ccamp-gmpls-recovery-terminology-02.txt

Protection

Static
Dynamic

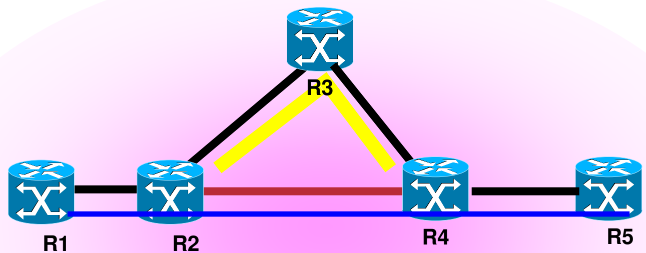
Protection Mode

L1 Only
L3 Only
L1 / L3 Independent
L1 / L3 Coordinated (Hold Off Timer)
L1 & L3 Interworking

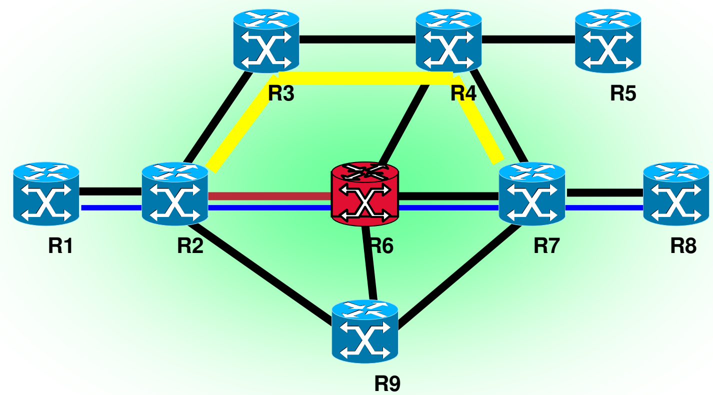
Protection Type

Node Protection
Link Protection

GMPLS Protection / Restoration Based on MPLS TE FRR



Link Protection



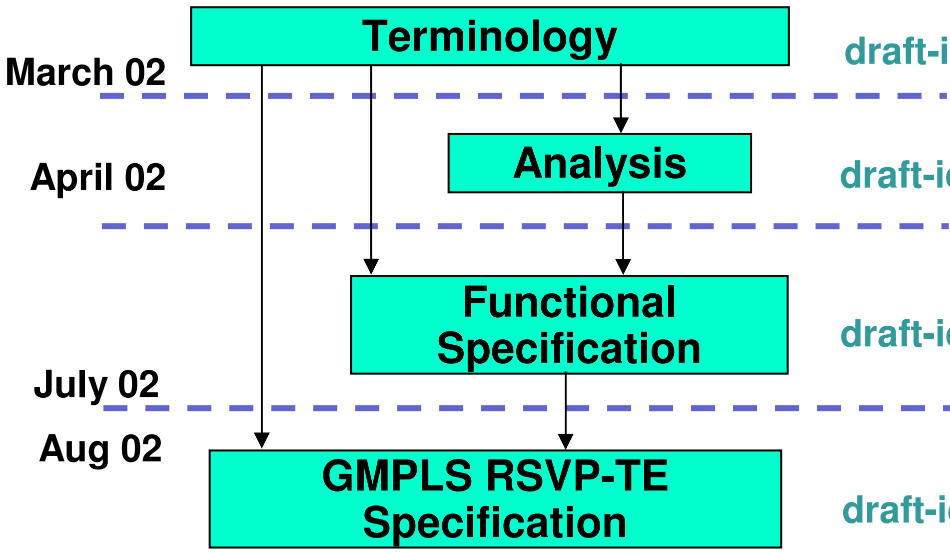
Node Protection

- **FRR mechanism to minimize packet loss during Link / Node Failure**
- **Pre-provisioned protection tunnels carry traffic when protected resource goes down**
- **MPLS-TE to signal FRR protection tunnels**

  MPLS TE traffic doesn't have to follow IGP shortest path
- **Can protect MPLS or IP traffic !**

GMPLS Based Recovery

Cisco.com



draft-ietf-ccamp-gmpls-recovery-terminology-02.txt

draft-ietf-ccamp-gmpls-recovery-analysis-02.txt

draft-ietf-ccamp-gmpls-recovery-functional-01.txt

draft-ietf-ccamp-gmpls-recovery-e2e-signaling-02.txt

- LSP Protection

  full LSP signaling (cross-connection) before failure occurrence

- Pre-Planned Rerouting (with shared rerouting as particular case)

  Pre-signaling before failure – LSP activation after failure – allows for low priority

- LSP Dynamic Rerouting (aka restoration)

  full LSP signaling after failure occurrence

GMPLS MIBs

- Based on MPLS MIBs - Revision 3 now ready

  http://www.olddog.co.uk/download

- Open issues

  Expand conformance statements for configuration/monitoring tunnel resources in GMPLS systems like SONET/SDH or G.709

  Extend performance tables for technology specific GMPLS LSPs

  Consider way to expose

    Tunnel heads

    Tunnel tail

    Tunnel transfer entries

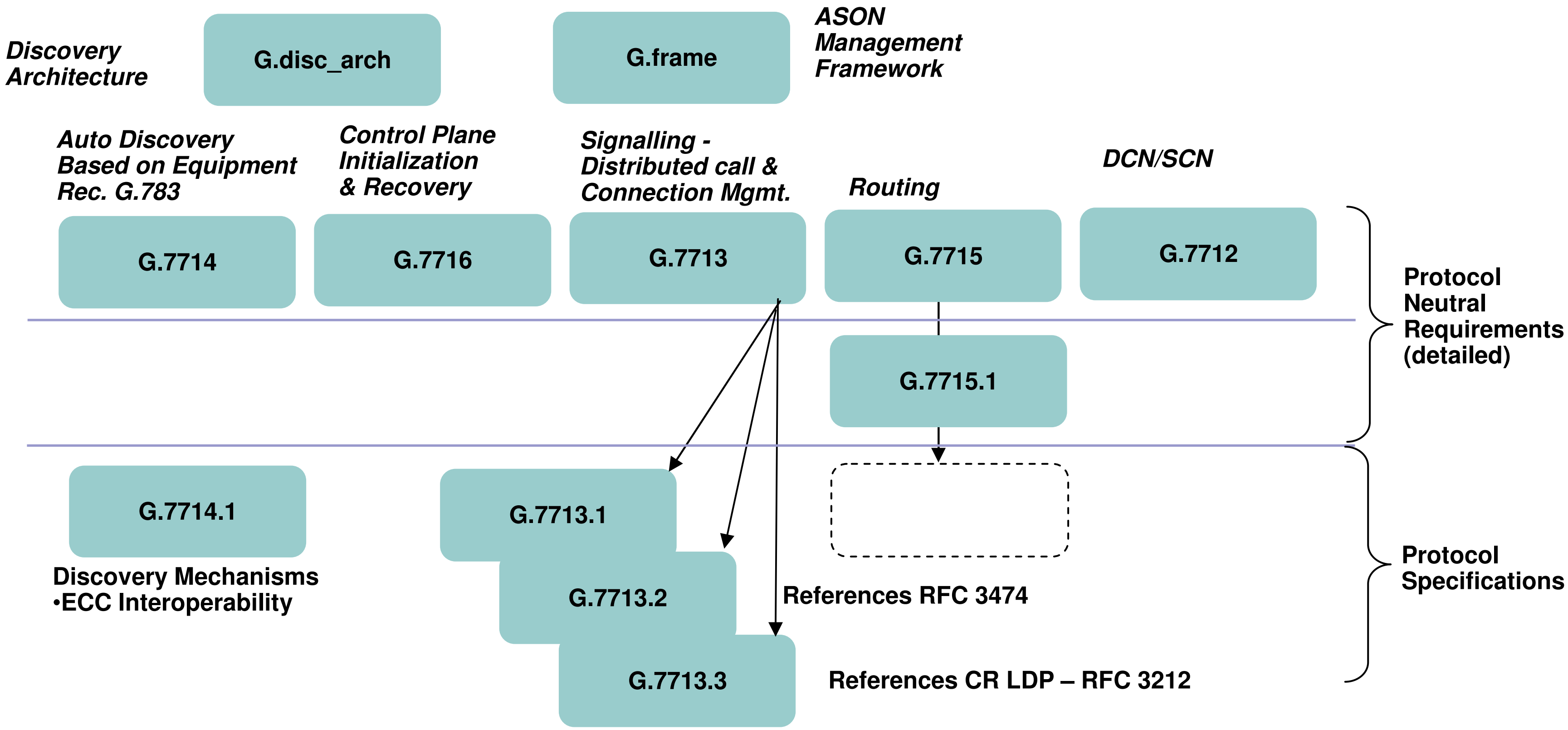
  Support for IF_ID control and error reporting

  LSR or interface config for Hellos and Restart

- draft-ccamp-ietf-gmpls-tc-mib-01.txt
- draft-ccamp-ietf-gmpls-lsr-mib-01.txt
- draft-ccamp-ietf-gmpls-te-mib-01.txt

ITU-T SG 15 Communications to IETF CCAMP Qestion14 – Optical Control Plane

Cisco.com



ITU-T SG 15, Question 14 - ASON Control & Management Recommendations

- **Recommendations G.7715.1 and living lists for G.7714.1 and G.7713**

  ftp://sg15opticalt:atxchange@ftp.itu.int/tsg15opticaltransport/COMMUNICATIONS/index.html

  http://www.ietf.org/iesg/liaison.html

GMPLS Extensions for ASON

- Extend GMPLS Signaling (RFC 3471 / RFC 3475)

    Must meet FULL functional requirements of ASON architecture in GMPLS

        provide call & connection mgmt (G.7713)

    Must be BACKWARD COMPATIBLE with current GMPLS RFCs

- ASON architecture includes

    Automated control plane supporting both call & connection mgmt (G.8080)

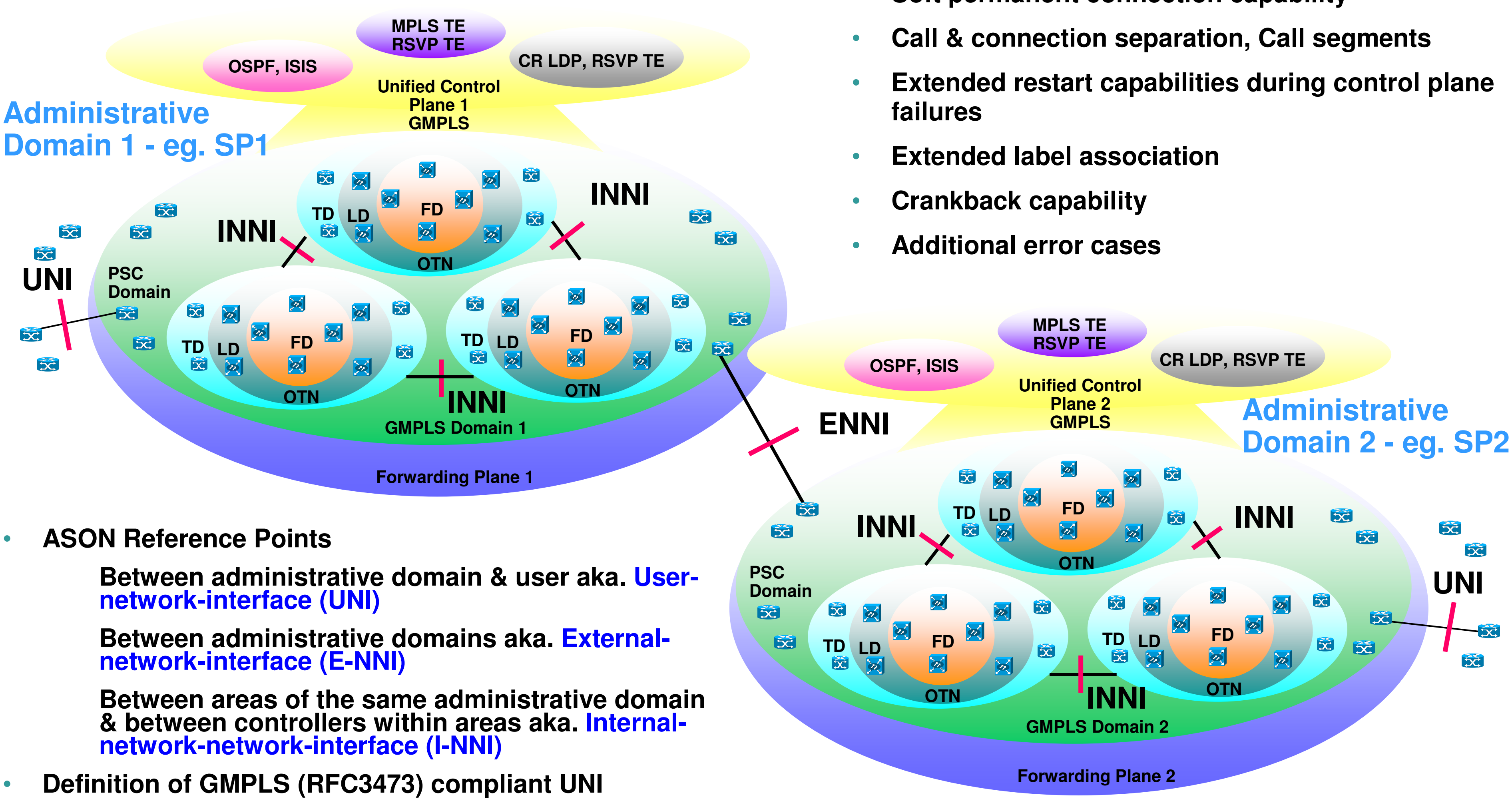
    Control plane applicable to different transport technologies (eg. SDH/SONET, OTN) & networking environments (eg. Inter-Carrier, Intra-Carrier)

    Refined reference point terminology (UNI, E-NNI, I-NNI)

- draft-ietf-ccamp-gmpls-ason-reqts-04.txt

GMPLS Extensions for ASON Reference Point Terminology - UNI, ENNI, INNI

- Soft permanent connection capability
- Call & connection separation, Call segments
- Extended restart capabilities during control plane failures
- Extended label association
- Crankback capability
- Additional error cases



- ASON Reference Points

  Between administrative domain & user aka. User-network-interface (UNI)

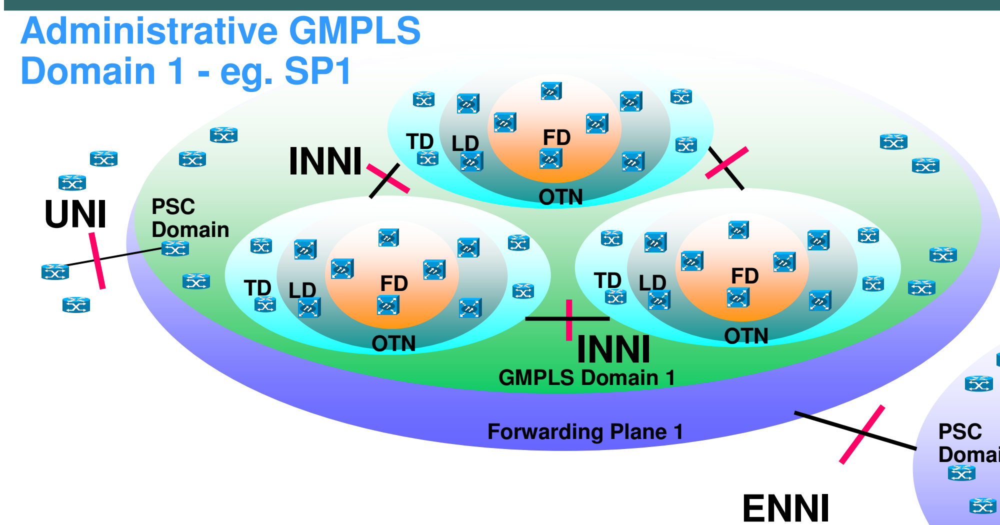
  Between administrative domains aka. External-network-interface (E-NNI)

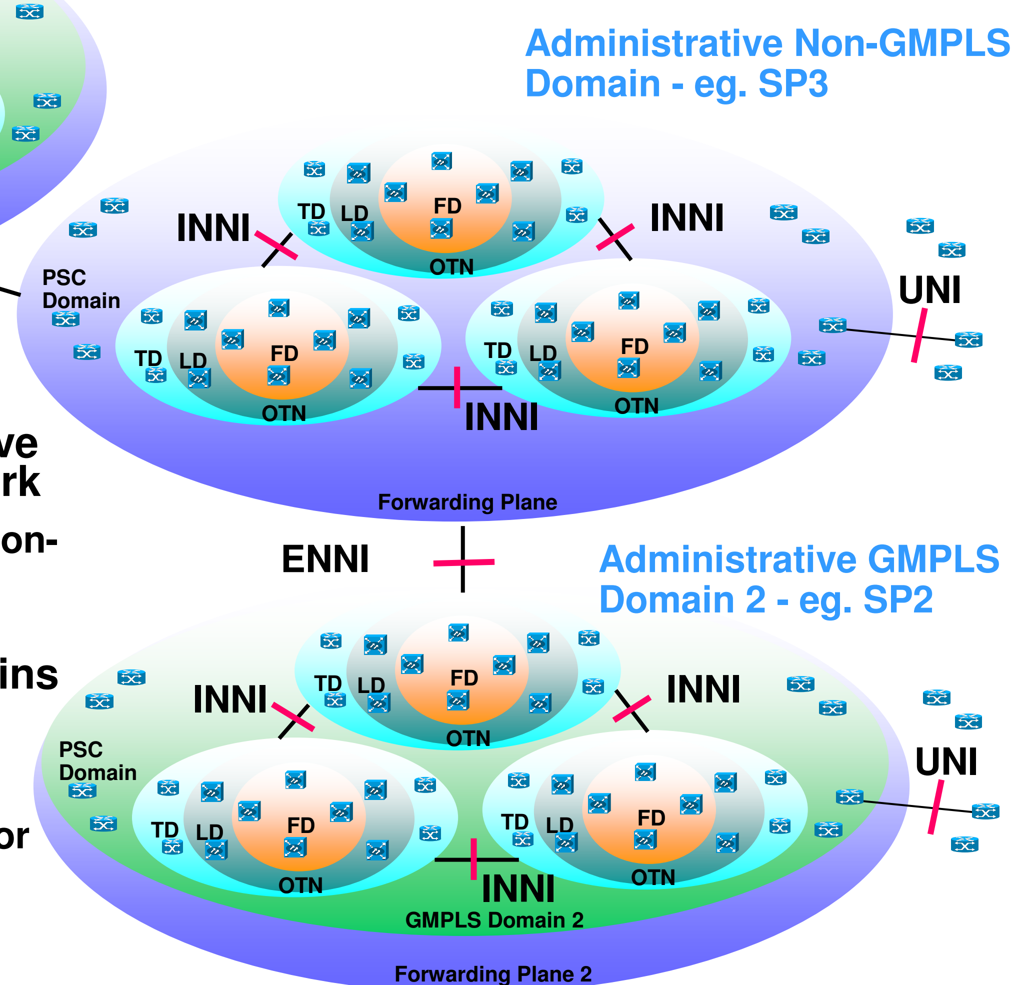
  Between areas of the same administrative domain & between controllers within areas aka. Internal-network-network-interface (I-NNI)

- Definition of GMPLS (RFC3473) compliant UNI
- GMPLS-OVERLAY & GMPLS-VPN

GMPLS Extensions for ASON
E2E Signaling over GMPLS and Non-GMPLS Domains

- No restricted use of other protocols within the control domain





- e2e signalling regardless of administrative boundaries & protocols within the network
  - Includes both GMPLS control domains & non-GMPLS control domains
- ASON support within a GMPLS control domain & between GMPLS control domains
- Backward compatibility with GMPLS signaling extensions for ASON
  - Regardless if transit nodes speak GMPLS or not

G.7713.2 / RFC3474 – RFC3473 Interworking

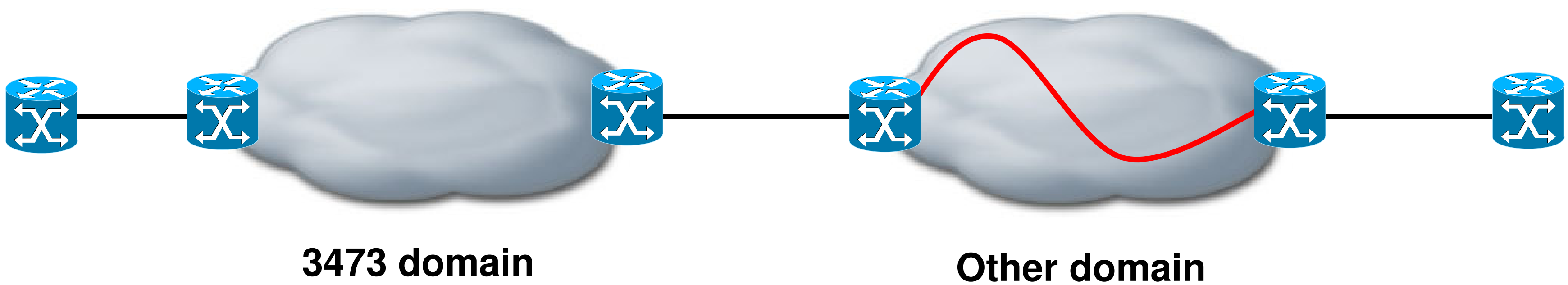
- **RFCs 3473 and 3474 interworking explained in**

  draft-ong-ccamp-3473-3474-iw-00.txt

  - Specifics are in the draft
  - More details and clarifications to be added

- **RFC 3474 Key Concepts**

  - Overlay or multiple domain model
    - Client interface (overlay)
    - ENNI (between domains)
  - Client address space (TNA)
    - Separate address space and format
  - Call-ID and related information
    - Carried transparently across intermediate nodes
  - Multi-session RSVP
    - e2e connection stitched together from multiple tunnels



GMPLS RSVP TE Signaling in Support of ASON

- Backward/Forward compatible with GMPLS RFCs (RFC 3471/73)
- Independence between UNI and E-NNI (agnosticism)
- Interworking (at UNI and/or E-NNI) must be impact free on GMPLS RFCs
- Intra-Domain and Inter-Domain Signaling
- Only define new object and procedures when strictly needed (max re-use principle)

Requirements	Info RFC 3474/76	Proposal
Soft Permanent Connection	Yes (SPC Label)	Yes (RFC 3473)
E2e Capability Negotiation	No	Yes
Call w/o Connection Setup	No	Yes
Call w/ (single) Connection Setup	Yes (limited to single hop sessions)	Yes
Multiple Connections per Call (add/remove)	No	Yes
Call Segments	No	Yes
Restart (CP failures)	Limited	Yes
Crankback Signaling	No	Ongoing
Backward Capability	No	Yes

draft-dimitri-ccamp-gmpls-rsvp-te-ason-01.txt

ASON Routing Requirements

- Requirements to support ASON routing
- Contains what’s missing in a “GMPLS ASON Routing Requirements” document
- Rules (same as for ASON signaling requirements)
  - No requirement that is not an ASON routing requirement (as decided by SG 15/Q12 and SG 15/Q14) will be considered in this document
- Functional Requirements
  - Support of multiple hierarchical levels
  - Support of multiple data plane layers
  - Support of architectural evolution
    - Levels, aggregation, segmentation

draft-alanqar-ccamp-gmpls-ason-routing-reqts-00.txt

Inter-Region / Inter-AS MPLS TE

- One common method for different "Regions"
- Requirements defined by TEWG
  - Inter-AS draft-ietf-tewg-interas-mpls-te-req-01.txt
  - Inter-area draft-boyle-tewg-interarea-reqts-00.txt
- Each Region may either nest or stitch the Inter-Region TE LSP into a "different" Intra-Region TE LSP to carry the ete Multi-Region TE LSP
  - RSVP-TE signaling based on LSP Hierarchy (for both nested and stitching)
  - Nesting of multiple inter-region LSPs into intra-region LSP
    - Control & forwarding plane scalability
- draft-ayyangar-inter-region-te-01.txt
  - Multiple LSP pieces nested or stitched together
  - Per region control
- draft-vasseur-inter-as-te-01.txt
  - Contiguous LSP ete
  - Head end control

Inter-AS MPLS TE

- draft-vasseur-inter-AS-TE-01.txt
- Defines signaling and routing mechanisms to make possible the creation of paths that span multiple IGP areas, multiple ASs, and multiple providers, including techniques for crankback ….
- Draft defines two cenarios for signaling and routing of TE LSP spanning multiple ASs
    - Per AS path computation
    - Distributed path computation between PSCs (ASBR)
- Can be used in combination with Hierarchical LSPs, crankback, …
- draft-vasseur-mpls-loose-path-reopt-01.txt proposes a set of mechanisms allowing a Head-end to exert a strict control on the TE LSP reoptimizing process and draft-ietf-mpls-nodeid-subobject-00.txt to support MPLS TE Fast Reroute

Two Scenarios

Scenario 1 - Per-AS TE LSP Path Computation

- No impact on RSVP/IGP scalability
- Semi-dynamic
- Small set of protocol extensions required
- No optimal e2e path
- Diverse path computation not always possible (path protection, load balancing)
- Call set up failure
- Support of e2e reoptimization (timer/event driven)
- Support of FRR Bypass for ASBR protection

Scenario 2 - Distributed Path Computation Server

- No impact on RSVP/IGP scalability
- Dynamic
- Implementation more complex
- Optimal e2e path
- Diverse path computation always possible (Path protection, load balancing)
- No call set up failure (not more than with single area/AS)
- Support of e2e reoptimization
- Support of FRR Bypass for ASBR protection
- TE LSP local protection recommended

Scenario 1 and 2 are both compliant with set of requirements defined in draft-ietf-tewg-interas-mpls-te-req-00.txt

Working Group Drafts

- **WG last call soon**
  - **GMPLS UNI**
    - **RSVP Support for Overlay Model**
    - draft-ietf-ccamp-gmpls-overlay-02.txt
    - **GMPLS Signaling Extensions for G.709 OTN Control**
    - draft-ietf-ccamp-gmpls-g709-04.txt
- **New revisions soon**
  - **Exclude Routers – Extensions to RSVP-TE**
    - draft-ietf-ccamp-rsvp-te-exclude-route-00.txt
- **Further discussions**
  - **ASON requirements** (draft-ietf-ccamp-gmpls-ason-reqts-04.txt)
  - **Protection and Recovery drafts**
  - **GMPLS MIBs**

Interaction with other WGs

- **TEWG**
  - **Multi-area AS requirements**
  - **draft-ietf-tewg-interas-mpls-te-req**
- **MPLS**
  - **Ptmp LSPs - requirements and solutions include all switching types**
  - **draft-yasukawa-mpls-p2mp-requirements)**
- **OSPF / IS-IS**
  - **GMPLS extensions complete**
  - **May interact for solutions to ASON routing requirements**
- **IPO**
  - **IP over Optical Networks – a framework**
  - **draft-ietf-ipo-framework**
  - **Just completing IESG review**

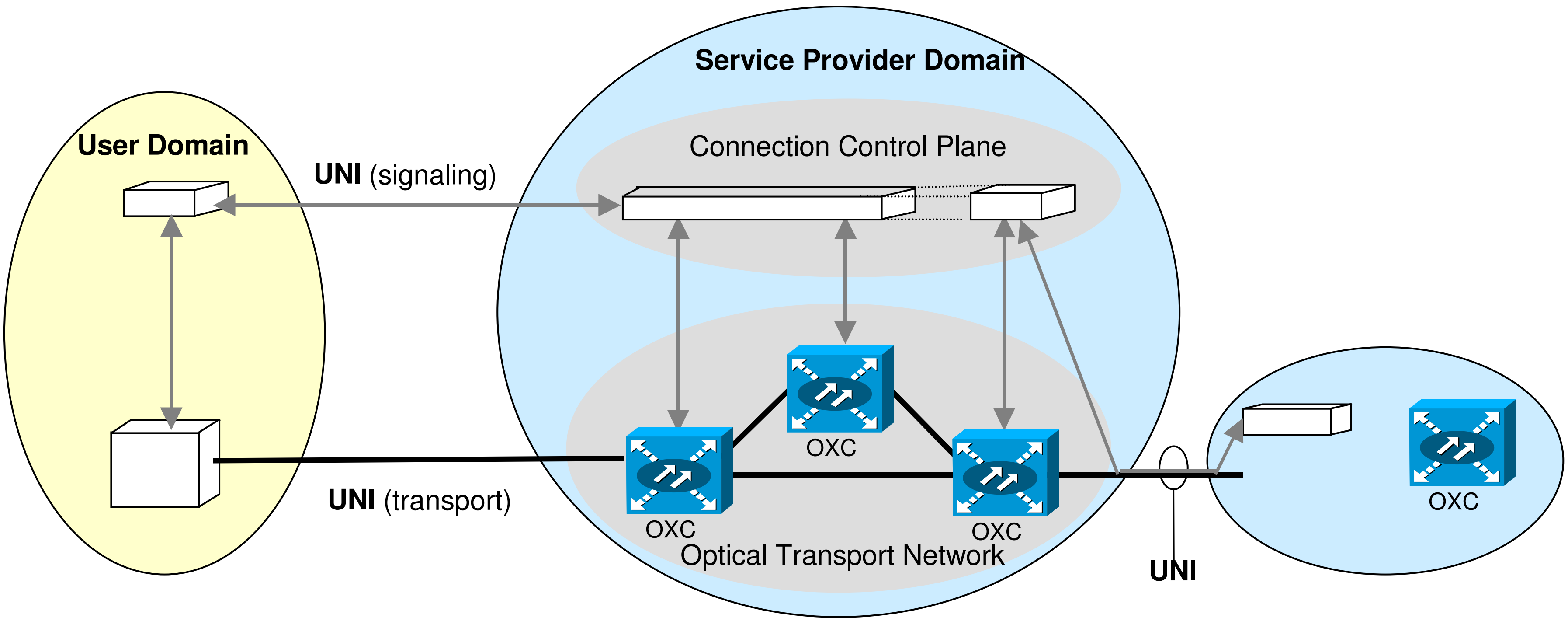
What is O-UNI ?

A Signaling Interface (demarcation) between the Optical User Equipment and the Service Provider Transport Network !

Optical User Equipment (Client)

- **Service Provider, Enterprise, Organization**
- **IP router, SONET/SDH, ATM NEs**

Where does O-UNI fit in the network ?



Enables Subscribers via signaling to request circuits from Service Provider Networks based on required service parameters

What is O-NNI ?

Cisco.com

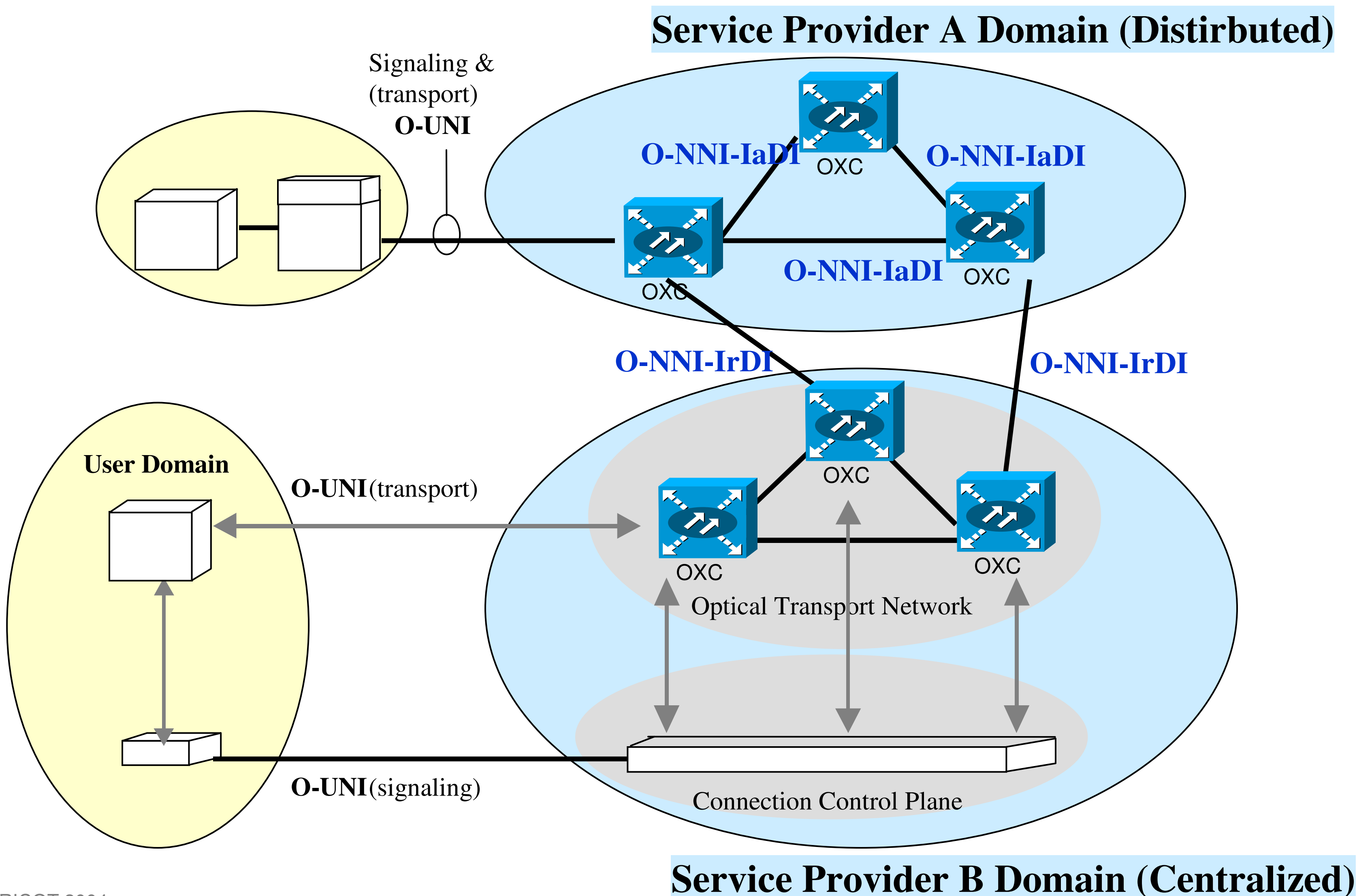
A signaling & routing interface between Optical Networking Elements in the same or different administrative domains !

<u>O-NNI Key Characteristics</u>

- Intra-Domain (IaDI) NNI interface
- Inter-Domain (IrDI) NNI interface
- Distributed Model, Centralize Model
- Examples of Optical Networking Elements with O-NNI include OXCs & OADMs

Where does O-NNI fit in the network ?

Cisco.com



O-UNI
Carrier Identified Potential Applications

- **Bandwidth On Demand**

  High bandwidth transient, time of day network reconfiguration, multiple optical client types

- **Optical Virtual Private Network**

  Shared optical infrastructure to provide virtual dedicated circuit network to customers with contracted range of control by customers

O-UNI Key Features

- Signaling Interface between Optical Network & Clients
  - IP routers, ATM switches, SONET ADMs
- UNI Functional Components
  - Neighbor Discovery & control channel maintenance
    - Control channel configuration
    - Hello initiation & link verification (up/down status)
    - Neighbor discovery information retrieval
  - Service discovery & address registration
    - Discovery of service attributes
    - Service Granularity (min, max bandwidth)
    - Signaling protocols (RSVP-TE/LDP)
  - Signaling Message Exchange
    - Connection Create, Delete, Status Inquiry

OIF O-UNI 1.0 Key Protocols

- **All signaling & control messages**
  - **IETF IP protocols used**
- **In-Fiber IP Control Channel**
  - **DCC: PPP in HDLC IETF RFC1662**
  - **Dedicated channel: PPP over SONET/SDH IETF RFC2615**
- **Signaling Protocol**
  - **IETF RSVP-TE, LDP-based**
- **Neighbor Discovery, Service Discovery**
  - **IETF LMP protocol (draft status) based**
- **Routing Protocol - Not Applicable**

OIF O-UNI 1.0 Key Connection Attributes

Key Connection Attributes beyond Src & Dst TNA & ports

Connection ID (M)	**Contract ID (O)**
Framing Type (M)	**Transparency (M)**
Bandwidth (M)	**Concatenation (M)**
Directionality (O)	**Payload (O)**
Service level (O)	**Diversity (O)**

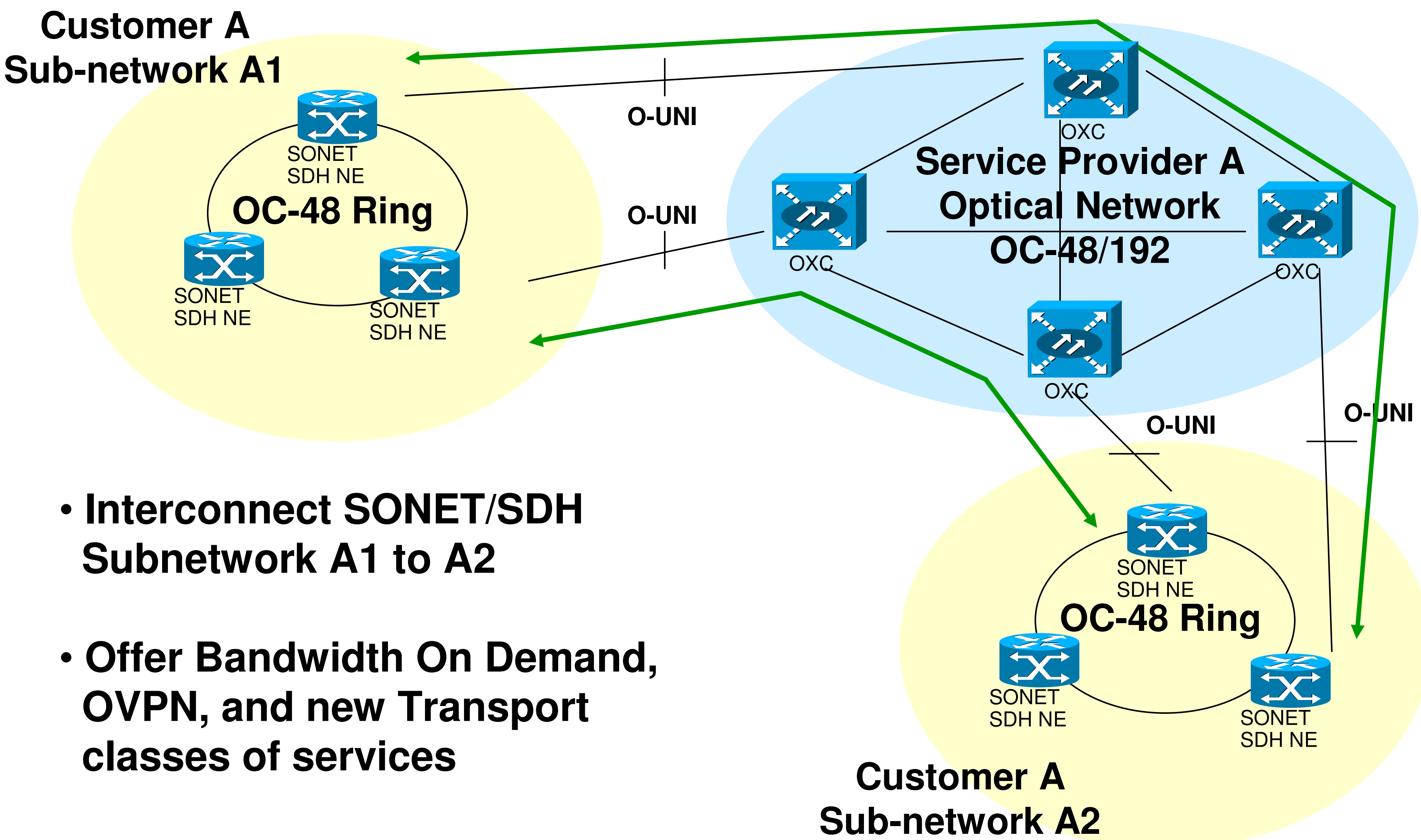
UNI 1.0 Security Provisions

Cryptographic Authentication as per RSVP-TE & LDP

thus provides original authentication and message integrity

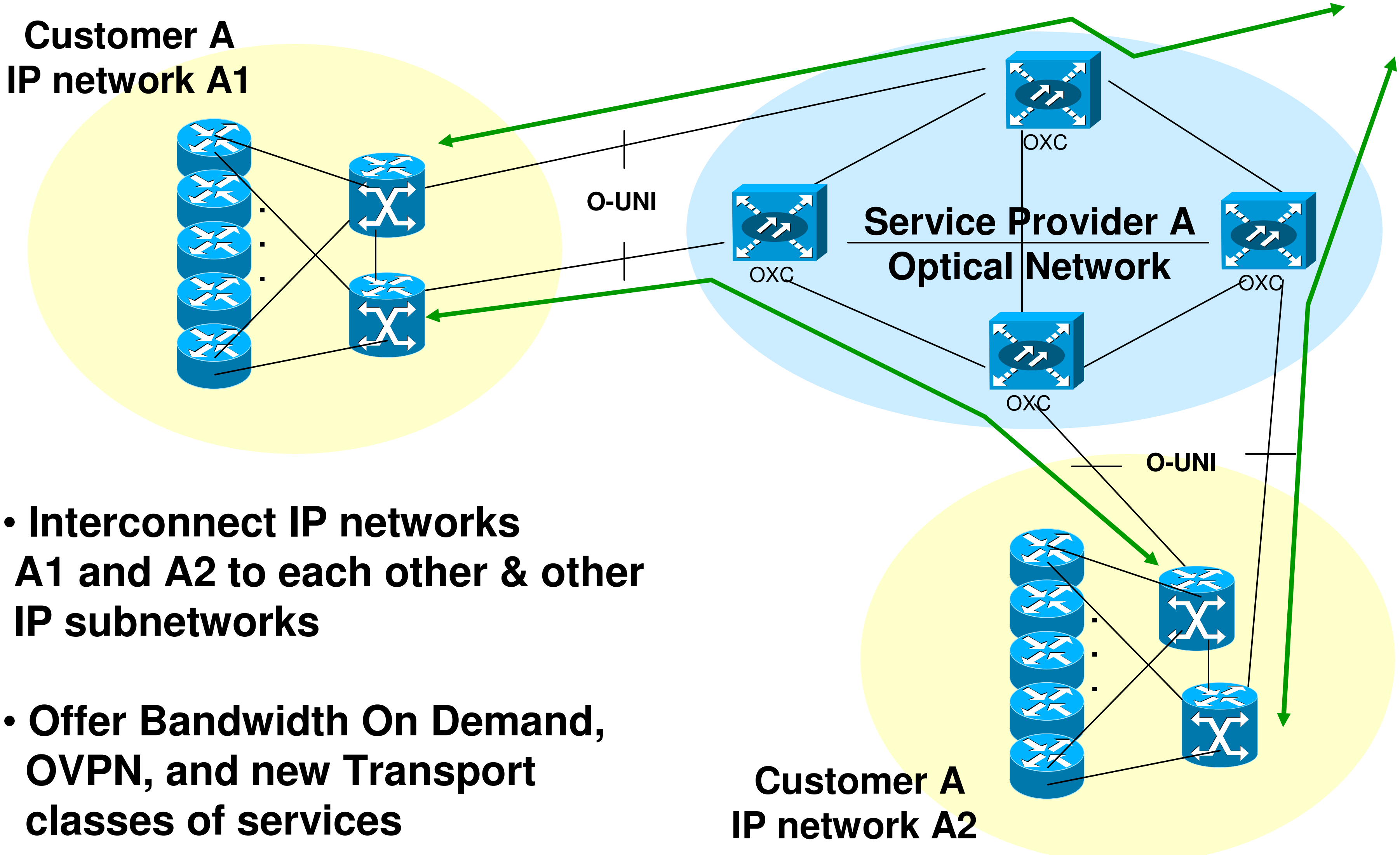
HMAC-MD5 is specified for UNI 1.0

O-UNI
Transport Network Applications



- Interconnect SONET/SDH Subnetwork A1 to A2
- Offer Bandwidth On Demand, OVPN, and new Transport classes of services

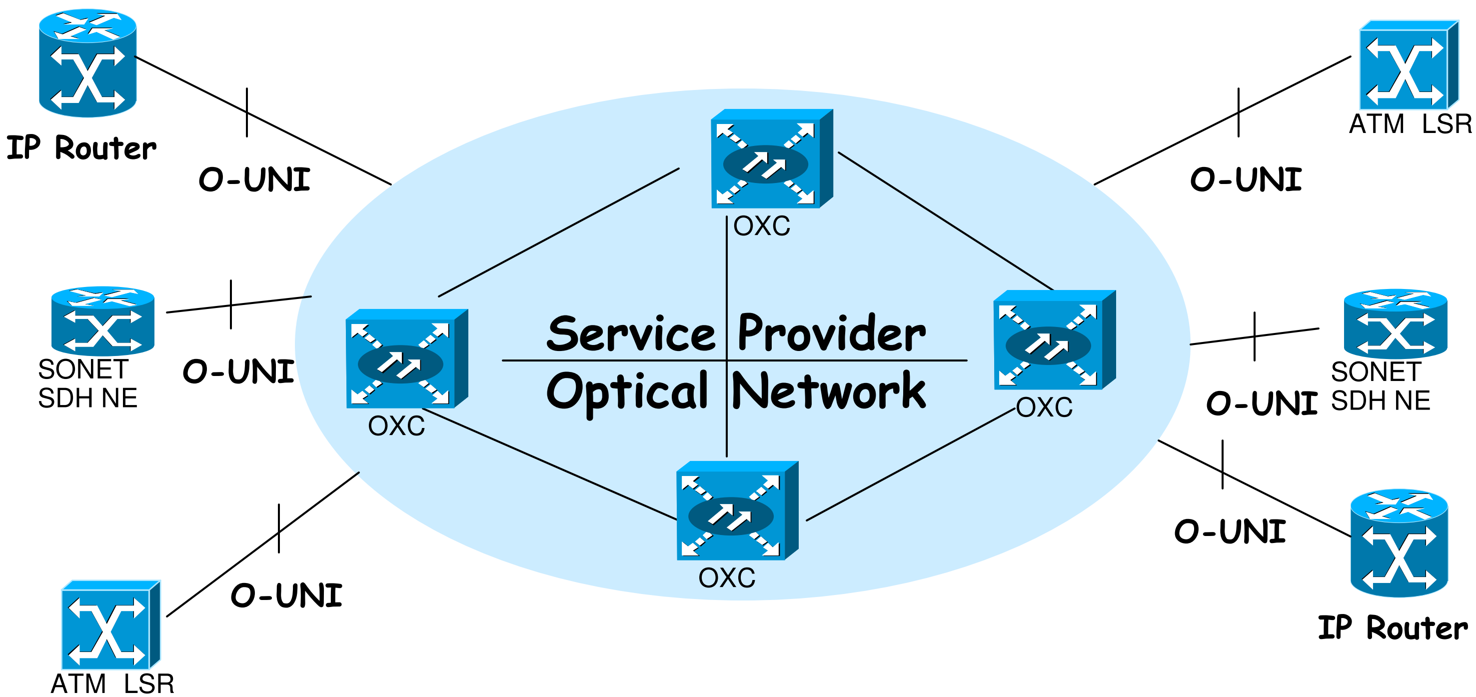
O-UNI IP Router Network Applications



- Interconnect IP networks A1 and A2 to each other & other IP subnetworks

- Offer Bandwidth On Demand, OVPN, and new Transport classes of services

O-UNI Multi-Service Network Applications



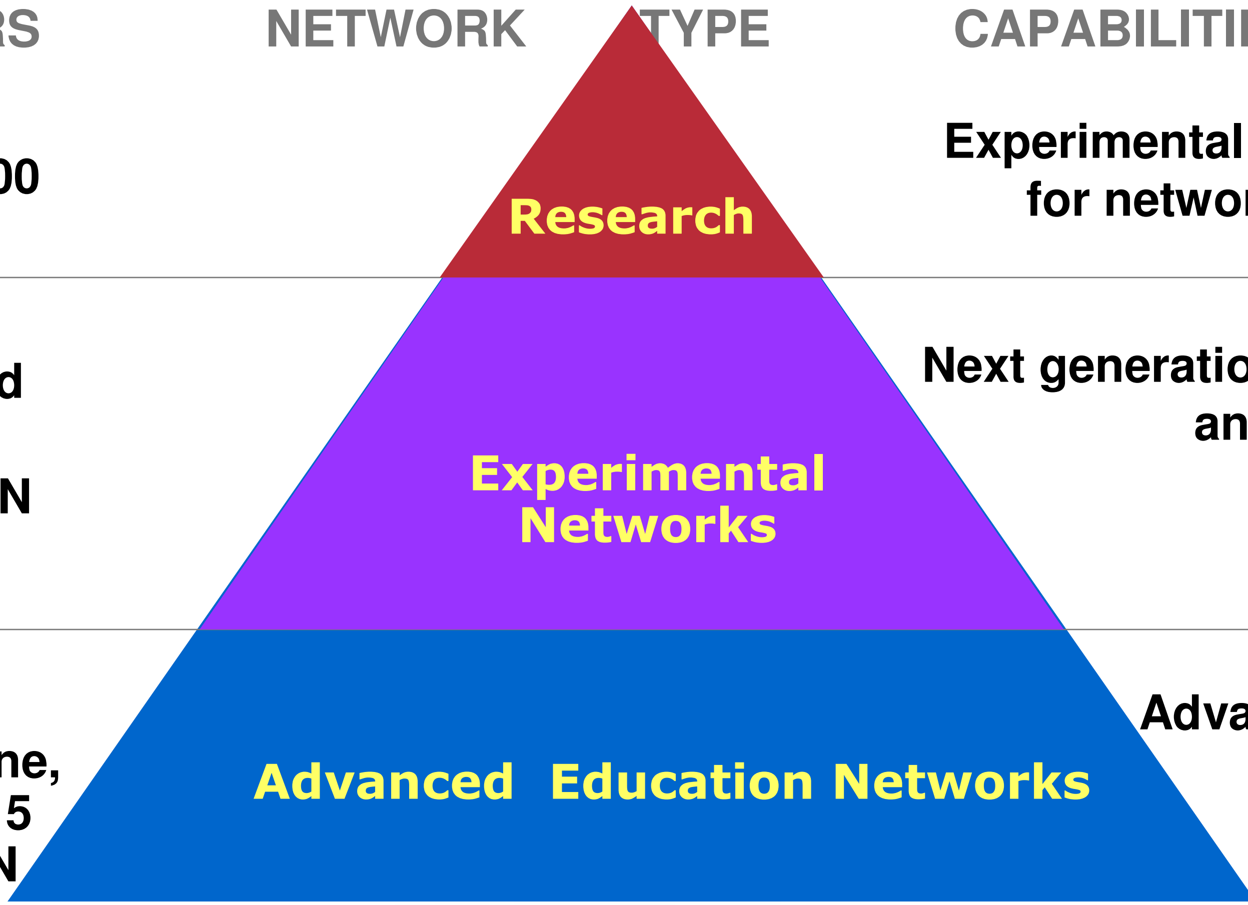
Service Provider offering dynamic optical paths for myriad of optical client equipment and networks

Offer Bandwidth On Demand, OVPN, and new Transport classes of services

Research & Education Network Tiers

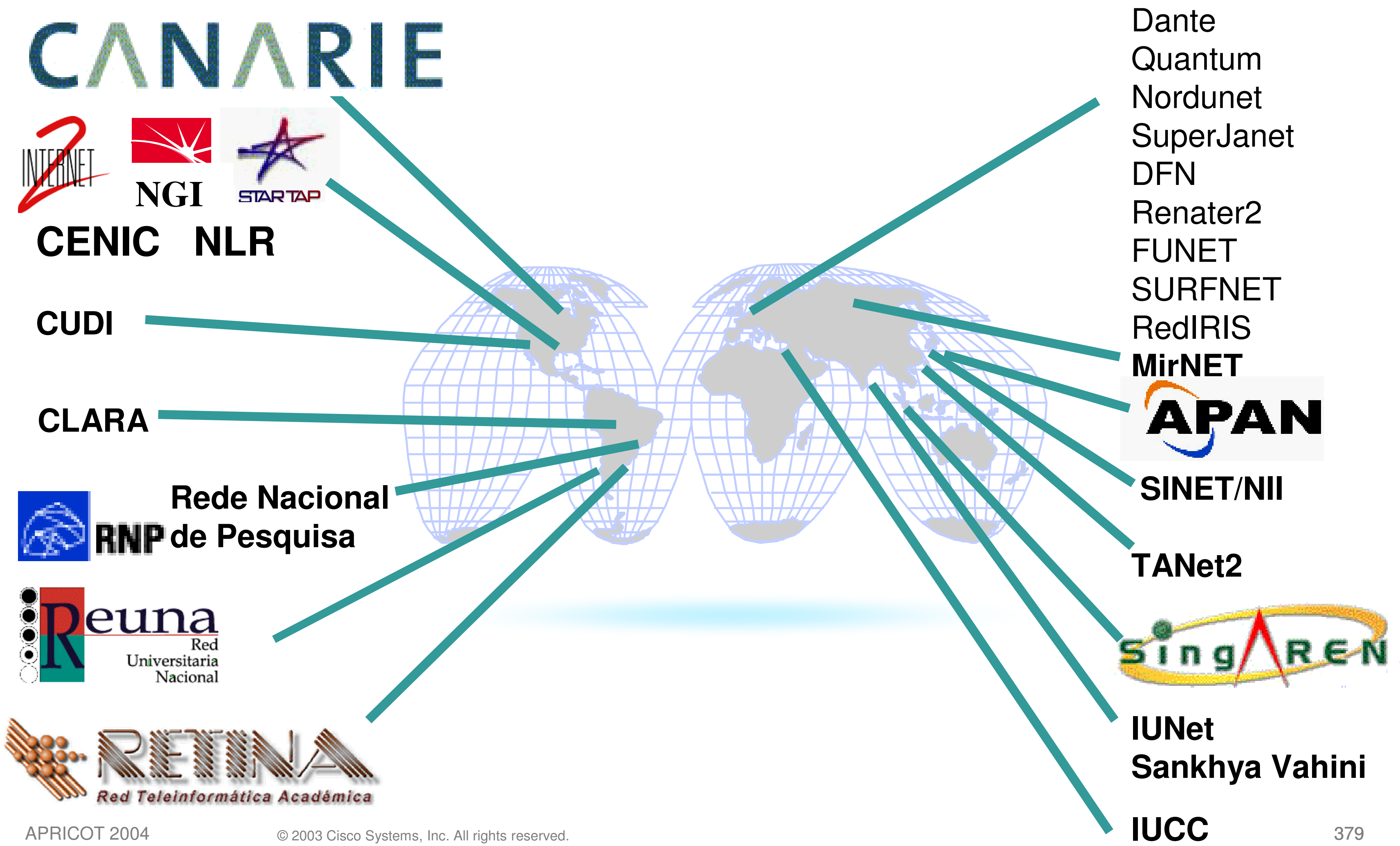
Cisco.com

LEADERS | NETWORK TYPE | CAPABILITIES/USERS



LEADERS	NETWORK TYPE	CAPABILITIES/USERS
Web100 NLR	Research	Experimental environments for network researchers
Teragrid WIDE CALREN NLR	Experimental Networks	Next generation architecture and applications for research community
I2-Abilene, SurfNet 5 CALREN	Advanced Education Networks	Advanced services for education
ISPs	Commodity Internet	General Use

Advanced Internet Initiatives



http:// … *Advanced Internets*

Cisco.com

www.canarie.ca

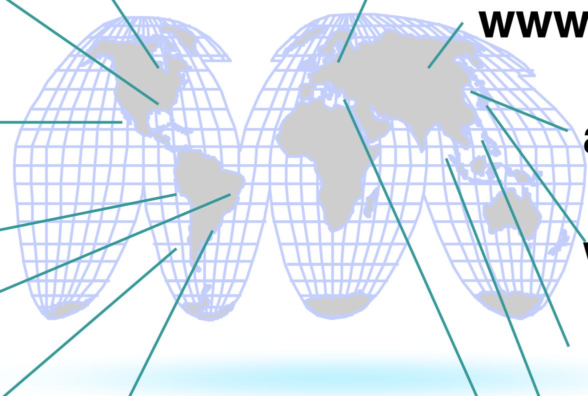
www.internet2.edu
www.ngi.gov
www.startap.net
www.cenic.org

www.cudi.edu.mx/

www.rnp.br/

www.reuna.cl/

www.retina.ar



www.dante.net/quantum.html
www.nordu.net
www.ukerna.ac.uk www.dfn.de
www.renater.fr www.surfnet.nl
www.csc.fi/english/funet

www.friends-partners.org/friends/mirnet/

apan.or.kr

www.nii.ac.jp

www.tanet2.net.tw/

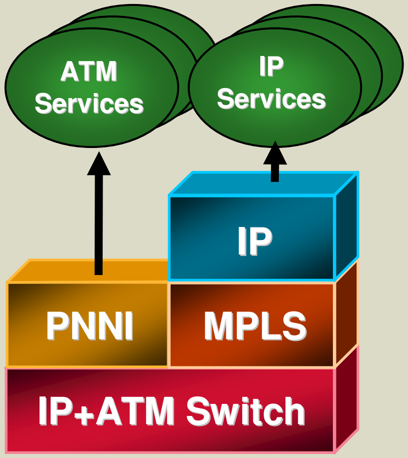
www.singaren.net.sg

www.machba.ac.il/index.html

Summary

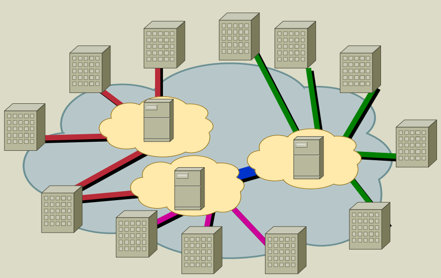
Azhar Sayeed

MPLS: The Key Technology for the delivery of L2 & L3 Services



IP+ATM: MPLS Brings IP and ATM Together

- eliminates IP "over" ATM overhead and complexity
- one network for Internet, Business IP VPNs, and transport

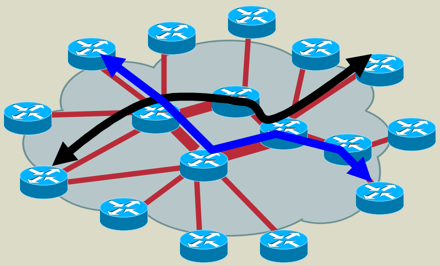


Network-Based VPNs with MPLS: a Foundation for Value Added Service Delivery

- flexible user and service grouping (biz-to-biz)
- flexibility of IP and the QoS and privacy of ATM
- enables application and content hosting inside each VPN
- transport independent
- low provisioning costs enable affordable managed services

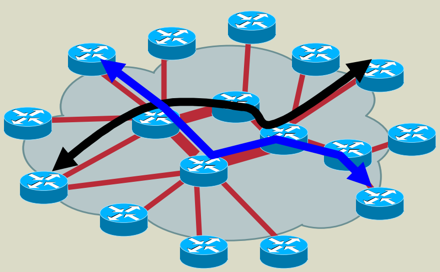
MPLS: The Key Technology for the delivery of L2 & L3 Services

Cisco.com



MPLS Traffic Engineering

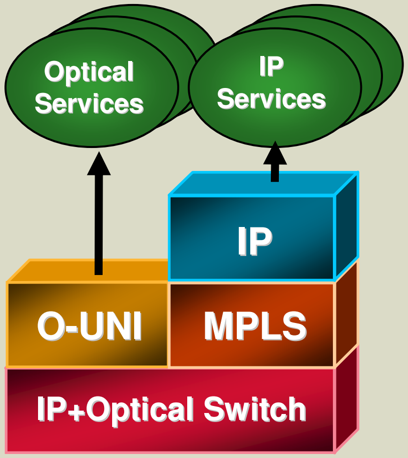
- Provides Routing on diverse paths to avoid congestion
- Better utilization of the network
- Better availability using Protection Solution (FRR)



Guaranteed Bandwidth Services

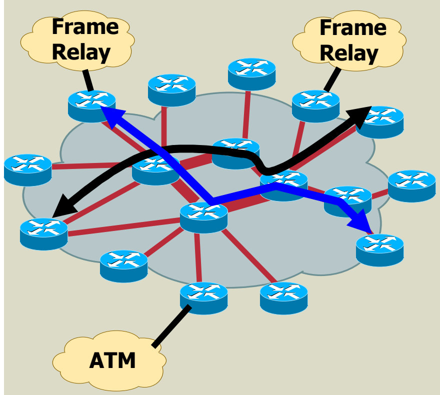
- Combine MPLS Traffic Engineering and QoS
- Deliver Point-to-point bandwidth guaranteed pipes
- Leverage the capability of Traffic Engineering
- Build Solution like Virtual leased line and Toll Trunking

MPLS: The Key Technology for the delivery of L3 Services



IP+Optical Integration

- eliminates IP “over” Optical Complexity
- Uses MPLS as a control Plane for setting up lightpaths (wavelengths)
- one control plane for Internet, Business IP VPNs, and optical transport



Any Transport over MPLS

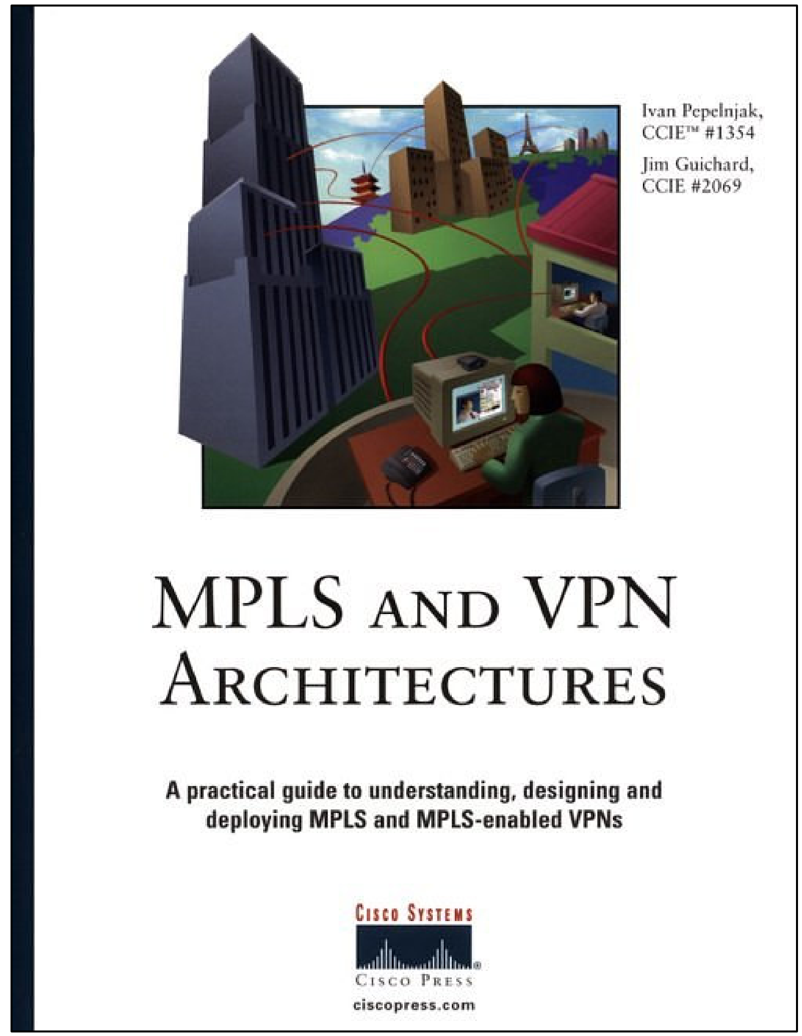
- Transport ATM, FR, Ethernet, PPP over MPLS
- Provide Services to existing installed base
- Protect Investment in the installed gear
- Leverage capabilities of the packet core
- Combine with other packet based services such as MPLS VPNs

Recommended Reading

Cisco.com

- **MPLS and VPN Architectures by Jim Guichard and Ivan Pepelnjak**

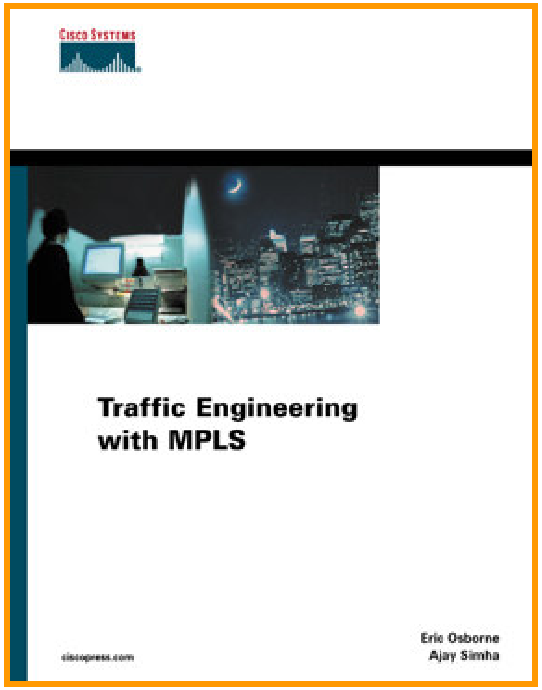
  ISBN: 1-58705-002-1



Recommended Reading

- **Traffic Engineering with MPLS**

  ISBN: 1-58705-031-5



Questions?

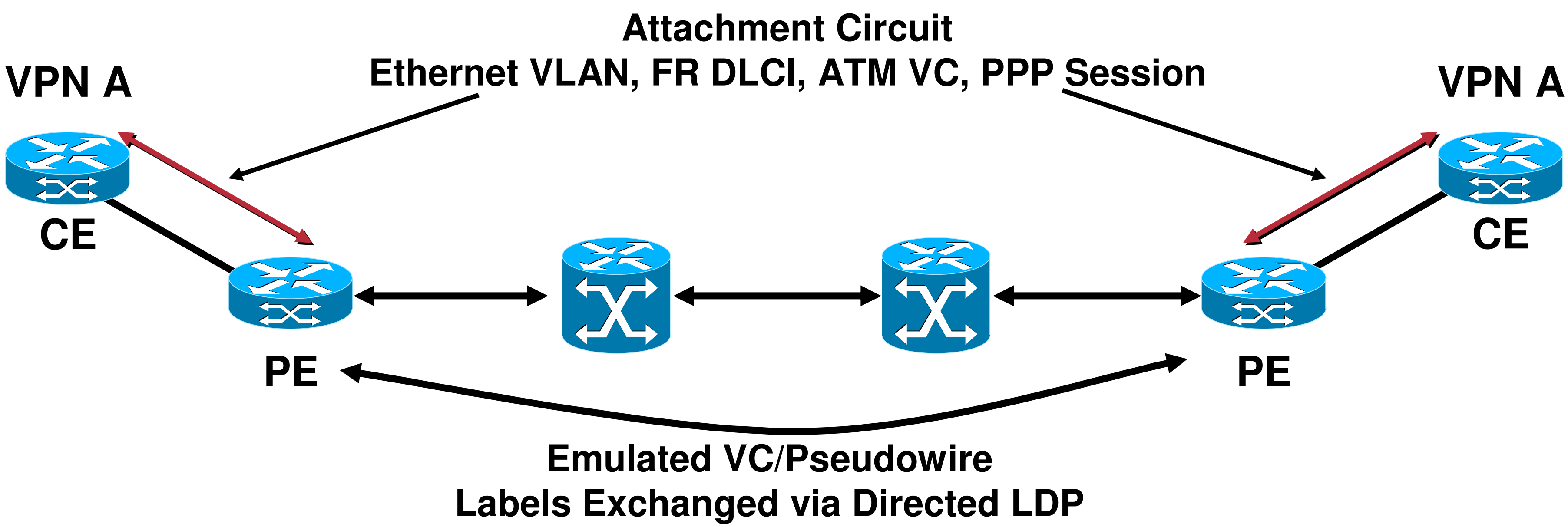
Layer 2 VPNs

Layer 2 VPNs

Similar to L3VPN

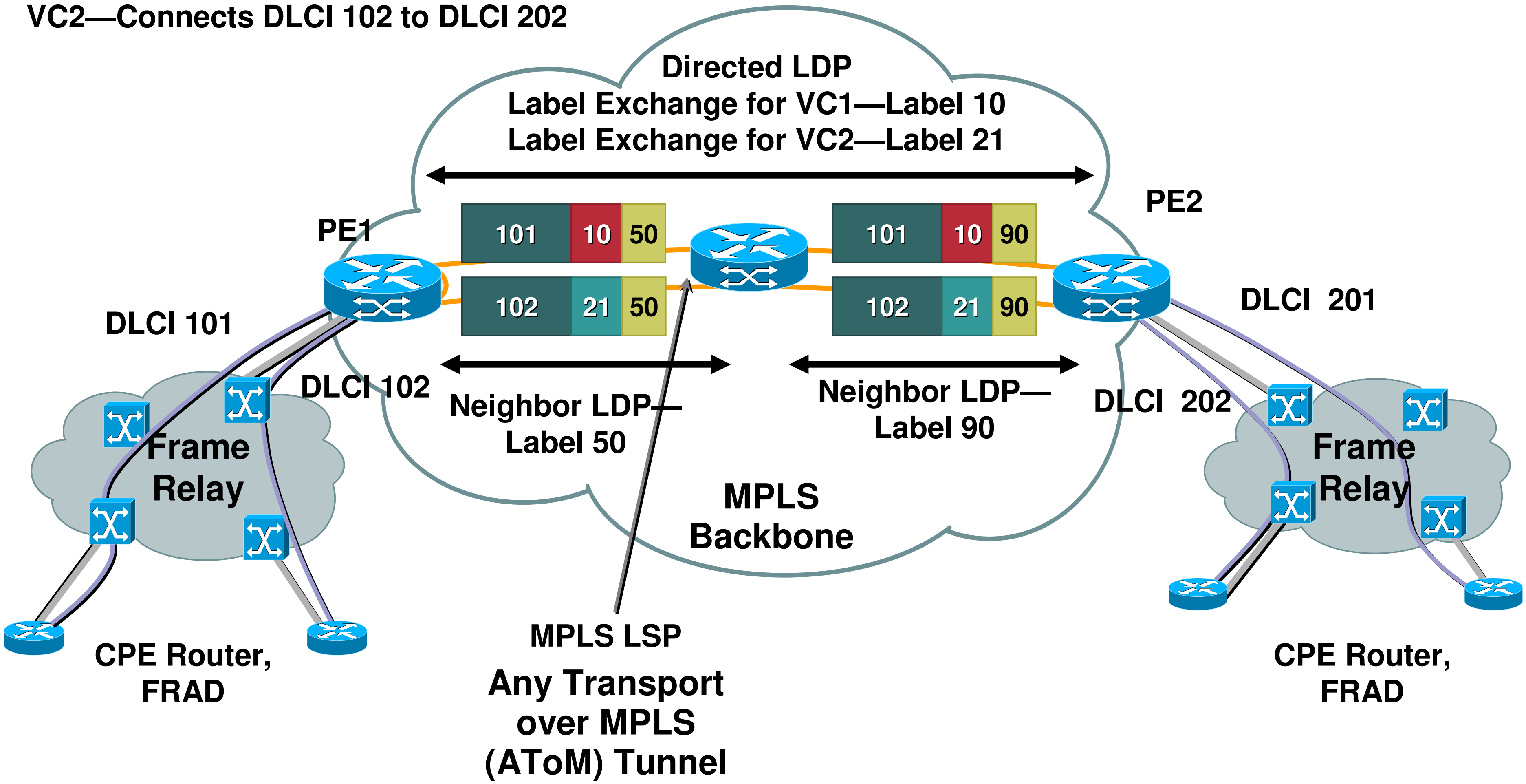
- Designate a label for the circuit
- Exchange that label information with the egress PE
- Encapsulate the incoming traffic (layer 2 frames)
- Apply label (learnt through the exchange)
- Forward the MPLS packet (l2 encapsulated to destination on an LSP)
- At the egress
  - Lookup the L2 label
  - Forward the packet onto the L2 attachment circuit

Architecture



Frame Relay over MPLS—Example

VC1—Connects DLCI 101 to DLCI 201
VC2—Connects DLCI 102 to DLCI 202



Summary

- Easy way of transporting layer 2 frames
- Can be used to transport ATM AAL5 frames, Cells, FR DLCI, PPP sessions, Ethernet VLANs
- Point-to-point transport with QoS guarantees
- Combine with TE and QoS to emulate layer 2 service over a packet infrastructure
- Easy migration towards network convergence