MPLS Tutorial SANOG

Introduction to MPLS Traffic Engineering

© 2005 Cisco Systems, Inc. All rights reserved.

Agenda

- Introduction
- Traffic Engineering by tweaking IGPs
- Limitations of the Overlay Model

© 2006 Cisco Systems, Inc. All rights reserved.

What is Traffic Engineering??

- **Preventing a situation where some parts of a service provider network are over-utilized (congested), while other parts under-utilized**
- **Reduce the overall cost of operations by more efficient use of bandwidth resources**

The ultimate goal is cost saving !

© 2006 Cisco Systems, Inc. All rights reserved.

ISSUES WITH IGP ROUTING

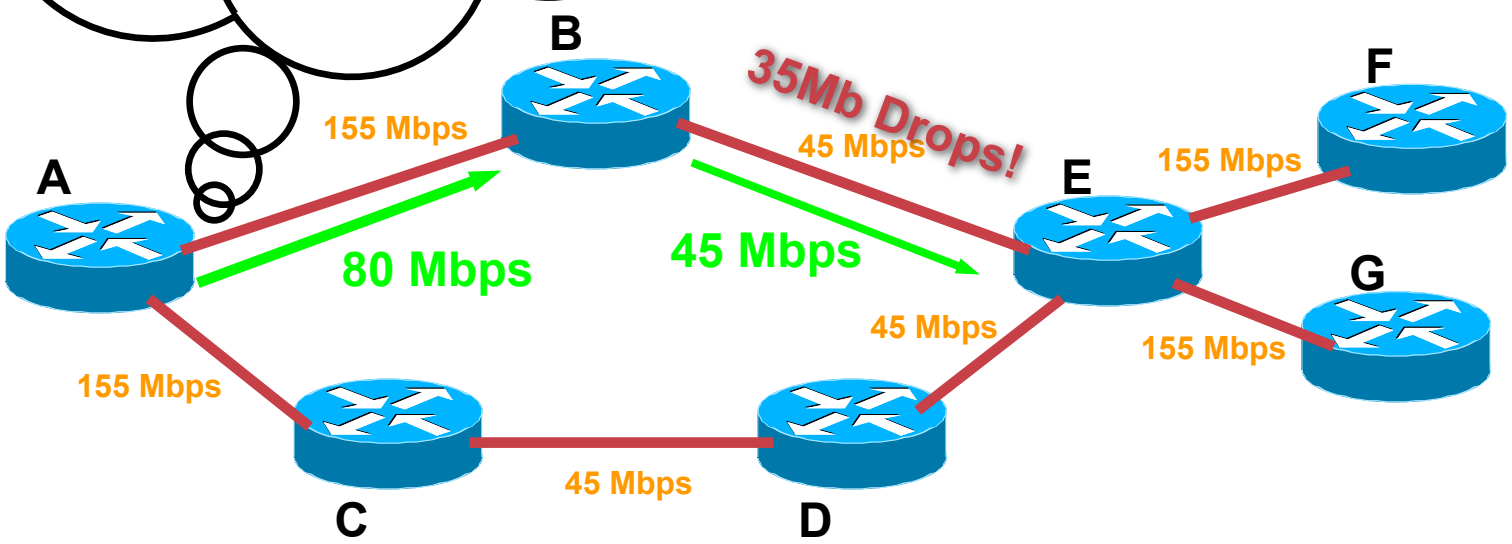
- IGPs forward packets based on shortest path (metric).
- Flows from multiple sources may go over some common link(s) causing congestion.
- Alternate longer and underutilized path will not be used.
- IGP metric change may have side effects.

© 2006 Cisco Systems, Inc. All rights reserved.

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

- Assume “A” has 40Mb of traffic for “F” and 40Mb of traffic for “G”
- Some links are 45 Mbps, some are 155 Mbps
- Massive (44%) packet loss between “B” and “E”
- Changing path to A->C->D->E won’t help

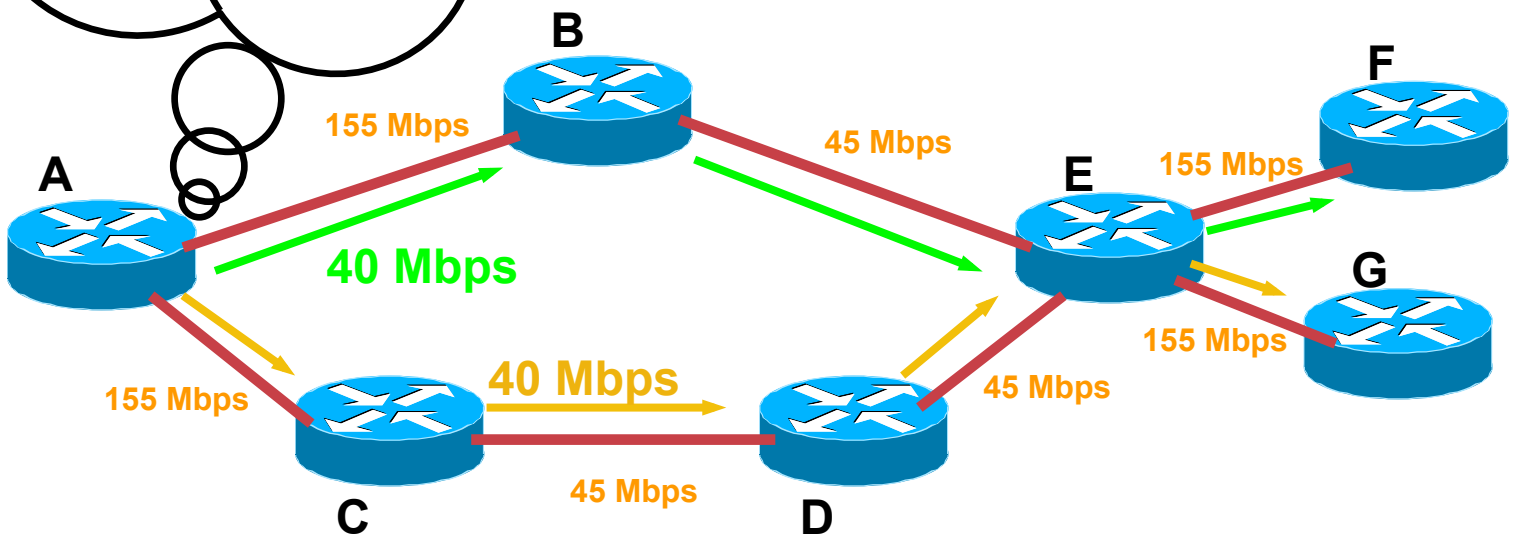


© 2006 Cisco Systems, Inc. All rights reserved.

MPLS-TE Example

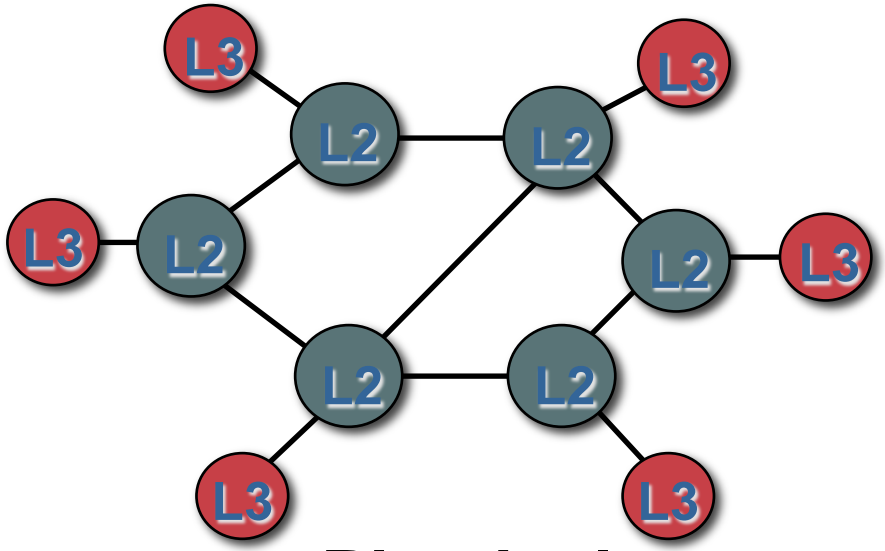
Node	Next-Hop	Cost
B	B	10
C	C	10
D	C	20
E	B	20
F	Tunnel0	30
G	Tunnel1	30

- Assume “A” has 40Mb of traffic for “F” and 40Mb of traffic for “G”
- “A” computes paths on properties other than just shortest cost (available bandwidth)
- No congestion!

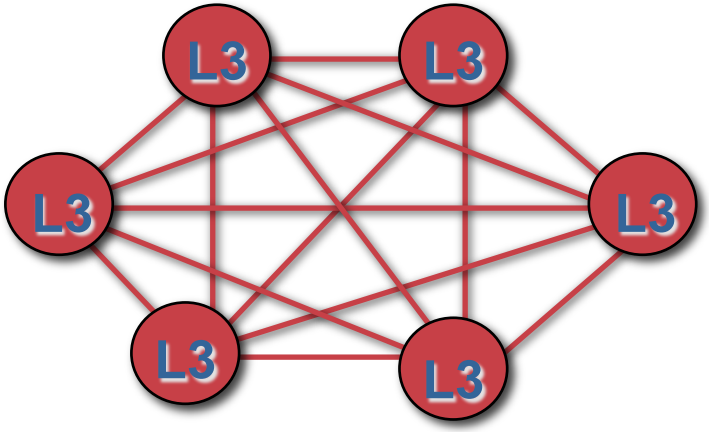


© 2006 Cisco Systems, Inc. All rights reserved.

The “Overlay” Solution



Physical



Logical

- Routing at layer 2 (ATM or FR) is used for traffic engineering
- Full mesh of VCs between routers. Each router has a direct VC to every other router in the mesh.

© 2006 Cisco Systems, Inc. All rights reserved.

“Overlay” solution: drawbacks

- **Extra network devices (cost)**
- **More complex network management (cost)**
  - **two-level network without integrated network management**
  - **additional training, technical support, field engineering**
- **IGP routing scalability issue for meshes**

© 2006 Cisco Systems, Inc. All rights reserved.

Traffic engineering with Layer 3 what is missing ?

- **Path Computation based just on IGP metric is not enough.**
- **Packet forwarding in IP network is done on a hop by hop basis, derived from IGP.**
- **Support for “explicit” routing (aka “source routing”) is not available.**

© 2006 Cisco Systems, Inc. All rights reserved.

Motivation for Traffic Engineering

- **Increase efficiency of bandwidth resources**
  - **Prevent over-utilized (congested) links whilst other links are under-utilized**
- **Ensure the most desirable/appropriate path for some/all traffic**
  - **Explicit-Path overrides the shortest path selected by the IGP**
- **Replace ATM/FR cores**
  - **PVC-like traffic placement without IGP full mesh and associated O(N^2) flooding**
- **The ultimate goal is COST SAVING**
  - **Service development also progressing**

© 2006 Cisco Systems, Inc. All rights reserved.

TE tunnel basics

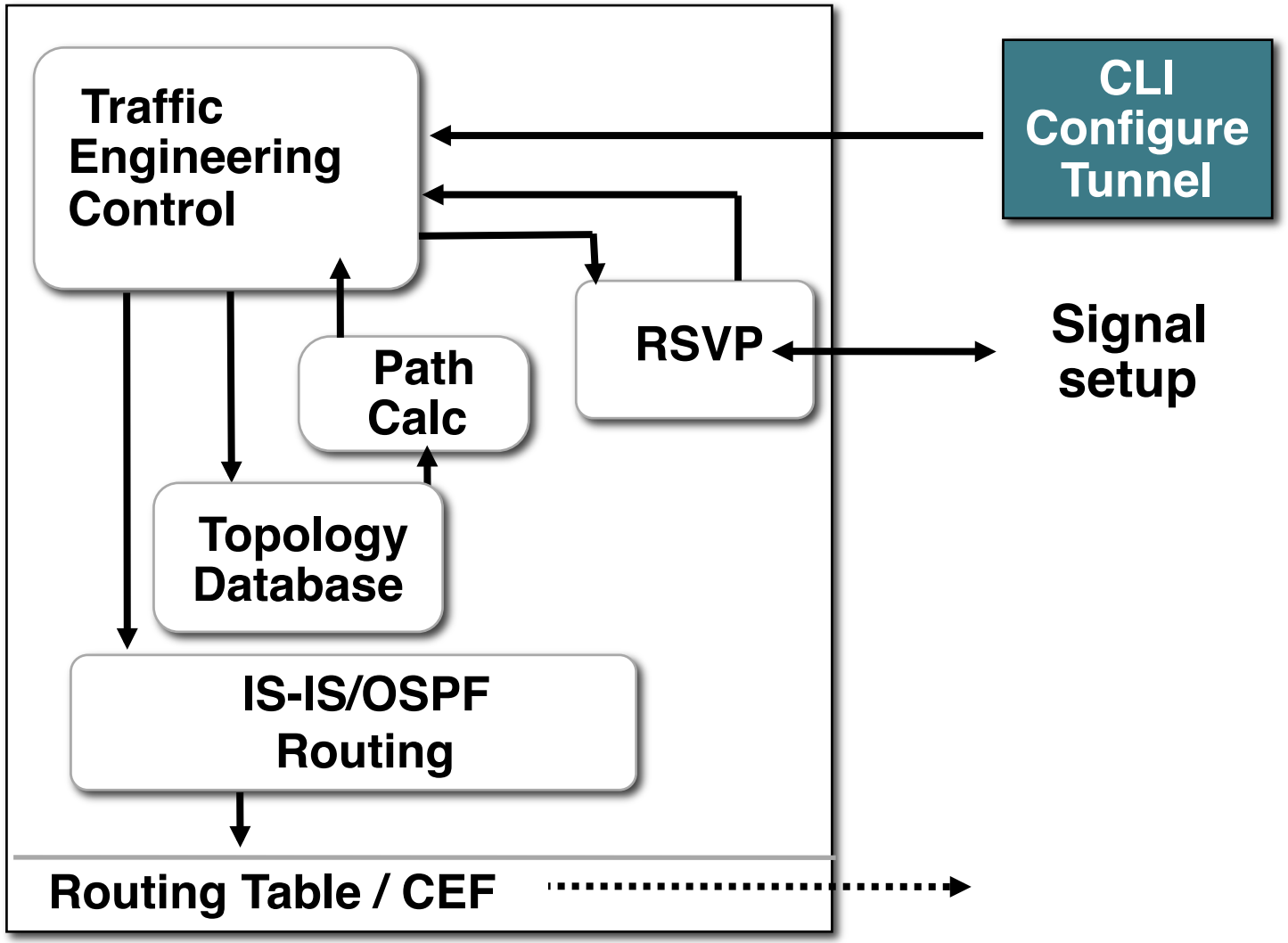
© 2005 Cisco Systems, Inc. All rights reserved.

Agenda

- MPLS-TE router operation
- Tunnel attributes:
  - Bandwidth
  - Priority
  - Metric selection
  - Affinity
- Tunnel Path selection

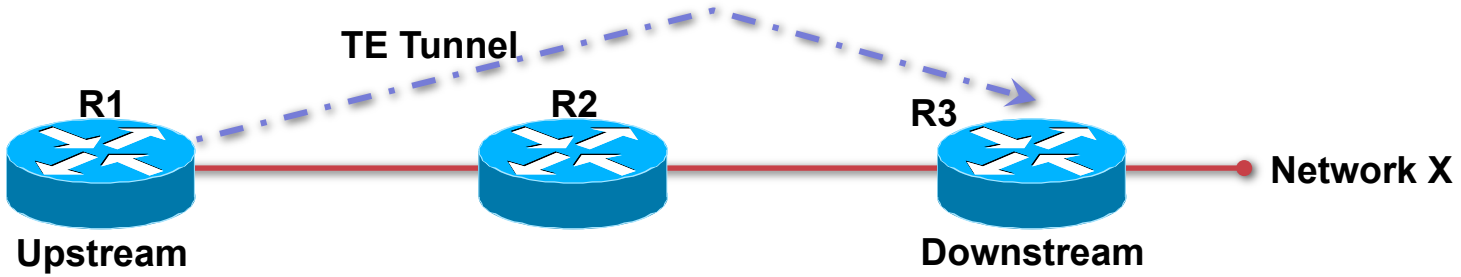
© 2006 Cisco Systems, Inc. All rights reserved.

Tunnel Setup



© 2006 Cisco Systems, Inc. All rights reserved.

A Terminology Slide—Head, Tail, LSP, etc.



- 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)

© 2006 Cisco Systems, Inc. All rights reserved.

Trunk Attributes

- Tunnel attributes are characteristics the tunnel requires to have on the links along the LSP.
- Configured at the head-end of the trunk
- These are:
  - Bandwidth
  - Priority
  - Metric selection (TE vs. IGP metric)
  - Affinity

```
interface Tunnel0
  tunnel mpls traffic-eng bandwidth Kbps
  tunnel mpls traffic-eng priority pri [hold-pri]
   tunnel mpls traffic-eng path-selection metric {te|igp}
   tunnel mpls traffic-eng affinity properties [mask]
```

© 2006 Cisco Systems, Inc. All rights reserved.

Tunnel Bandwidth

```
tunnel mpls traffic-eng bandwidth Kbps
```

- Bandwidth required by the tunnel across the network
- If not configured, tunnel is requested with zero bandwidth.

© 2006 Cisco Systems, Inc. All rights reserved.

Priority

```
tunnel mpls traffic-eng <S> {H}
```

- Configured on tunnel interface
- S = setup priority (0–7)
- H = holding priority (0–7)
- Lower number means higher priority

© 2006 Cisco Systems, Inc. All rights reserved.

Priority

- Setup priority of new tunnel on a link is compared to the hold priority of an existing tunnel
- New tunnel with better setup priority will force preemption of already established tunnel with lower holding priority
- Preempted tunnel will be torn down and will experience traffic black holing. It will have to be re-signaled
- Recommended that S=H; if a tunnel can setup at priority “X”, then it should be able to hold at priority “X” too!
- Configuring S > H is illegal; tunnel will most likely be preempted
- Default is S = 7, H = 7

© 2006 Cisco Systems, Inc. All rights reserved.

Metric Selection (TE vs. IGP metric)

```
tunnel mpls traffic-eng path-
  selection metric {te|igp}
```

- Configure admin weight == interface delay
- Configure VoIP tunnels to use TE metric to calculate the path cost
- Can be used as a Delay-sensitive metric

© 2006 Cisco Systems, Inc. All rights reserved.

Tunnel Affinity

- **Tunnel is characterized by a**
  - **Tunnel Affinity: 32-bit resource-class affinity**
  - **Tunnel Mask: 32-bit resource-class mask (0= don't care, 1= care)**

Link is characterized by a 32-bit resource-class attribute string called Link Affinity

Default-value of tunnel/link bits is 0

Default value of the tunnel mask = 0x0000FFFF

© 2006 Cisco Systems, Inc. All rights reserved.

Tunnel Affinity (Cont.)

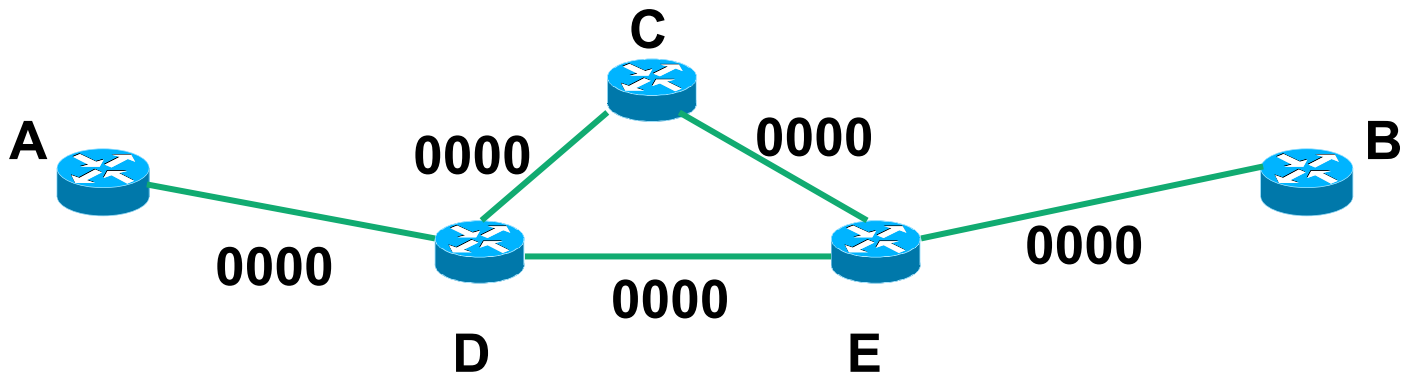
- **Affinity helps select which tunnels will go over which links**
- **A network with OC-12 and Satellite links will use affinities to prevent tunnels with VoIP traffic from taking the satellite links**

Tunnel can only go over a link if

(Tunnel Mask) AND (Link Affinity) == Tunnel Affinity

© 2006 Cisco Systems, Inc. All rights reserved.

Example0: 4-bit string, default



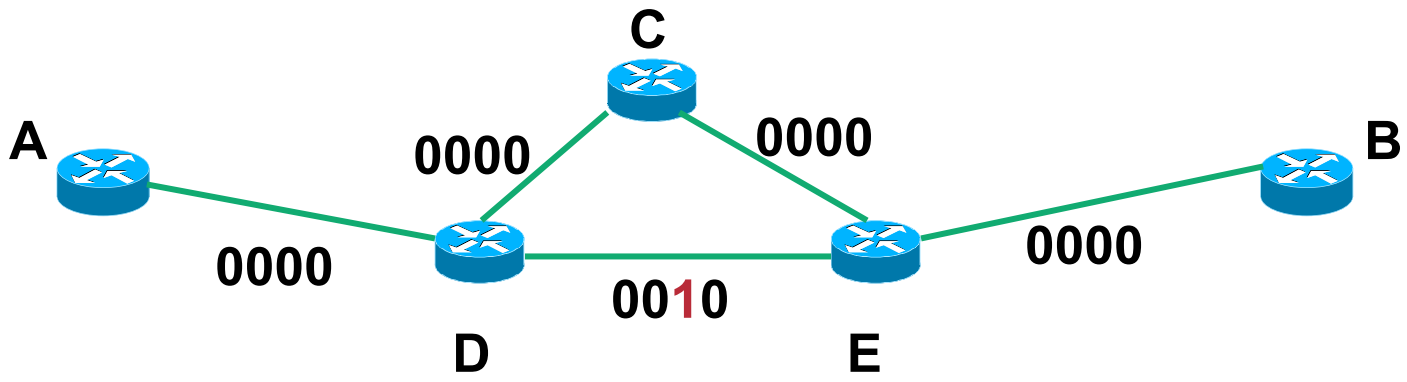
- **Trunk A to B:**

  tunnel = 0000, t-mask = 0011

- **ADEB and ADCEB are possible**

© 2006 Cisco Systems, Inc. All rights reserved.

Example1a: 4-bit string

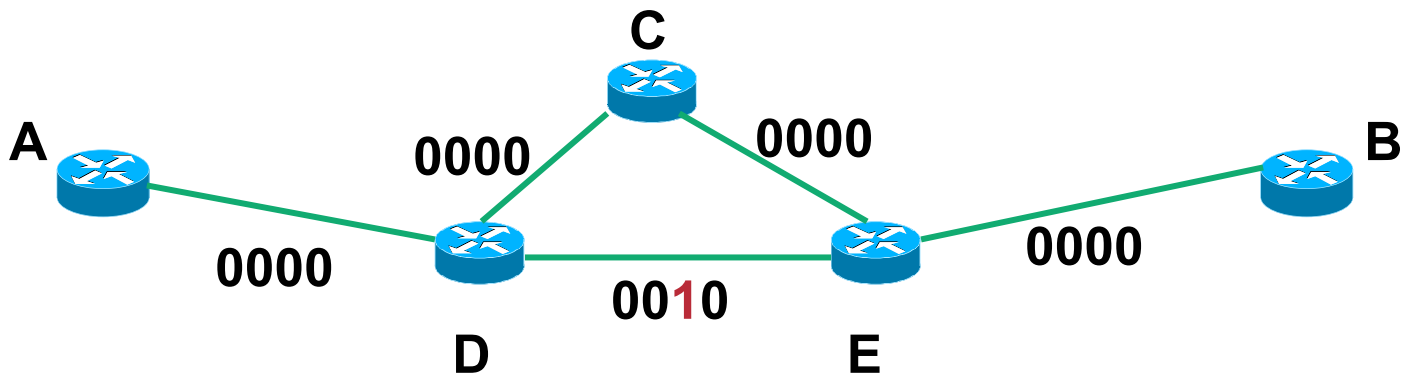


- Setting a link bit in the lower half drives all tunnels off the link, except those specially configured
- Trunk A to B:

  tunnel = 0000, t-mask = 0011
- Only ADCEB is possible

© 2006 Cisco Systems, Inc. All rights reserved.

Example1b: 4-bit string

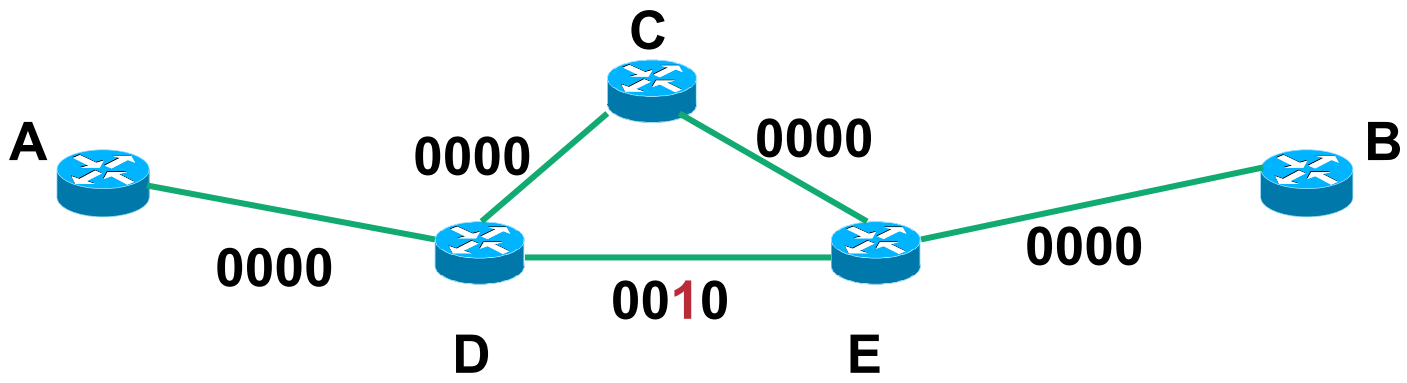


- A specific tunnel can then be configured to allow such links by clearing the bit in its affinity attribute mask
- Trunk A to B:

  tunnel = 0000, t-mask = 0001
- Again, ADEB and ADCEB are possible

© 2006 Cisco Systems, Inc. All rights reserved.

Example1c: 4-bit string

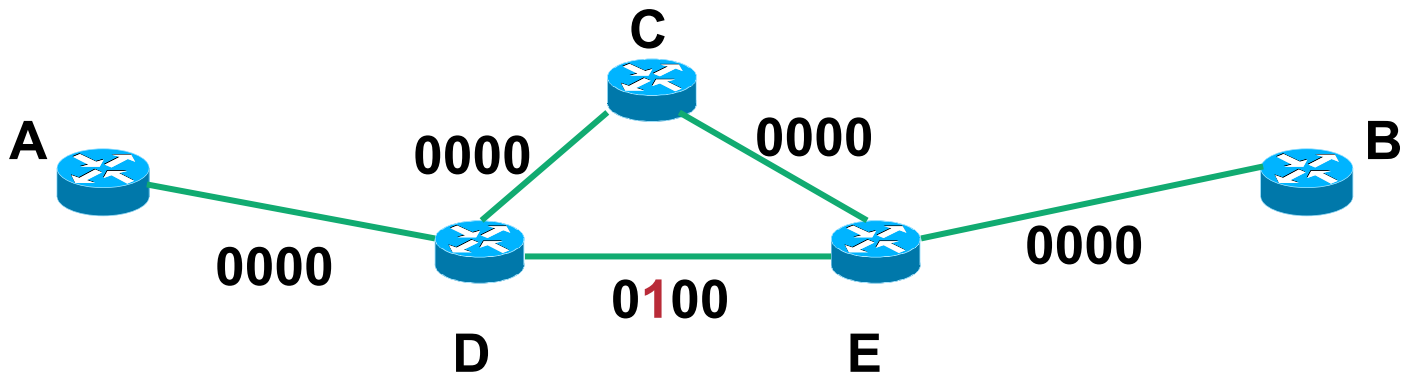


- A specific tunnel can be restricted to only such links by instead turning on the bit in its affinity attribute bits
- Trunk A to B:

  tunnel = 0010, t-mask = 0011
- No path is possible

© 2006 Cisco Systems, Inc. All rights reserved.

Example2a: 4-bit string

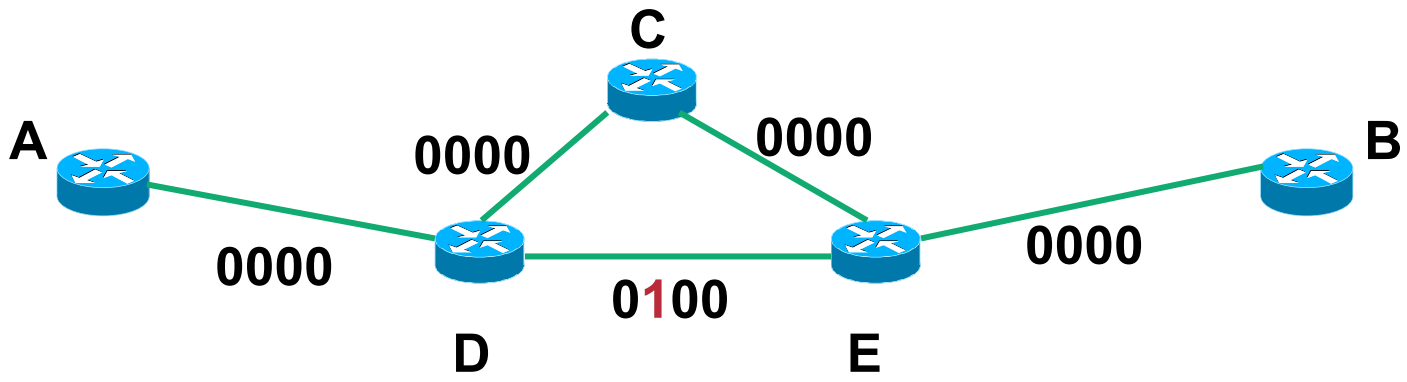


- Setting a link bit in the upper half drives has no immediate effect
- Trunk A to B:

  tunnel = 0000, t-mask = 0011
- ADEB and ADCEB are both possible

© 2006 Cisco Systems, Inc. All rights reserved.

Example2b: 4-bit string

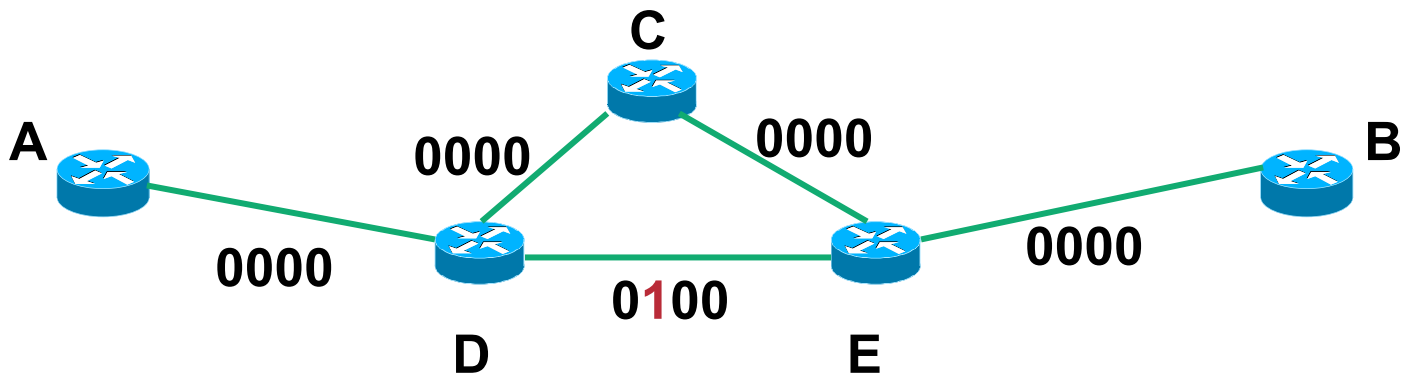


- A specific tunnel can be driven off the link by setting the bit in its mask
- Trunk A to B:

  tunnel = 0000, t-mask = 0111
- Only ADCEB is possible

© 2006 Cisco Systems, Inc. All rights reserved.

Example2c: 4-bit string



- A specific tunnel can be restricted to only such links
- Trunk A to B:

  tunnel = 0100, t-mask = 0111
- No path is possible

© 2006 Cisco Systems, Inc. All rights reserved.

Tunnel Path Selection

- **Tunnel has two path options**
  1. **Dynamic**
  2. **Explicit**
- **Path is a set of next-hop addresses (physical or loopbacks) to destination**
- **This set of next-hops is called Explicit Route Address (ERO)**

© 2006 Cisco Systems, Inc. All rights reserved.

Dynamic Path Option

```
tunnel mpls traffic-eng path-option <prio>
dynamic
```

- dynamic = router calculates path using TE topology database
- Router will take best IGP path that meets BW requirements
- If BW=0, tunnel could take the IGP path

© 2006 Cisco Systems, Inc. All rights reserved.

Explicit Path Option

```
tunnel mpls traffic-eng path-option
<prio> explicit <id|name> [ID|NAME]>
```

- **explicit = take specified path**
- **Strict source-routing of IP traffic**

© 2006 Cisco Systems, Inc. All rights reserved.

Explicit Path Option (Cont.)

```
ip explicit-path <id|name> [ID|NAME]
    next-address 192.168.1.1
    next-address 192.168.2.1 {loose}
    ...
```

- explicit = take specified path
- Router sets up path you specify
- Strict source-routing of IP traffic
- Each hop is a physical interface or loop back

© 2006 Cisco Systems, Inc. All rights reserved.

How does ERO come into play?

- If dynamic path-option is used, TE topology database is used to COMPUTE the Explicit Path
- If explicit path-option is used, TE topology database is used to VERIFY the Explicit Path

© 2006 Cisco Systems, Inc. All rights reserved.

MPLS-TE: Link attributes, IGP enhancements, CSPF

© 2005 Cisco Systems, Inc. All rights reserved.

Agenda

- **Link Attributes**
- **Information flooding**
- **IGP Enhancements for MPLS-TE**
- **Path Computation (C-SPF)**

© 2006 Cisco Systems, Inc. All rights reserved.

Link Attributes

- **Link attributes**
  - **Bandwidth per priority (0-7)**
  - **Link Affinity**
  - **TE-specific link metric**

© 2006 Cisco Systems, Inc. All rights reserved.

Bandwidth

```
ip rsvp bandwidth <x> <y>
```

- Per-physical-interface command
- X = amount of reservable BW, in K
- Y = not used by MPLS-TE

© 2006 Cisco Systems, Inc. All rights reserved.

Link Affinity

```
mpls traffic-eng attribute-flags
<0x0-0xFFFFFFFF>
```

- **Per-physical-interface command**

© 2006 Cisco Systems, Inc. All rights reserved.

Administrative Weight

```
mpls traffic-eng administrative-
weight <X>
```

- Per-physical-interface command
- X = 0–4,294,967,295
- Gives a metric that be considered for use instead of the IGP metric
- This can be used as a per-tunnel delay-sensitive metric for doing VoIP TE
- By default TE metric is used. However, when no TE metric is configured,

  IGP metric => TE metric

© 2006 Cisco Systems, Inc. All rights reserved.

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

© 2006 Cisco Systems, Inc. All rights reserved.

Information Distribution

- A link-state protocol has to be used as the IGP (IS-IS or OSPF)
- A Link-state protocol is not a requirement for other MPLS applications (e.g. VPNs)

© 2006 Cisco Systems, Inc. All rights reserved.

Need for a Link-State Protocol

Why is a link-state protocol required?

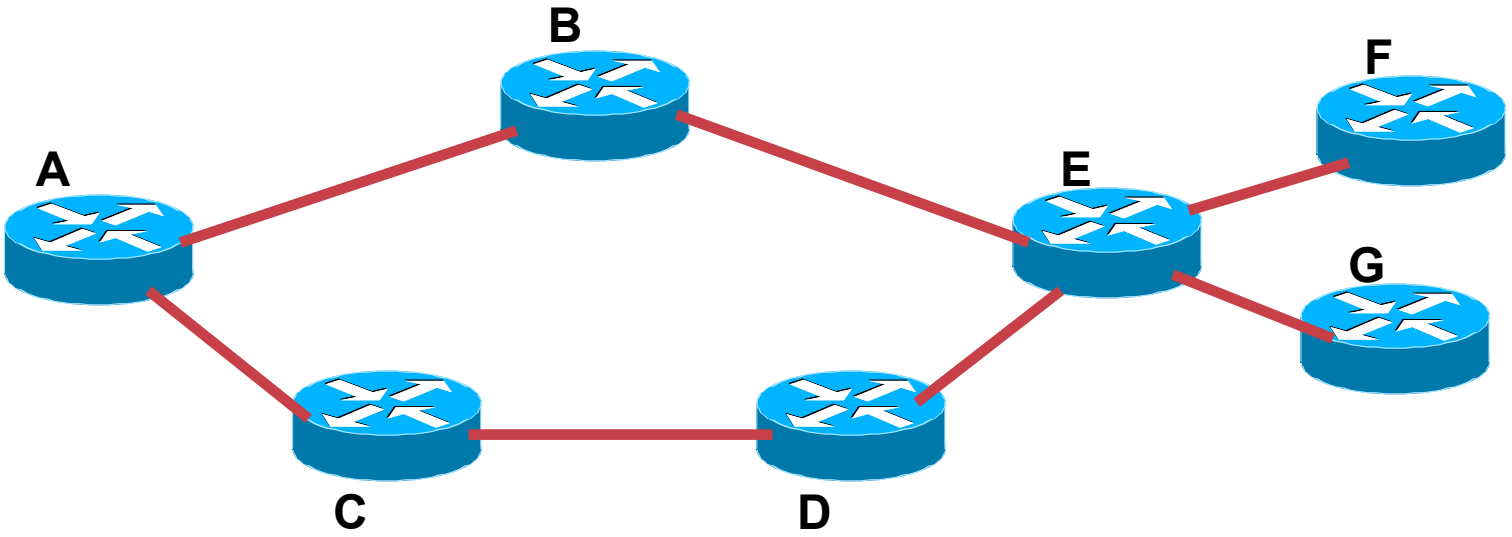
- Path is computed at the source
- Source needs entire picture (topology) of the network to make routing decision
- Only link-state protocols flood link information to build a complete network topology

© 2006 Cisco Systems, Inc. All rights reserved.

Need for a Link-State Protocol

Consider the following network:

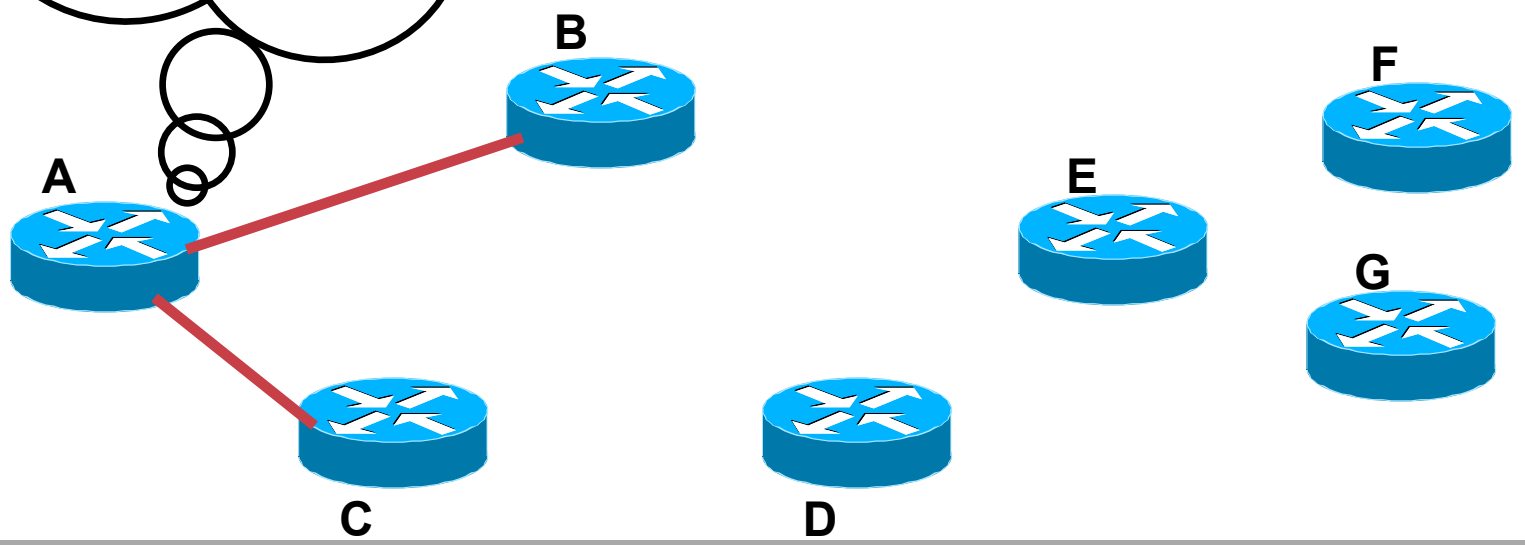
- **All links have a cost of 10**
- **Path from "A" to "E" is A->B->E, cost 20**
- **All traffic from "A" to {E,F,G} goes A->B->E**



© 2006 Cisco Systems, Inc. All rights reserved.

What a Distance Vector Protocol Sees

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

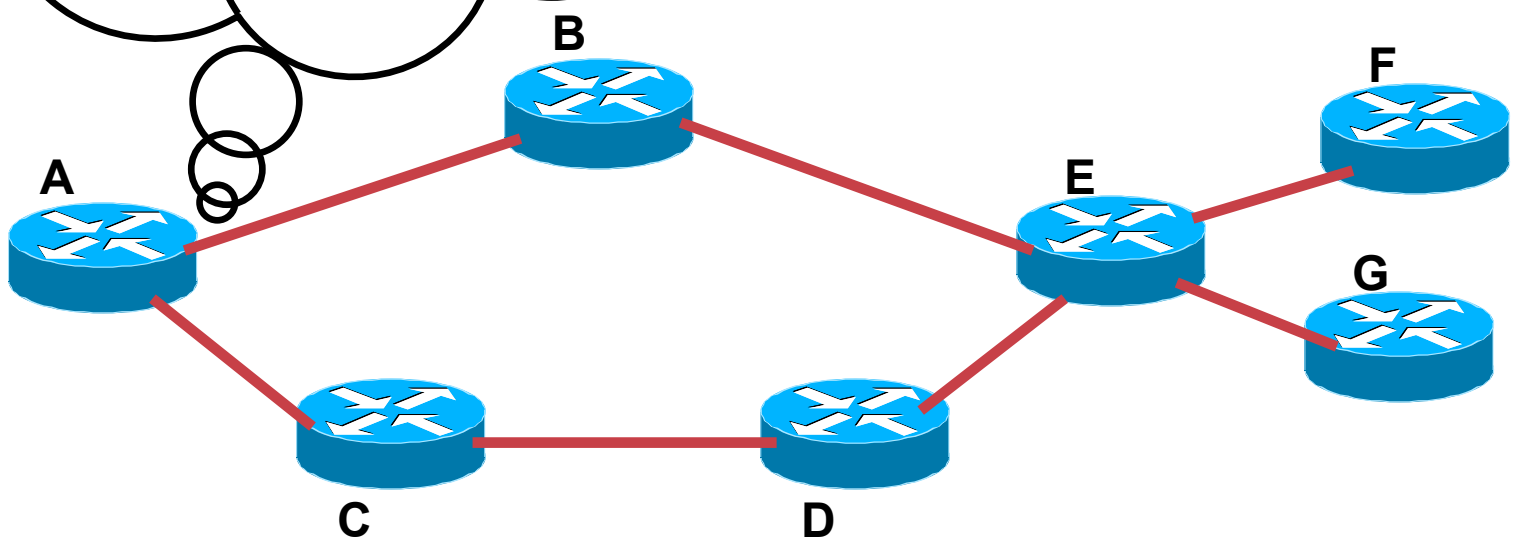


- "A" doesn't see all the links
- "A" *knows* about the shortest path
- Protocol limitation by design

© 2006 Cisco Systems, Inc. All rights reserved.

What a Link-State Protocol Sees

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



- "A" sees all links
- "A" *computes* the shortest path
- Routing table doesn't change

© 2006 Cisco Systems, Inc. All rights reserved.

Link-State Protocol Extensions/ IGP Flooding

- TE finds paths other than shortest-cost
- To do this, TE must have more info than just per-link cost
- OSPF and IS-IS have been extended to carry additional information
  - Physical bandwidth
  - RSVP configured bandwidth
  - RSVP Available bandwidth
  - Link TE metric
  - Link affinity

© 2006 Cisco Systems, Inc. All rights reserved.

OSPF Extensions

- **OSPF**
  - **Uses Type 10 (Opaque Area-Local) LSAs**
  - **See draft-katz-yeung-ospf-traffic**

© 2006 Cisco Systems, Inc. All rights reserved.

IS-IS Extensions

- **IS-IS**
  - **Uses Type 22 TLVs**
  - **See draft-ietf-isis-traffic**

© 2006 Cisco Systems, Inc. All rights reserved.

ISIS Extensions (contd)

- **Extended IS neighbor subTLVs**
  - **subTLV #3 - administrative group (color)**
  - **subTLV #6 - IPv4 interface address**
  - **subTLV #8 – IPv4 neighbor address**
  - **subTLV #9 - maximum link bandwidth**
  - **subTLV#10 - maximum reservable link BW**
  - **subTLV #11 - current bandwidth reservation**
  - **subTLV #18 - default TE metric**

© 2006 Cisco Systems, Inc. All rights reserved.

Information Distribution

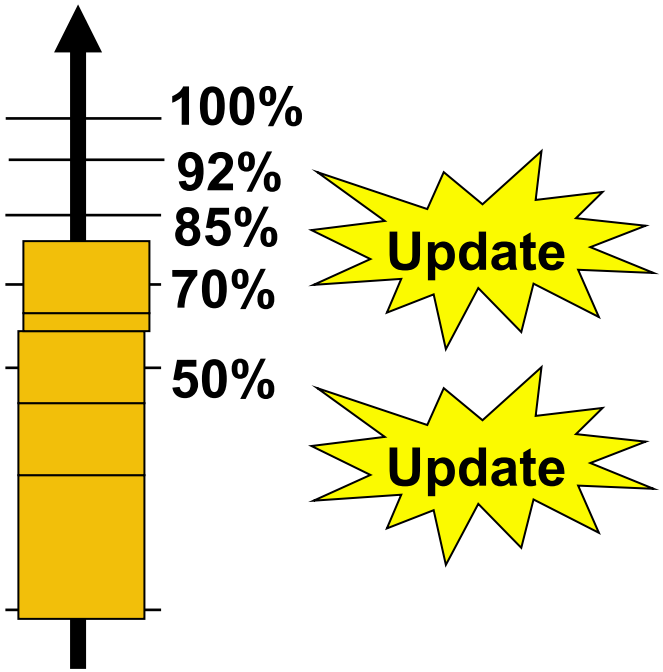
- Dynamics of ISIS and OSPF are unchanged

  Periodic flooding

  Hold-down timer to constrain the frequency of advertisements
- Current constraint information sent when IGP decides to re-flood
- TE admission control requests re-flooding on significant changes
  - *significant* is determined by a configurable set of thresholds
  - On link configuration changes
  - On link state changes
  - On LSP Setup failure
  - TE refresh timer expires (180 seconds default)

© 2006 Cisco Systems, Inc. All rights reserved.

Significant Change



- Each time a threshold is crossed, an update is sent
- Denser population as utilization increases
- Different thresholds for UP and Down

```
router#sh mpls traffic-eng link bandwidth-allocation pos4/0
 .............<snip>.....................
     Up Thresholds:         15 30 45 60 75 80 85 90 95 96 97 98 99 100 (default)
   Down Thresholds:       100 99 98 97 96 95 90 85 80 75 60 45 30 15 (default)
.............<snip>.....................
```

© 2006 Cisco Systems, Inc. All rights reserved.

Per-Priority Available BW

T=0	D advertises: AB(0)=100=…= AB(7)=100 AB(i) = 'Available Bandwidth at priority I"
T=1	Setup of a tunnel over L at priority=3 for 30 units
T=2	D advertises: AB(0)=AB(1)=AB(2)=100 AB(3)=AB(4)=…=AB(7)=70
T=3	Setup of an additional tunnel over L at priority=5 for 30 units
T=4	D advertises: AB(0)=AB(1)=AB(2)=100 AB(3)=AB(4)=70 AB(5)=AB(6)=AB(7)=40

This means that another tunnel having the piority < 3 and Bw > 70M would preempt the previous installed tunnel

© 2006 Cisco Systems, Inc. All rights reserved.

Constrained-based Path Computation (C-SPF)

© 1999, Cisco Systems, Inc.
© 2003 Cisco Systems, Inc. All rights reserved.

Path Calculation

- Modified Dijkstra at tunnel head-end
- Often referred to as CSPF

  Constrained SPF
- …or PCALC (path calculation)
- Final result is explicit route meeting desired constrain

© 2006 Cisco Systems, Inc. All rights reserved.

Path Calculation (C-SPF)

- **Shortest-cost path is found that meets administrative constraints**
- **These constraints can be**
  - **bandwidth**
  - **link attribute (aka color, resource group)**
  - **priority**
- **The addition of constraints is what allows MPLS-TE to use paths other than *just* the shortest one**

© 2006 Cisco Systems, Inc. All rights reserved.

Path Computation

“On demand” by the trunk’s head-end:

- **for a new trunk**
- **for an existing trunk whose (current) LSP failed**
- **for an existing trunk when doing re-optimization**

© 2006 Cisco Systems, Inc. All rights reserved.

Path Computation

Input:

- configured attributes of traffic trunks originated at this router
- attributes associated with resources
  - available from IS-IS or OSPF
- topology state information
  - available from IS-IS or OSPF

© 2006 Cisco Systems, Inc. All rights reserved.

Path Computation

- **Prune links if:**

  insufficient resources (e.g., bandwidth)

  violates policy constraints

- **Compute shortest distance path**

  TE uses its own metric

- **Tie-break:**

  1. **Path with the highest available bandwidth**
  2. **Path with the smallest hop-count**
  3. **Path found first in TE topology database**

© 2006 Cisco Systems, Inc. All rights reserved.

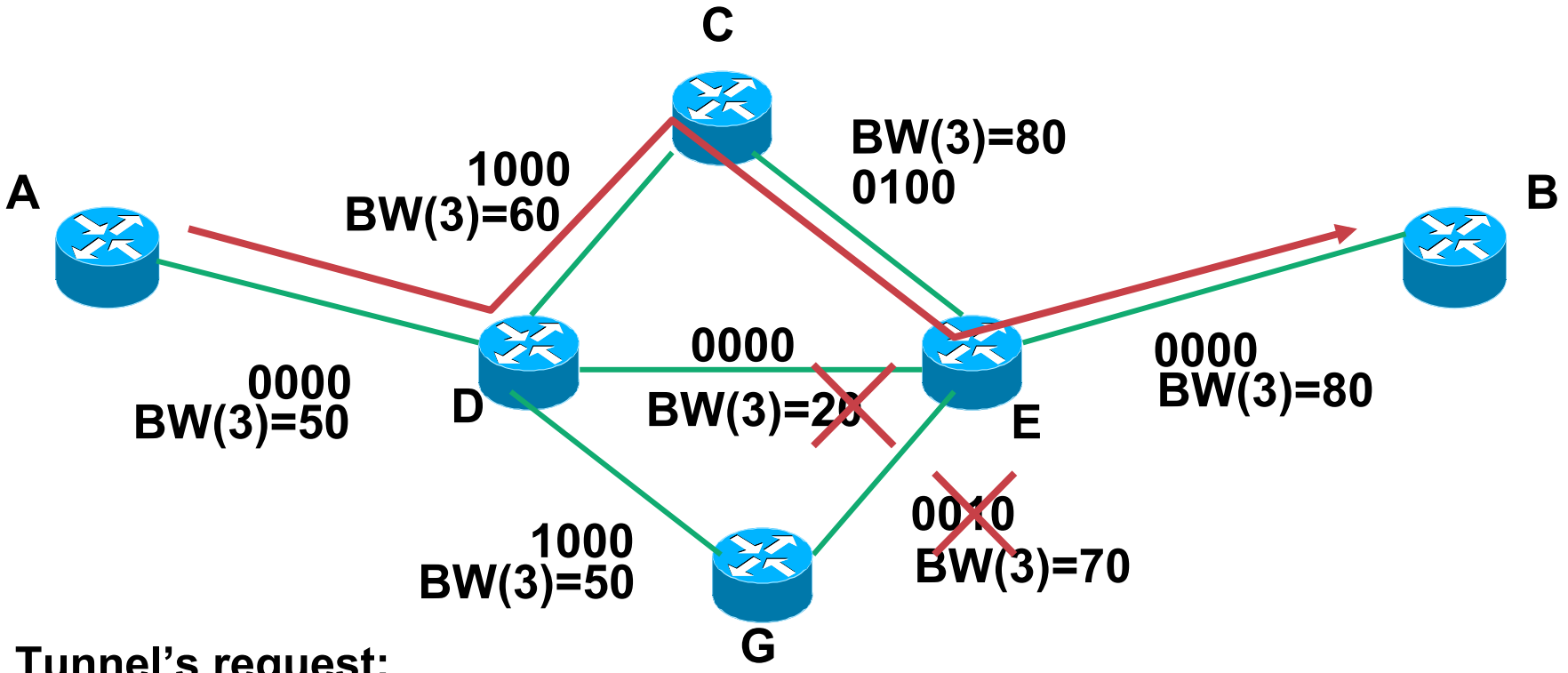
Path Computation

Output:

- **explicit route - expressed as a sequence of router IP addresses**
  - **interface addresses for numbered links**
  - **loopback address for unnumbered links**
- **used as an input to the path setup component**

© 2006 Cisco Systems, Inc. All rights reserved.

BW/Policy Example



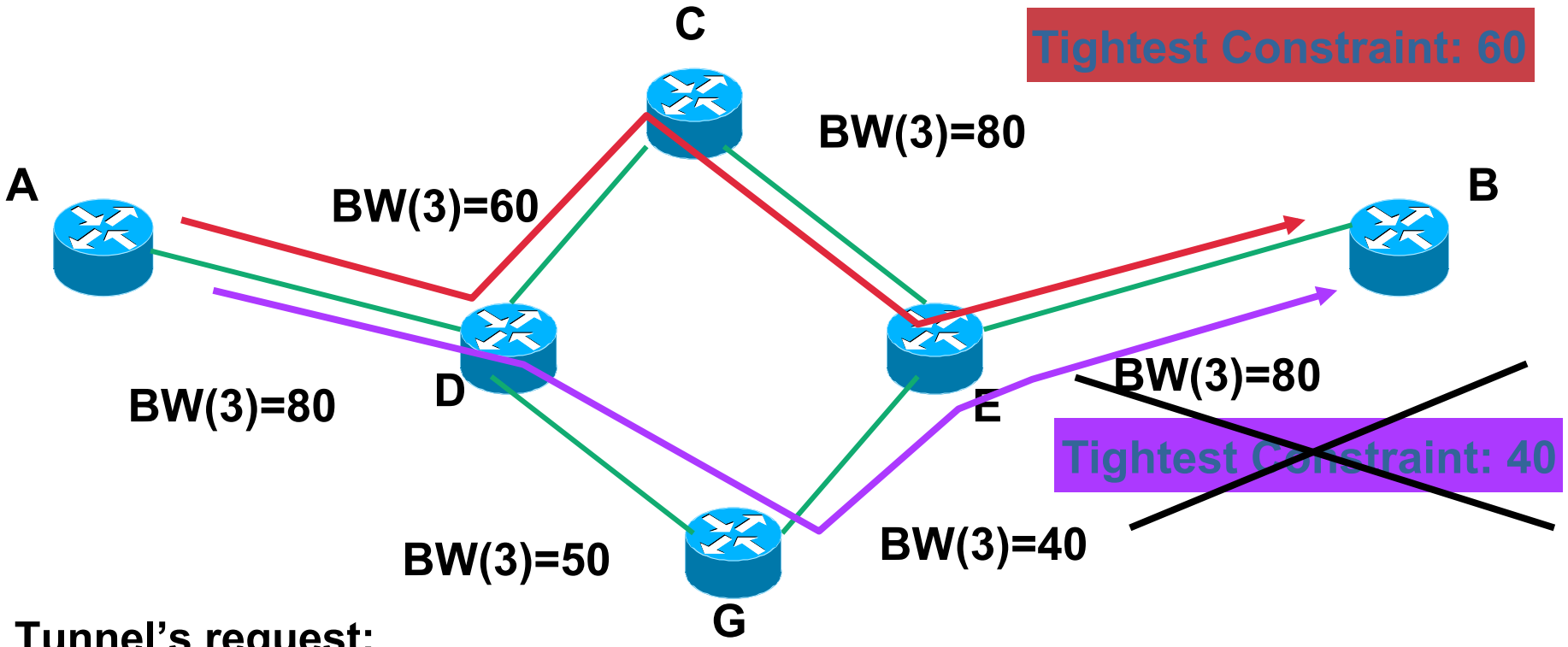
- **Tunnel's request:**

  Priority 3, BW = 30 units,

  Policy string: 0000, mask: 0011

© 2006 Cisco Systems, Inc. All rights reserved.

Maximizing the Tightest Constraint



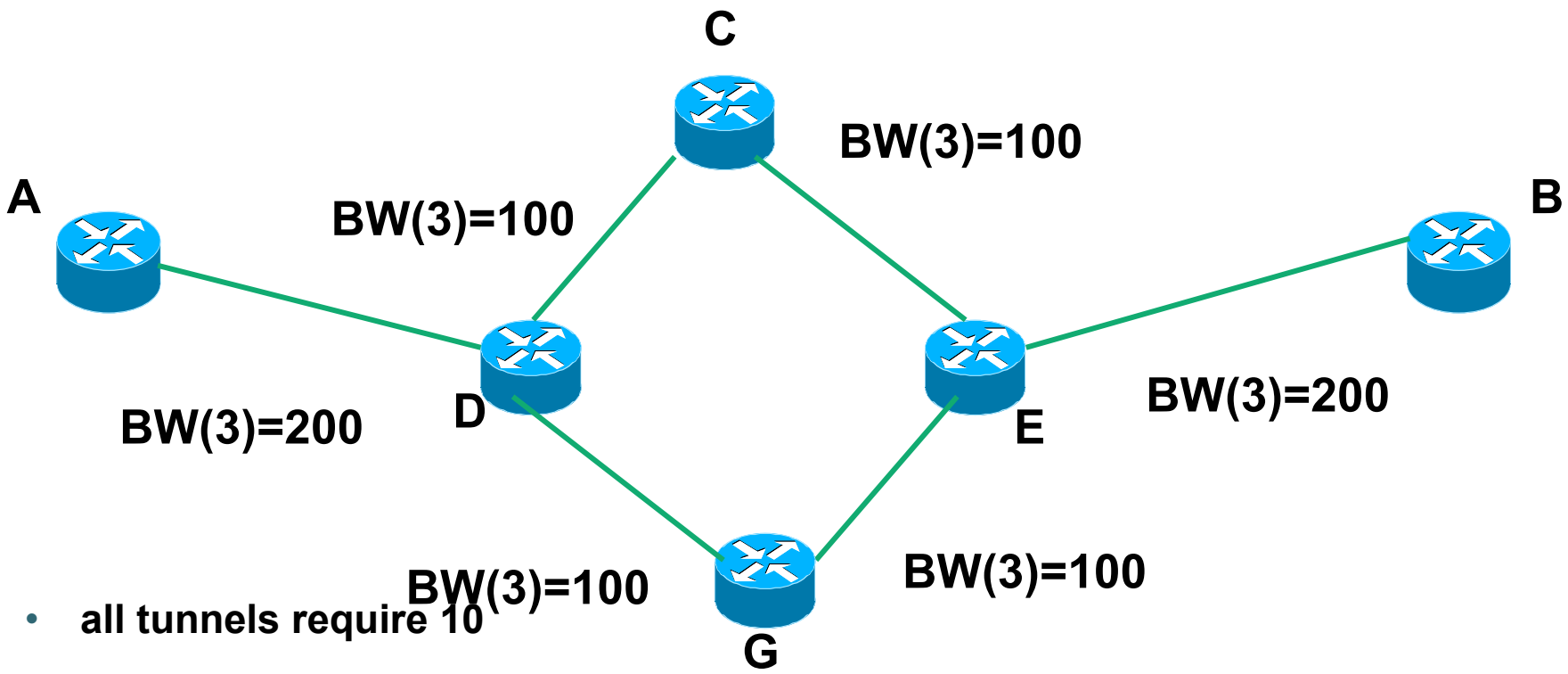
- **Tunnel's request:**

  Priority 3, BW = 30 units,

  Policy string: 0000, mask: 0011

© 2006 Cisco Systems, Inc. All rights reserved.

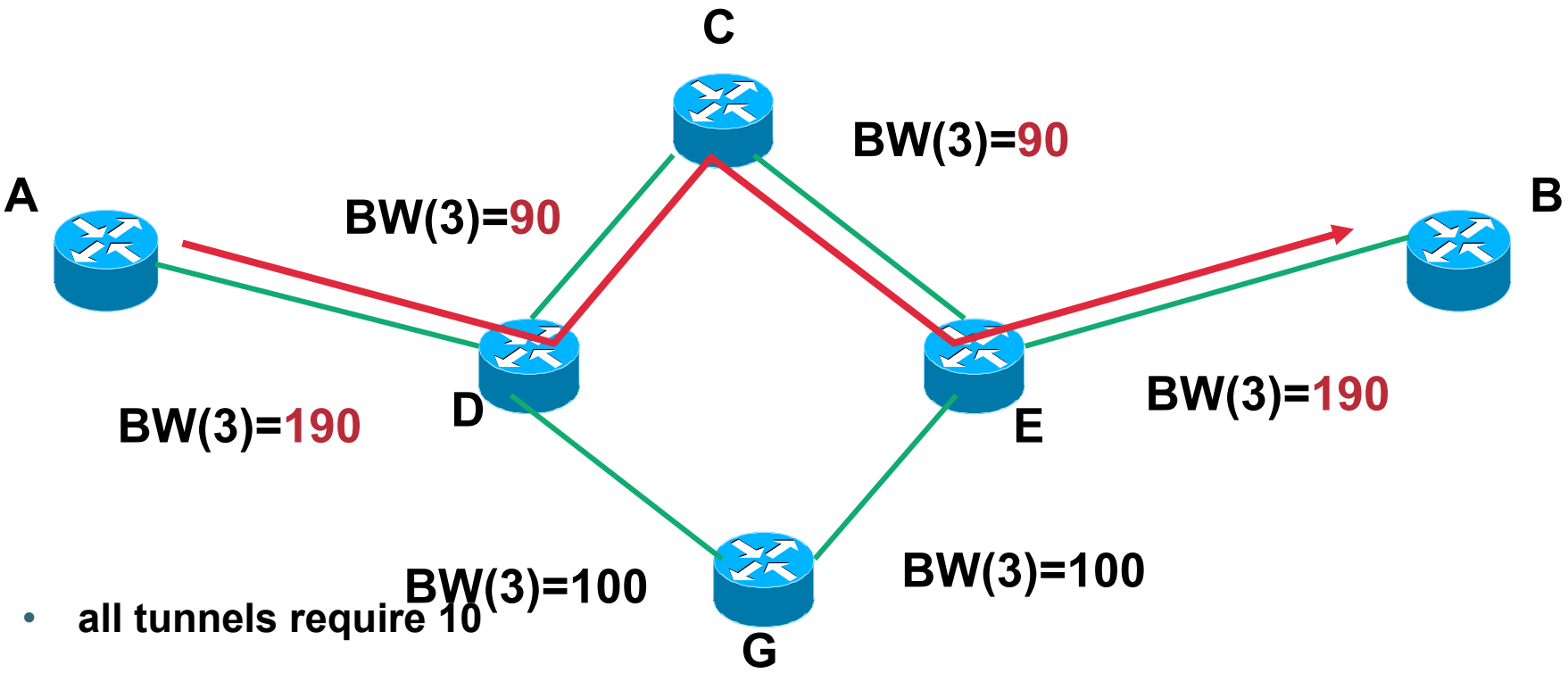
Load-Balancing tunnels



- all tunnels require 10

© 2006 Cisco Systems, Inc. All rights reserved.

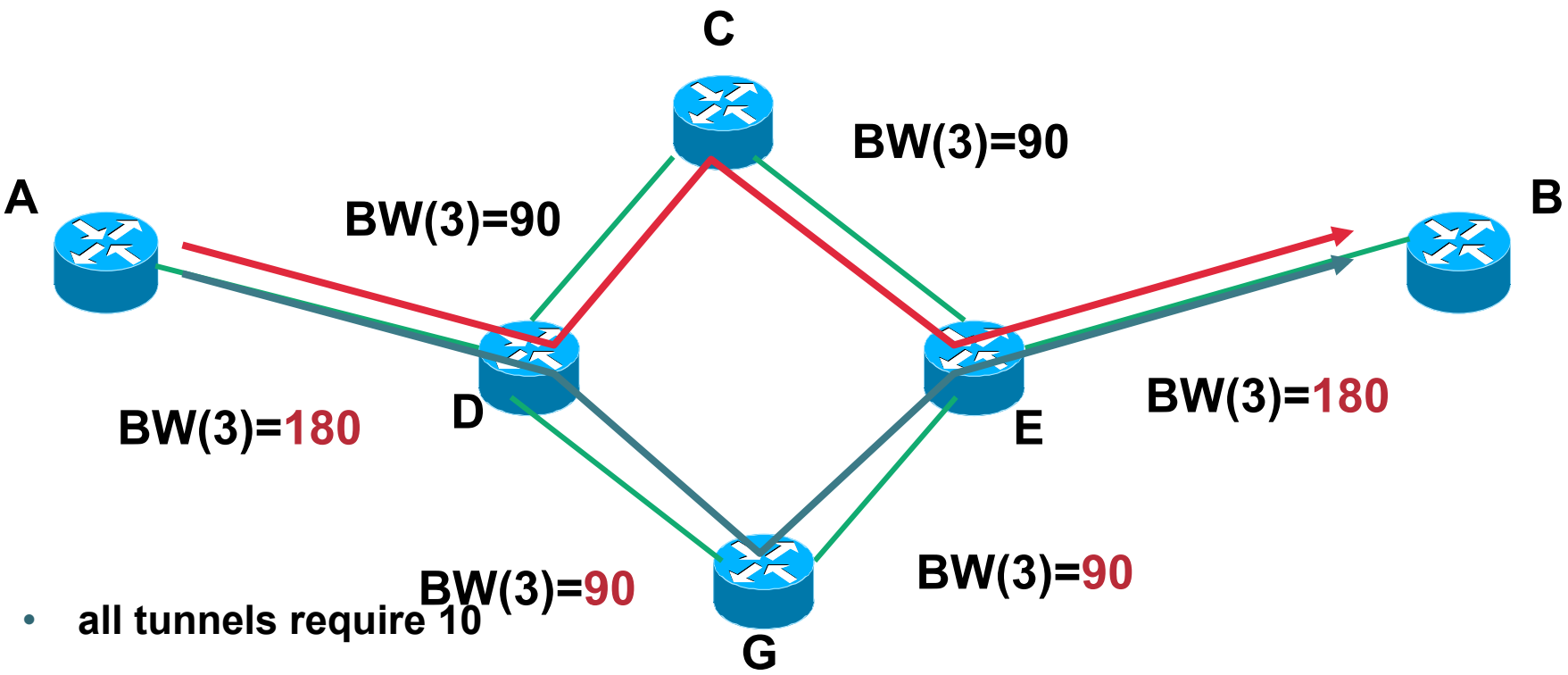
Load-Balancing tunnels



- all tunnels require 10

© 2006 Cisco Systems, Inc. All rights reserved.

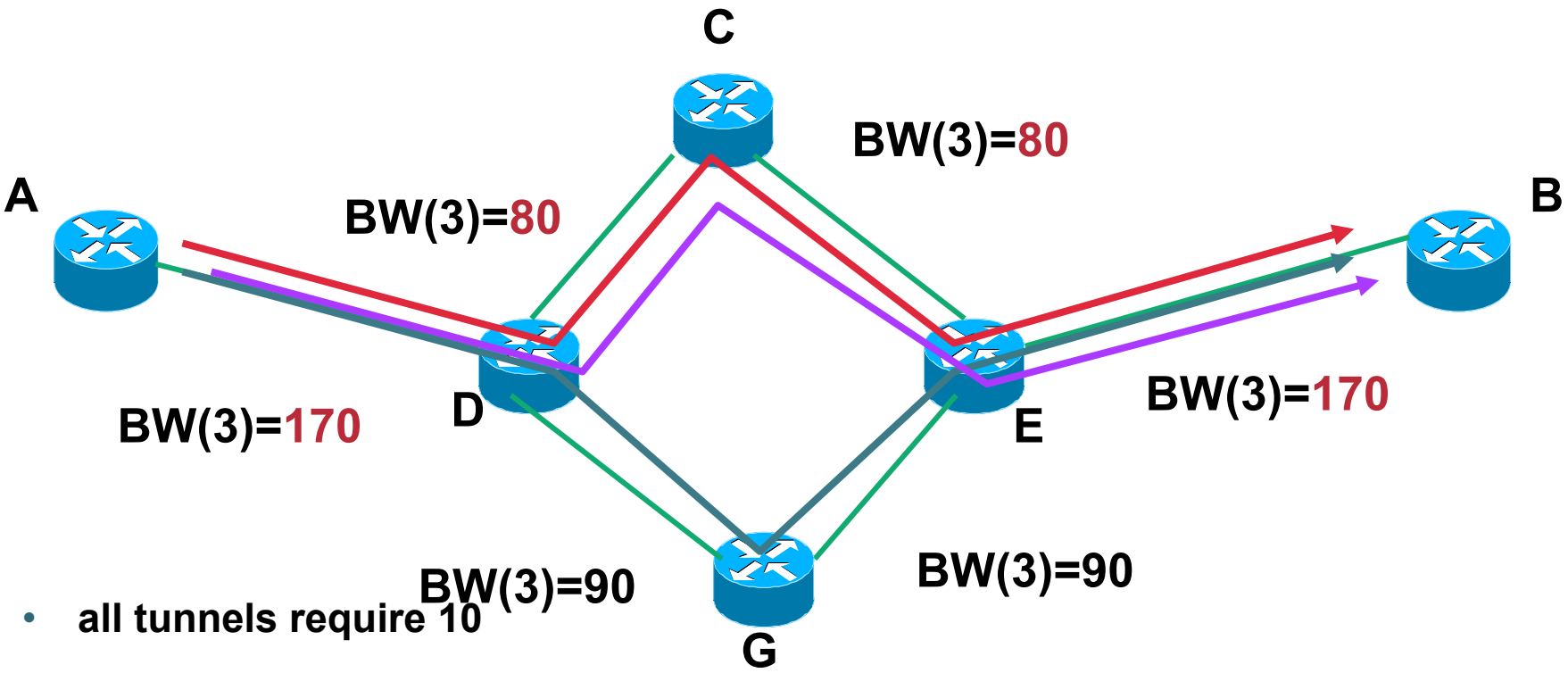
Load-Balancing tunnels



- all tunnels require 10

© 2006 Cisco Systems, Inc. All rights reserved.

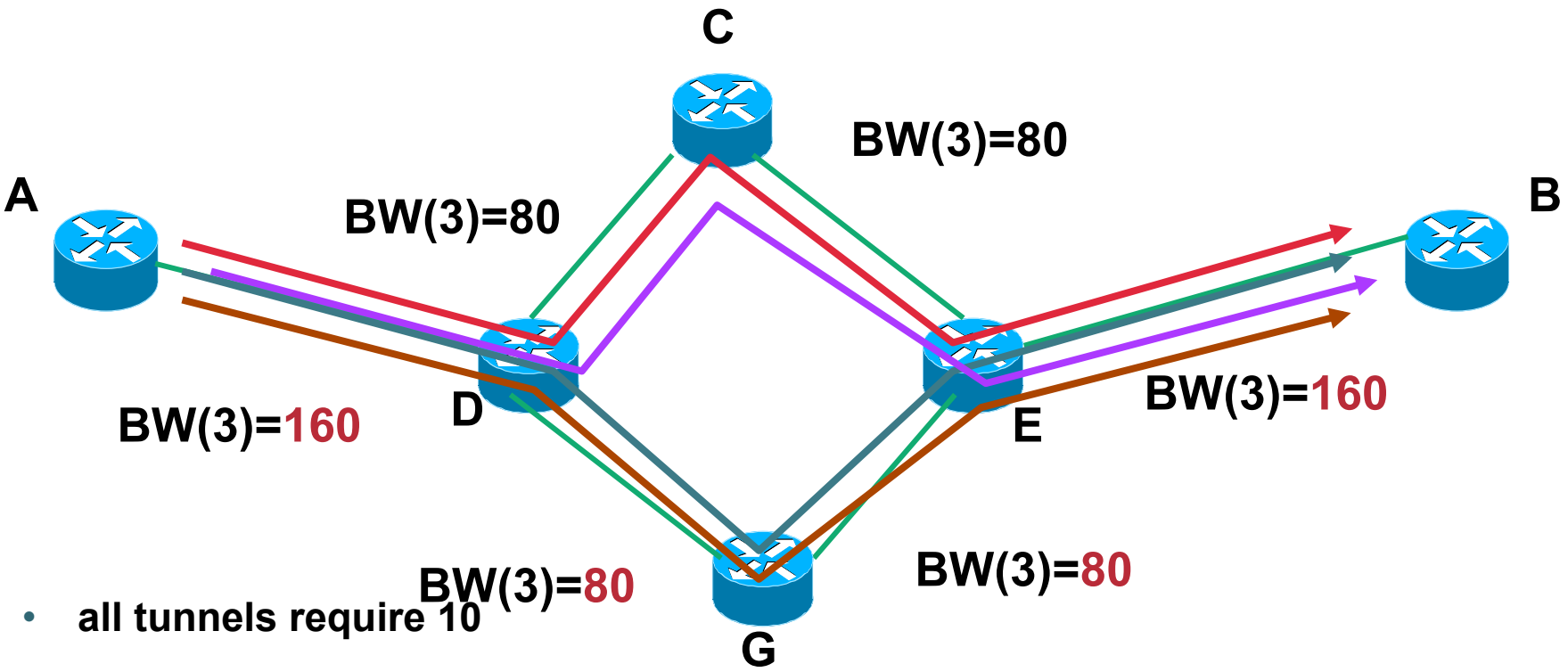
Load-Balancing tunnels



- all tunnels require 10

© 2006 Cisco Systems, Inc. All rights reserved.

Load-Balancing tunnels



- all tunnels require 10

© 2006 Cisco Systems, Inc. All rights reserved.

MPLS-TE: RSVP extensions, tunnel signaling and tunnel maintenance

© 2005 Cisco Systems, Inc. All rights reserved.

Agenda

- Path Setup (RSVP Extensions)
- Path maintenance
- Reoptimization
- Mapping Traffic to Tunnels
- Using metrics with tunnels
- Load balancing with TE tunnels

© 2006 Cisco Systems, Inc. All rights reserved.

Path Setup (RSVP Extensions)

© 1999, Cisco Systems, Inc.
© 2003 Cisco Systems, Inc. All rights reserved.

Path Setup

- After we calculate a path, we need to build an LSP across that path
- Path setup is done at the head-end of a trunk with RSVP + TE extensions
- RSVP sends PATH messages out, gets RESV messages back
- RFC2205, plus RFC 3209

© 2006 Cisco Systems, Inc. All rights reserved.

RSVP Extensions to RFC2205 for LSP Tunnels

- Downstream-on-demand label distribution
- Instantiation of explicit label switched paths
- Allocation of network resources (e.g., Bandwidth) to explicit lsps
- Re-routing of established lsp-tunnels in a smooth fashion using the concept of make-before-break
- Tracking of the actual route traversed by an lsp-tunnel
- Diagnostics on lsp-tunnels
- Pre-emption options that are administratively controllable

© 2006 Cisco Systems, Inc. All rights reserved.

RSVP Extensions for TE

	PATH	RESV
LABEL_REQUEST	X	
LABEL		X
EXPLICIT_ROUTE	X	
RECORD_ROUTE	X	X
SESSION_ATTRIBUTE	X	

© 2006 Cisco Systems, Inc. All rights reserved.

RSVP Label Allocation

- Labels are distributed from down-stream to up-stream
- Label Binding via PATH message - LABEL_REQUEST object
- Labels are allocated & distributed via RESV message using LABEL Object.

© 2006 Cisco Systems, Inc. All rights reserved.

RSVP - ERO

- ERO - Explicit Route Object
- “PATH” message carries ERO (concatenation of hops which constitute explicitly routed path) given by the Head-End Router
- This is used in setting up for the LSP
- The path can be administratively specified or dynamically computed

© 2006 Cisco Systems, Inc. All rights reserved.

RSVP - Record Route

- Added to the PATH message by the head-end Router.
- Every Router along the path records its IP address in the RRO.
- Used by the Head-End Router on how the actual LSP has traversed.
- Used for Loop Detection

© 2006 Cisco Systems, Inc. All rights reserved.

RSVP - Session Attribute

- Added to “PATH” message by Head-End router to aid in session identification & diagnostics
  - setup priority
  - hold priorities
  - resource affinities

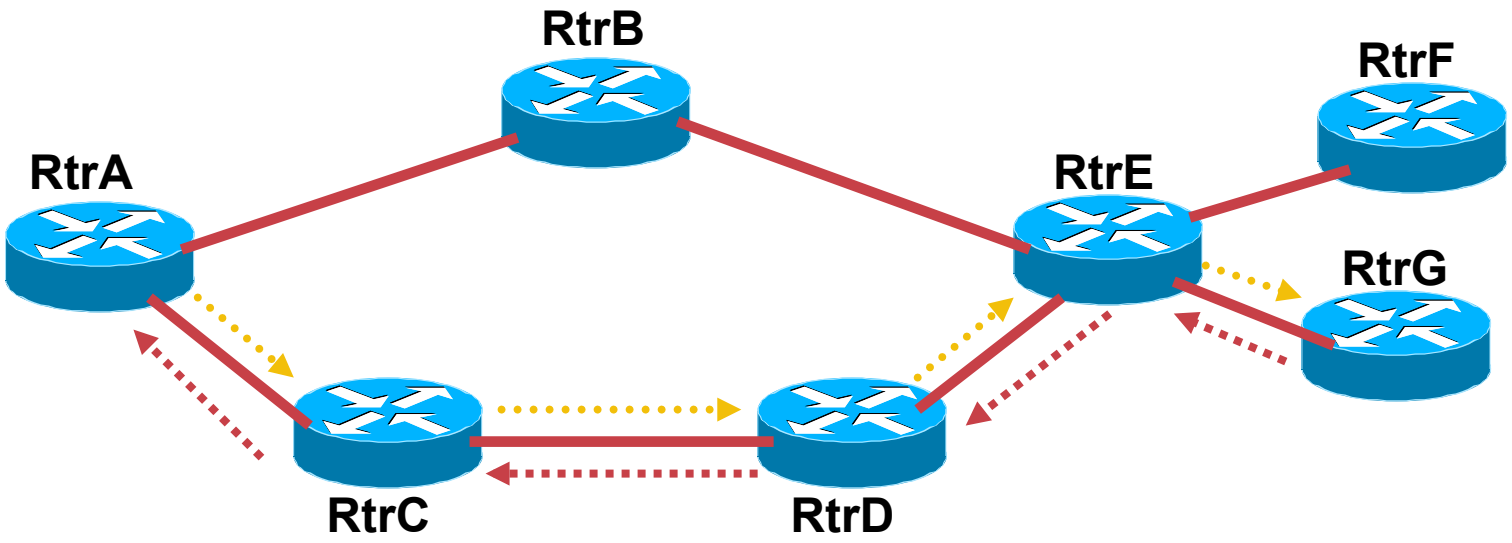
© 2006 Cisco Systems, Inc. All rights reserved.

Path Setup

- 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 messages are refreshed every 30 seconds

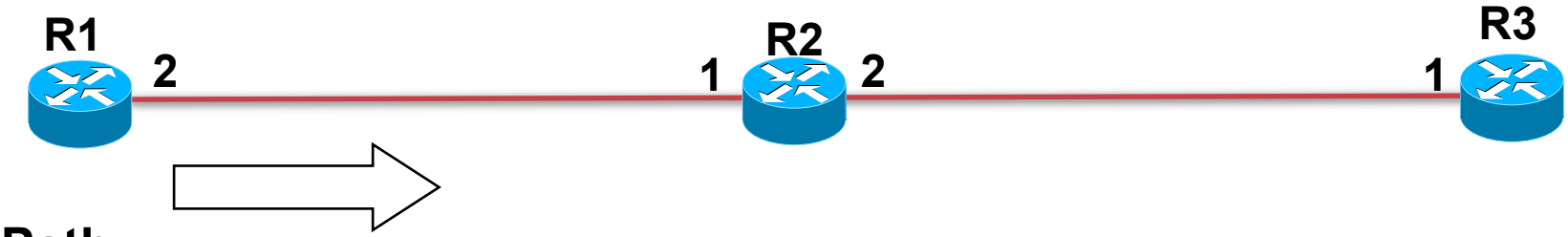
= PATH messages

= RESV messages



© 2006 Cisco Systems, Inc. All rights reserved.

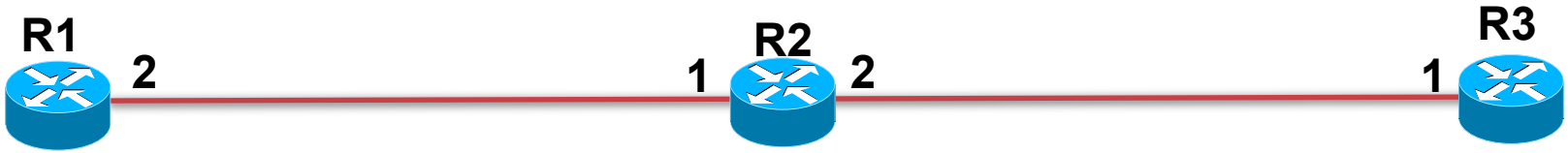
Path Setup - more details



Path:
Common_Header
Session(R3-lo0, 0, R1-lo0)
PHOP(R1-2)
Label_Request(IP)
ERO (R2-1, R3-1)
Session_Attribute (S(3), H(3), 0x04)
Sender_Template(R1-lo0, 00)
Sender_Tspec(2Mbps)
Record_Route(R1-2)

© 2006 Cisco Systems, Inc. All rights reserved.

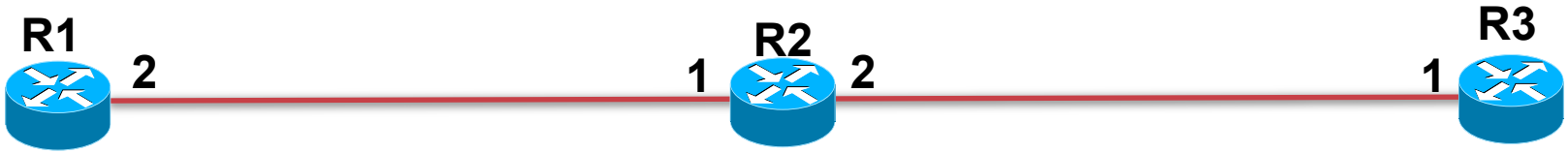
Path Setup - more details



Path State:
Session(R3-lo0, 0, R1-lo0)
PHOP(R1-2)
Label_Request(IP)
ERO (R2-1, R3-1)
Session_Attribute (S(3), H(3), 0x04)
Sender_Template(R1-lo0, 00)
Sender_Tspec(2Mbps)
Record_Route (R1-2)

© 2006 Cisco Systems, Inc. All rights reserved.

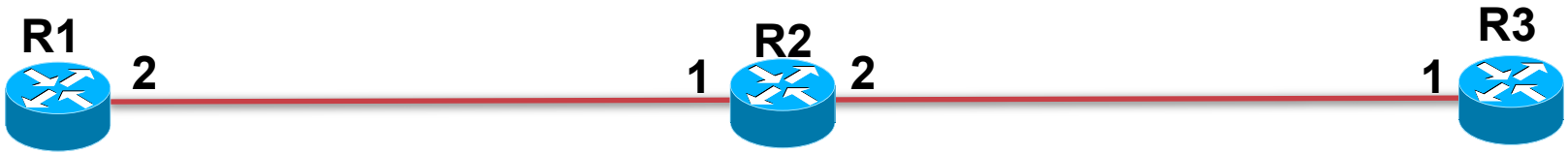
Path Setup - more details



Path:
Common_Header
Session(R3-lo0, 0, R1-lo0)
PHOP(R2-2)
Label_Request(IP)
ERO (R3-1)
Session_Attribute (S(3), H(3), 0x04)
Sender_Template(R1-lo0, 00)
Sender_Tspec(2Mbps)
Record_Route (R1-2, R2-2)

© 2006 Cisco Systems, Inc. All rights reserved.

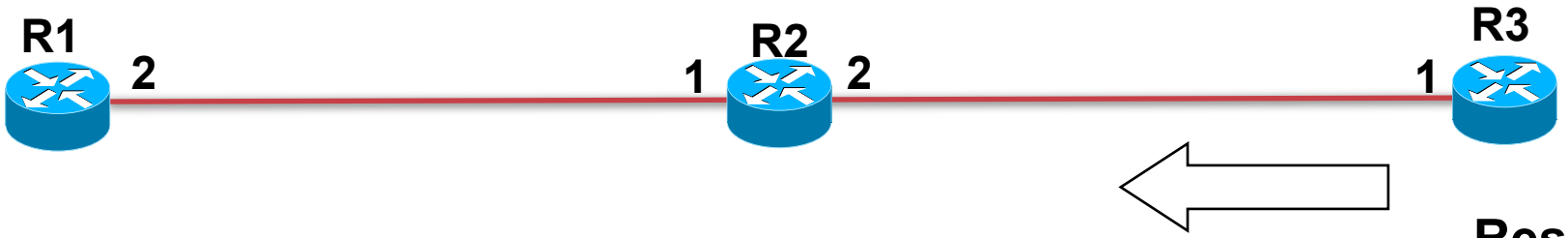
Path Setup - more details



Path State:
Session(R3-lo0, 0, R1-lo0)
PHOP(R2-2)
Label_Request(IP)
ERO ()
Session_Attribute (S(3), H(3), 0x04)
Sender_Template(R1-lo0, 00)
Sender_Tspec(2Mbps)
Record_Route (R1-2, R2-2, R3-1)

© 2006 Cisco Systems, Inc. All rights reserved.

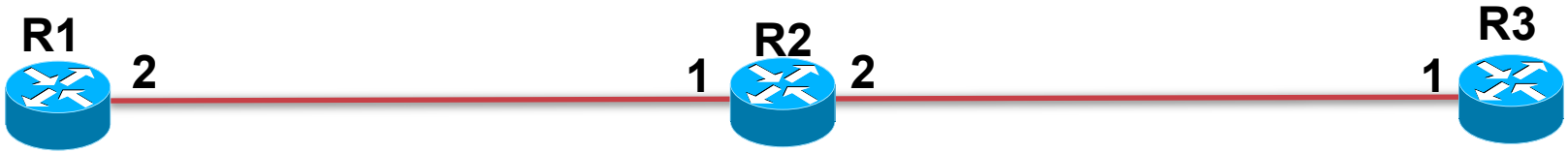
Path Setup - more details



Resv:
Common_Header
Session(R3-lo0, 0, R1-lo0)
PHOP(R3-1)
Style=SE
FlowSpec(2Mbps)
Sender_Template(R1-lo0, 00)
Label=POP
Record_Route(R3-1)

© 2006 Cisco Systems, Inc. All rights reserved.

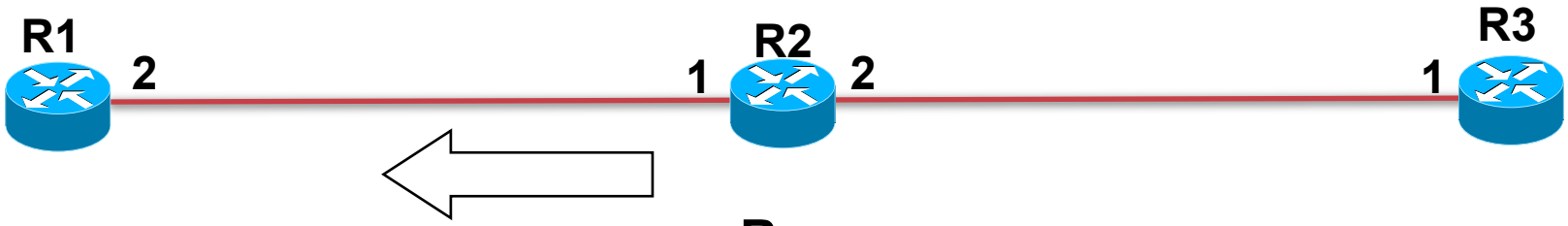
Path Setup - more details



Resv State
Session(R3-lo0, 0, R1-lo0)
PHOP(R3-1)
Style=SE
FlowSpec (2Mbps)
Sender_Template(R1-lo0, 00)
OutLabel=POP
IntLabel=5
Record_Route(R3-1)

© 2006 Cisco Systems, Inc. All rights reserved.

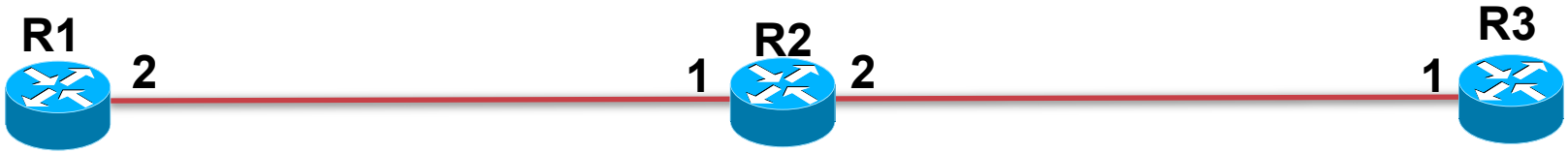
Path Setup - more details



Resv:
Common_Header
Session(R3-lo0, 0, R1-lo0)
PHOP(R2-1)
Style=SE
FlowSpec (2Mbps)
Sender_Template(R1-lo0, 00)
Label=5
Record_Route(R2-1, R3-1)

© 2006 Cisco Systems, Inc. All rights reserved.

Path Setup - more details



Resv state:
Session(R3-lo0, 0, R1-lo0)
PHOP(R2-1)
Style=SE
FlowSpec (2Mbps)
Sender_Template(R1-lo0, 00)
Label=5
Record_Route(R1-2, R2-1, R3-1)

© 2006 Cisco Systems, Inc. All rights reserved.

Trunk Admission Control

- Performed by routers along a Label Switched Path (LSP)
- Determines if resources are available
- May tear down (existing) LSPs with a lower priority
- Does the local accounting
- Triggers IGP information distribution when resource thresholds are crossed
- Since TE tunnels are unidirectional, we do admission control on outbound interfaces only

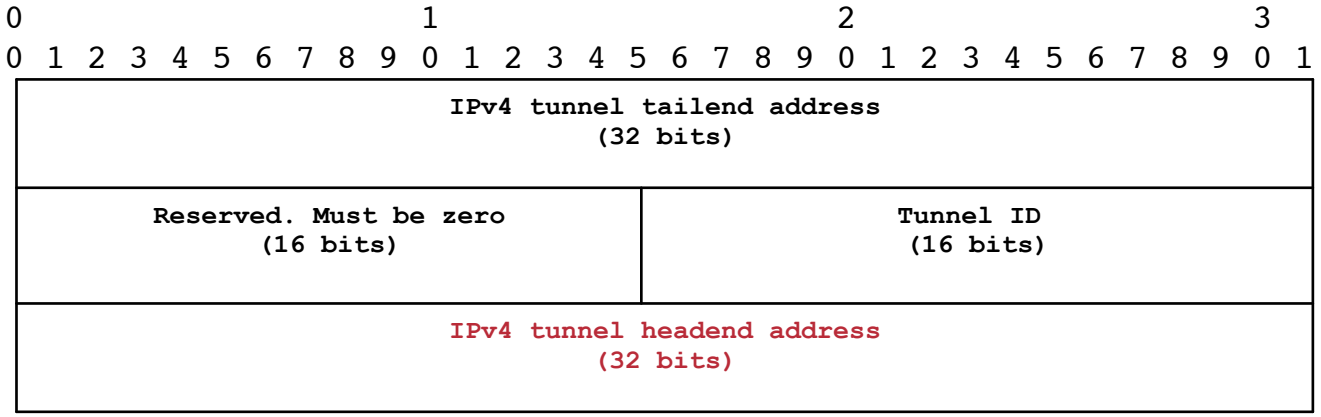
© 2006 Cisco Systems, Inc. All rights reserved.

Path maintenance

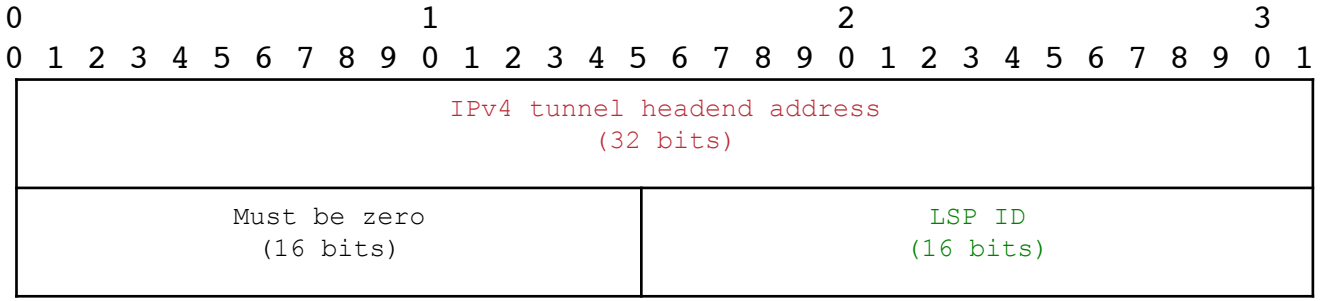
© 2005 Cisco Systems, Inc. All rights reserved.

Identifying TE-tunnels

SESSION Object



SENDER_TEMPLATE / FILTER_SPEC



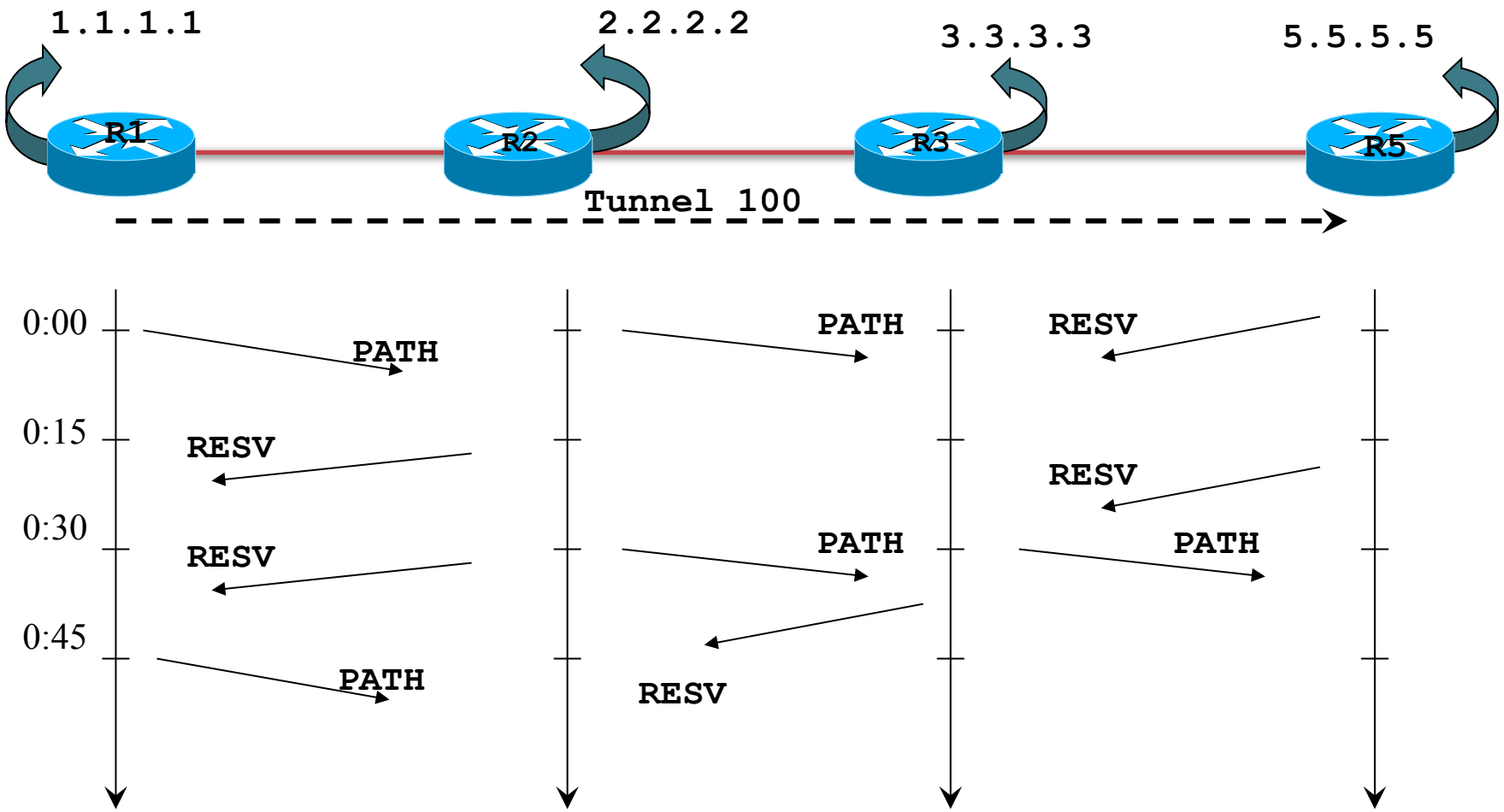
© 2006 Cisco Systems, Inc. All rights reserved.

Path Maintenance

- Once the TE tunnel is setup, PATH and RESV messages are used to maintain the tunnel state
- RSVP is a soft-state protocol, relying on PATH & RESV messages for state refresh
- PATH & RESV messages are sent out on average, every 30 seconds
- If we miss 4 consecutive PATH or RESV messages, we consider the RSVP reservation dead

© 2006 Cisco Systems, Inc. All rights reserved.

Path Maintenance in action



© 2006 Cisco Systems, Inc. All rights reserved.

Re-optimization

© 2005 Cisco Systems, Inc. All rights reserved.

Make-Before-Break objectives

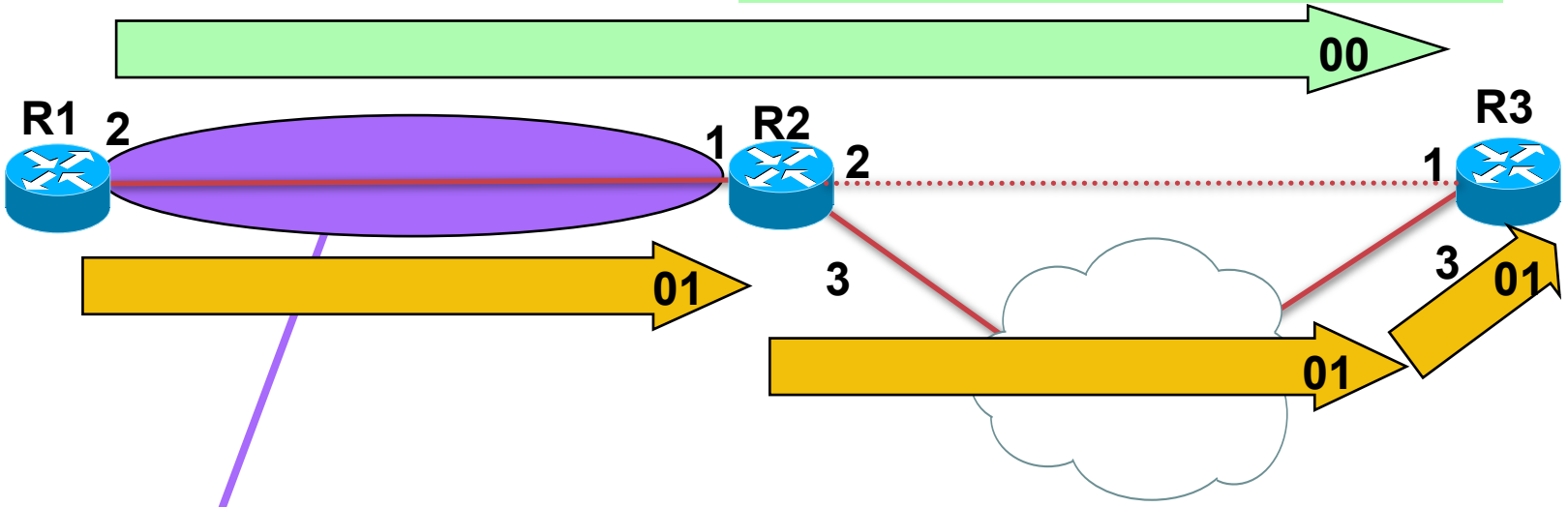
- Avoid tearing tunnel before the new tunnel instance comes up. This could cause traffic disruption
- Avoid double counting bandwidth on the common link carrying the new and the old tunnel

© 2006 Cisco Systems, Inc. All rights reserved.

Make before break in action

Session(R3-lo0, 0, R1-lo0)

ERO (R2-1, R3-1)
Sender_Template(R1-lo0, 00)

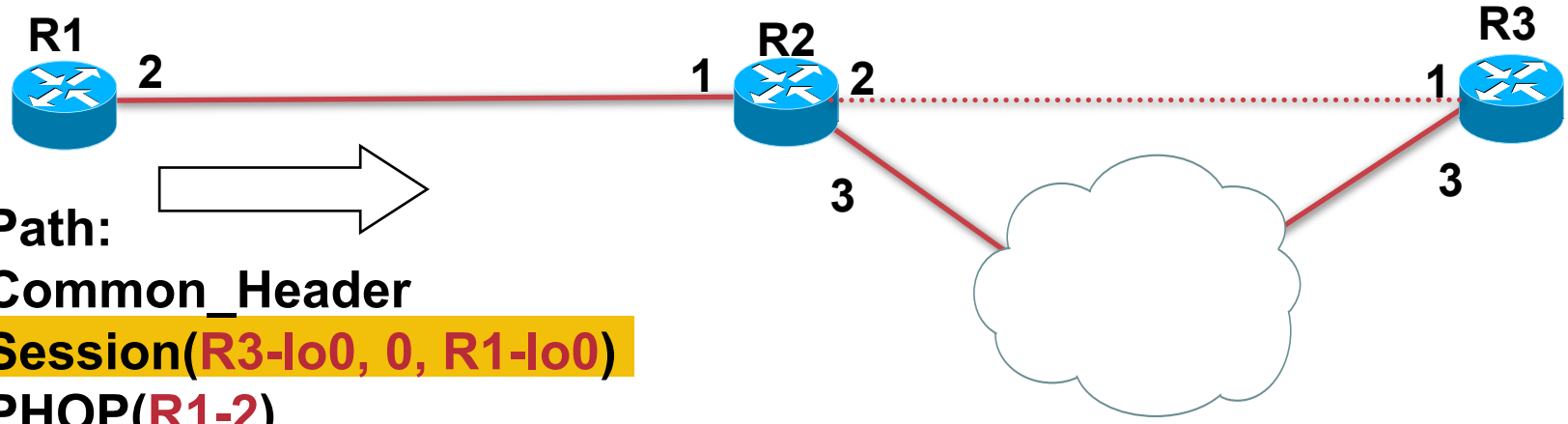


Resource Sharing

ERO (R2-1, …, R3-3)
Sender_Template(R1-lo0, 01)

© 2006 Cisco Systems, Inc. All rights reserved.

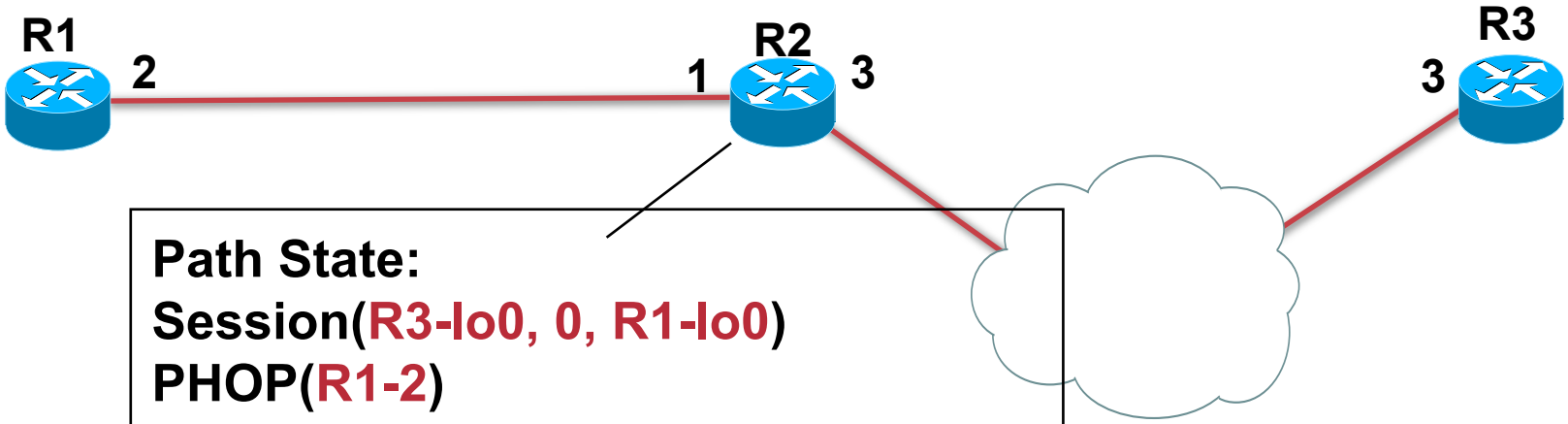
Make before break in action



Path:
Common_Header
Session(R3-lo0, 0, R1-lo0)
PHOP(R1-2)
Label_Request(IP)
ERO (R2-1, …,R3-3)
Session_Attribute (S(3), H(3), 0x04)
Sender_Template(R1-lo0, 01)
Sender_Tspec(3Mbps)
Record_Route(R1-2)

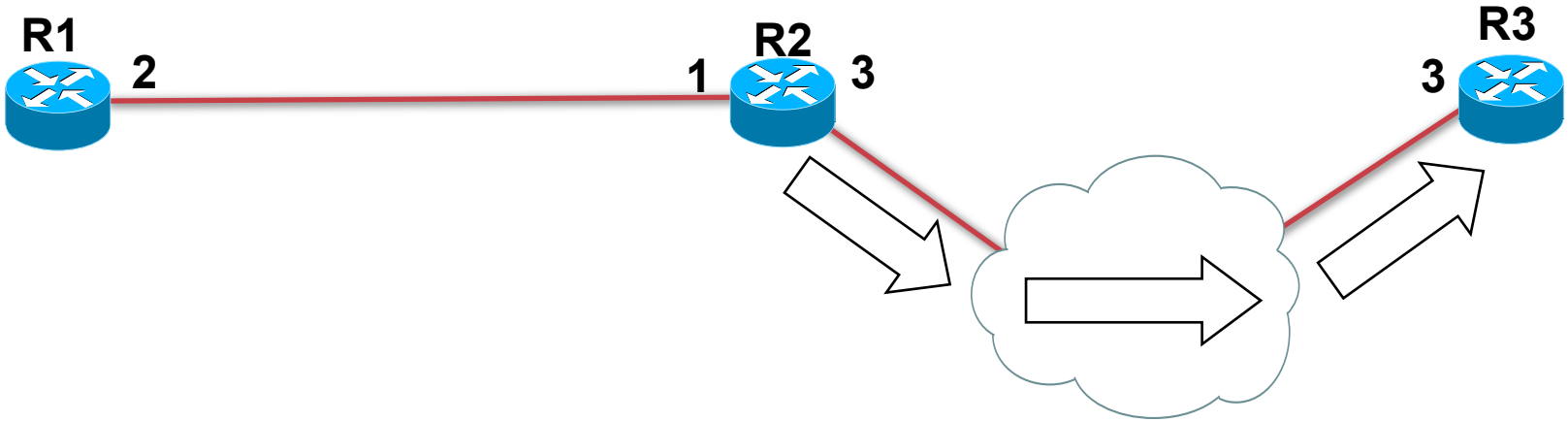
© 2006 Cisco Systems, Inc. All rights reserved.

Make before break in action



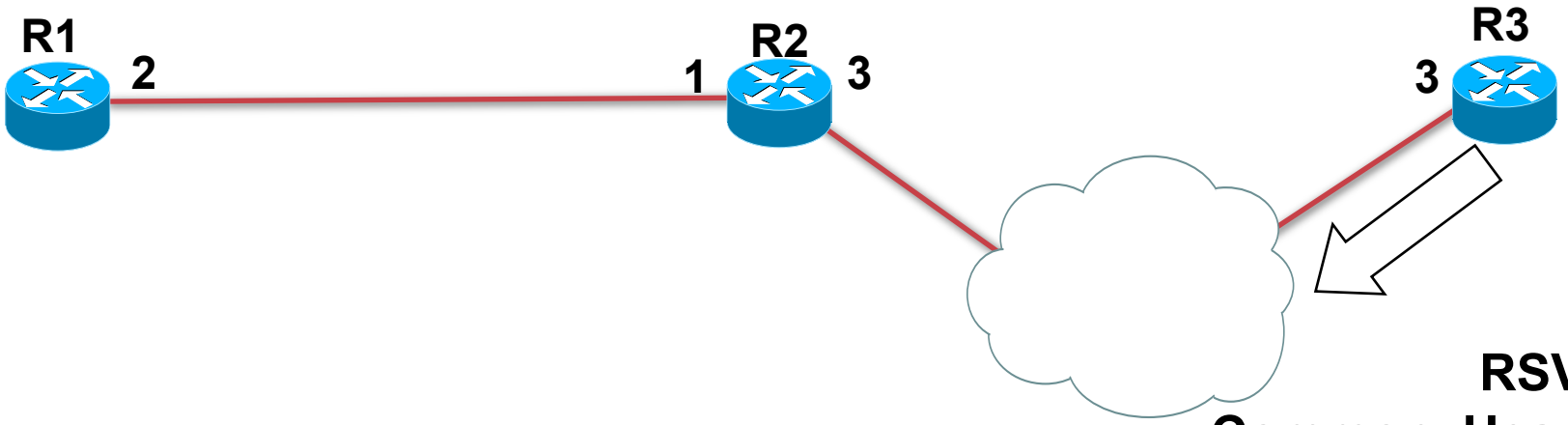
© 2006 Cisco Systems, Inc. All rights reserved.

Make before break in action



© 2006 Cisco Systems, Inc. All rights reserved.

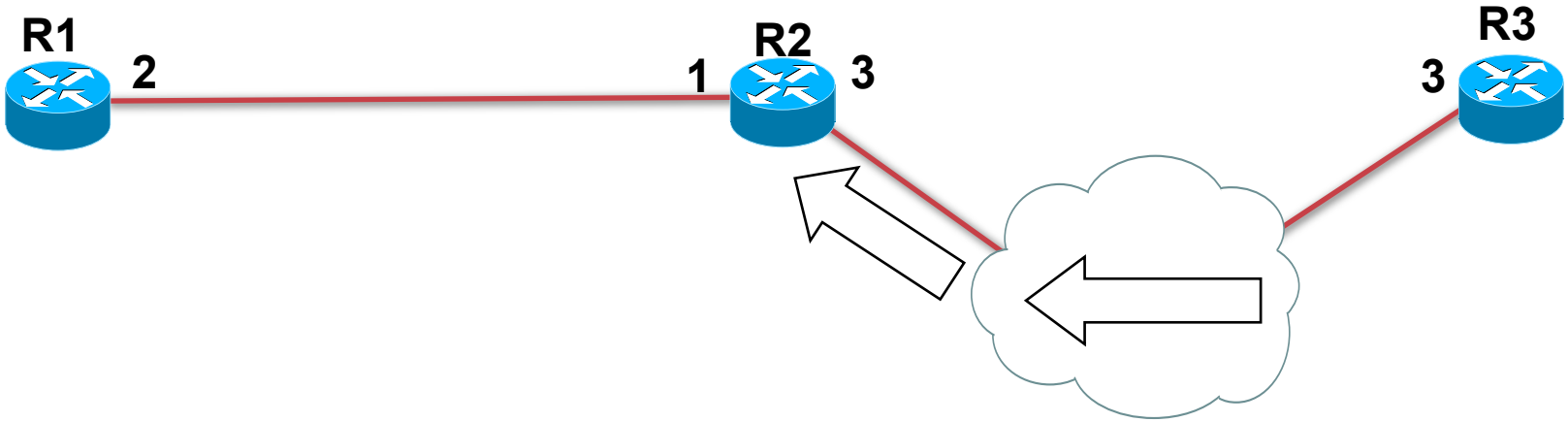
Make before break in action



RSVP:
Common_Header
Session(R3-lo0, 0, R1-lo0)
PHOP(R3-3)
Style=SE
FlowSpec(3Mbps)
Sender_Template(R1-lo0, 01)
Label=POP
Record_Route(R3-3)

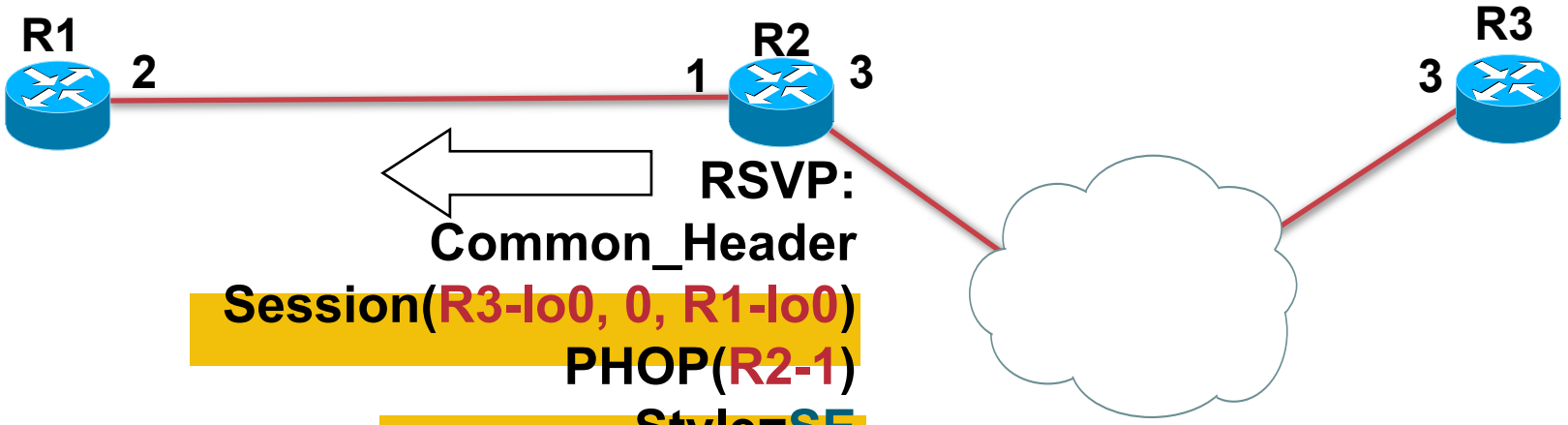
© 2006 Cisco Systems, Inc. All rights reserved.

Make before break in action



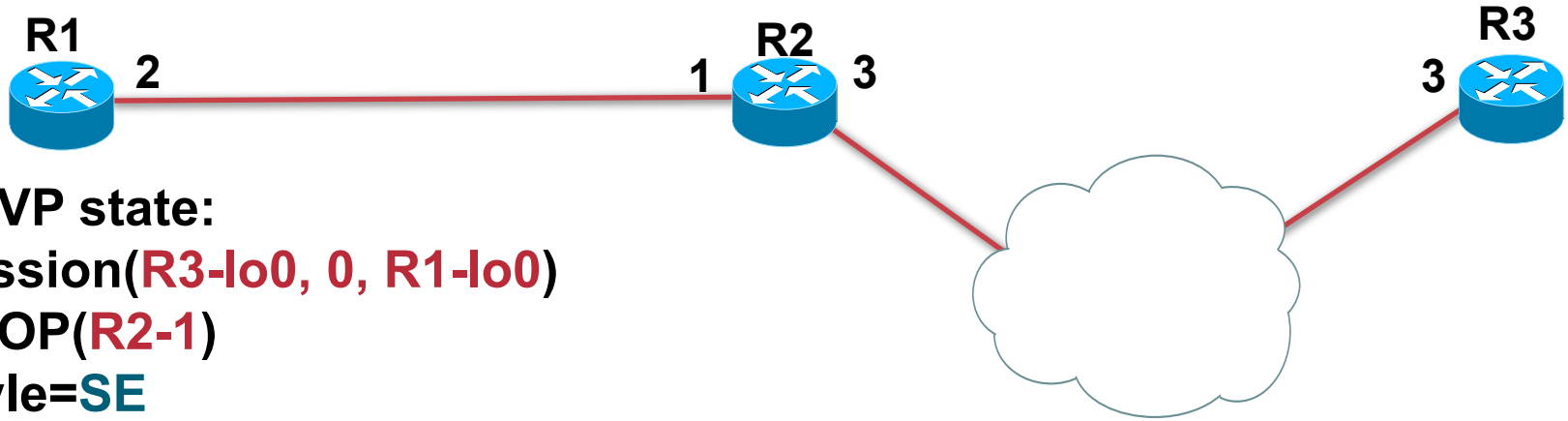
© 2006 Cisco Systems, Inc. All rights reserved.

Make before break in action



© 2006 Cisco Systems, Inc. All rights reserved.

Make before break in action



RSVP state:
Session(R3-lo0, 0, R1-lo0)
PHOP(R2-1)
Style=SE
FlowSpec (3Mbps)
Sender_Template(R1-lo0, 01)
Label=6
Record_Route(R2-1, …, R3-3)
Sender_Template(R1-lo0, 00)
Label=5
Record_Route(R2-1, R3-1)

© 2006 Cisco Systems, Inc. All rights reserved.

Re-optimization

- Periodically, a tunnel can rerun PCALC to see if a better path exists to destination.
- Better path will have a lower IGP metric or fewer hops
- If better path exists, headend signals the tunnel via the better path using “make before break”
- Reoptimization happens in the orer of tunnel ID

© 2006 Cisco Systems, Inc. All rights reserved.

Re-optimization Triggers

- **Periodic:** by default triggered every 3600 seconds (or CLI configured period) for all TE tunnels in the

  order of priority (0 thru 7)

  within each priority based on the tunnel ID

```
mpls traffic-eng reoptimize timers frequency <1-604800 sec>
```

- **Event triggered:** event such as a link coming up will trigger reoptimization
- **Manual:** reoptimize one or all tunnels at the command prompt

```
mpls traffic-eng reoptimize   (all tunnels)
mpls traffic-eng reoptimize Tunnel <0-2147483647>  (per tunnel)
```

© 2006 Cisco Systems, Inc. All rights reserved.

Disabling Re-optimization

- **One or all tunnels can be disabled for reoptimization if we think that the tunnel does not need reoptimization**

```
mpls traffic-eng reoptimize timers frequency 0 (disables all
   tunnels)


interface tunnel0

tunnel mpls traffic-eng path-option 1 dynamic lockdown (disable
   tunnel0)
```

© 2006 Cisco Systems, Inc. All rights reserved.

MPLS-TE: traffic aspects of TE tunnels

© 2005 Cisco Systems, Inc. All rights reserved.

Agenda

- Mapping Traffic to Paths
- Using metrics with tunnels
- Load balancing with TE tunnels
- Monitoring traffic with TE tunnels

© 2006 Cisco Systems, Inc. All rights reserved.

Mapping Traffic to Path

© 2005 Cisco Systems, Inc. All rights reserved.

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

© 2006 Cisco Systems, Inc. All rights reserved.

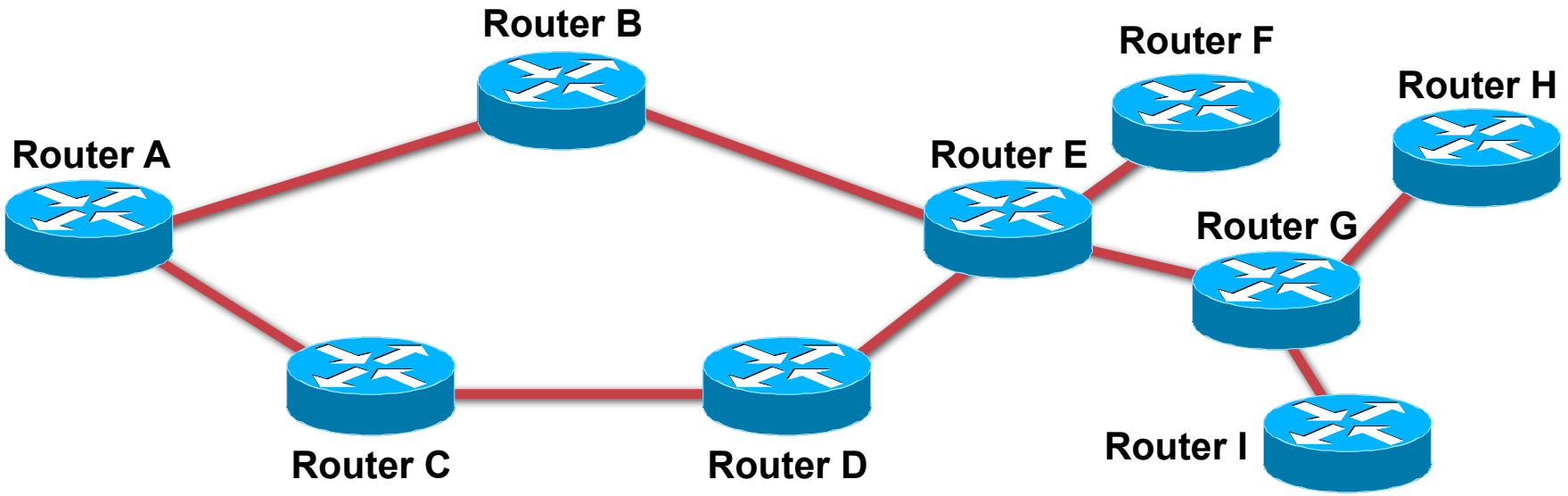
Autoroute

- 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

© 2006 Cisco Systems, Inc. All rights reserved.

Autoroute

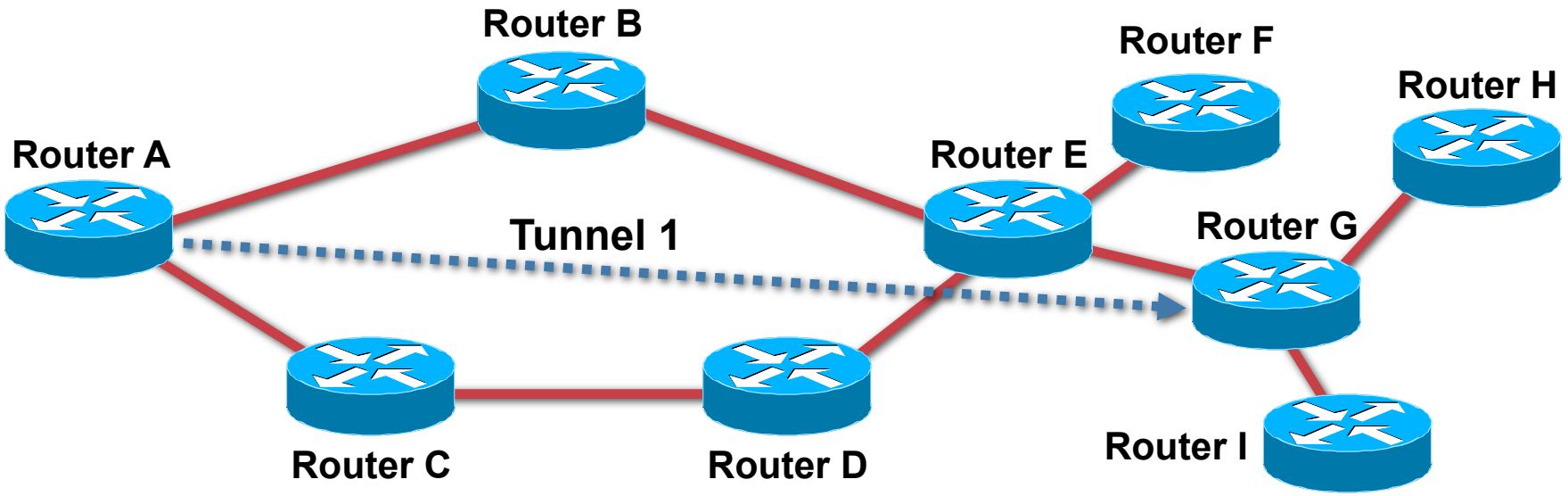
This Is the Physical Topology



© 2006 Cisco Systems, Inc. All rights reserved.

Autoroute

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

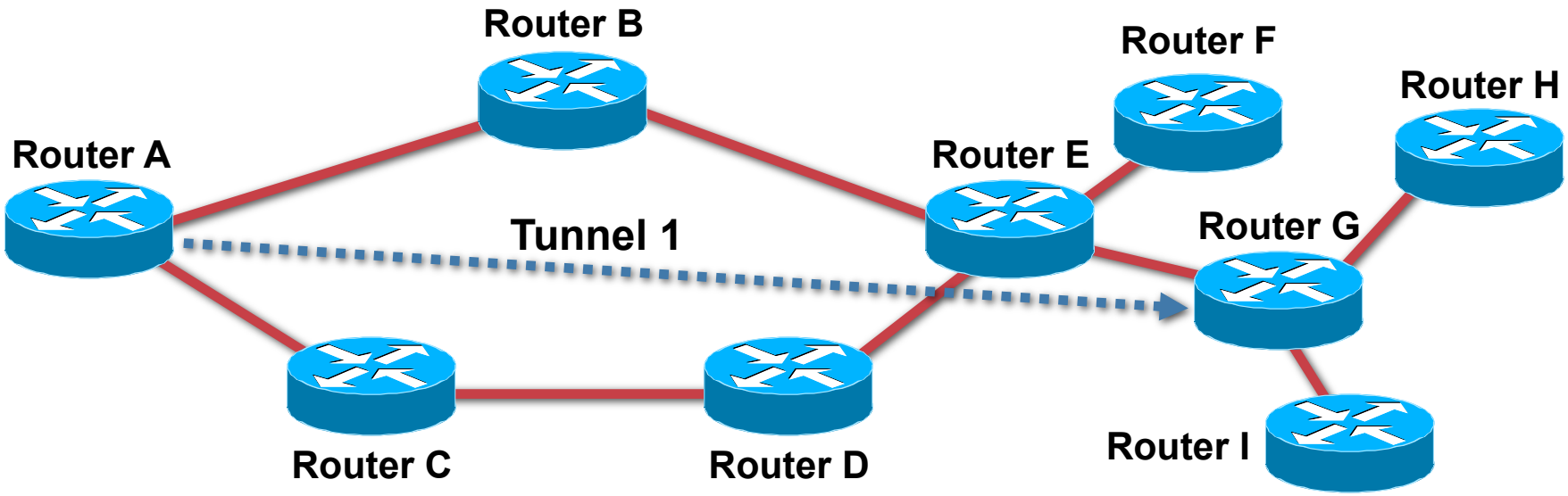


© 2006 Cisco Systems, Inc. All rights reserved.

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

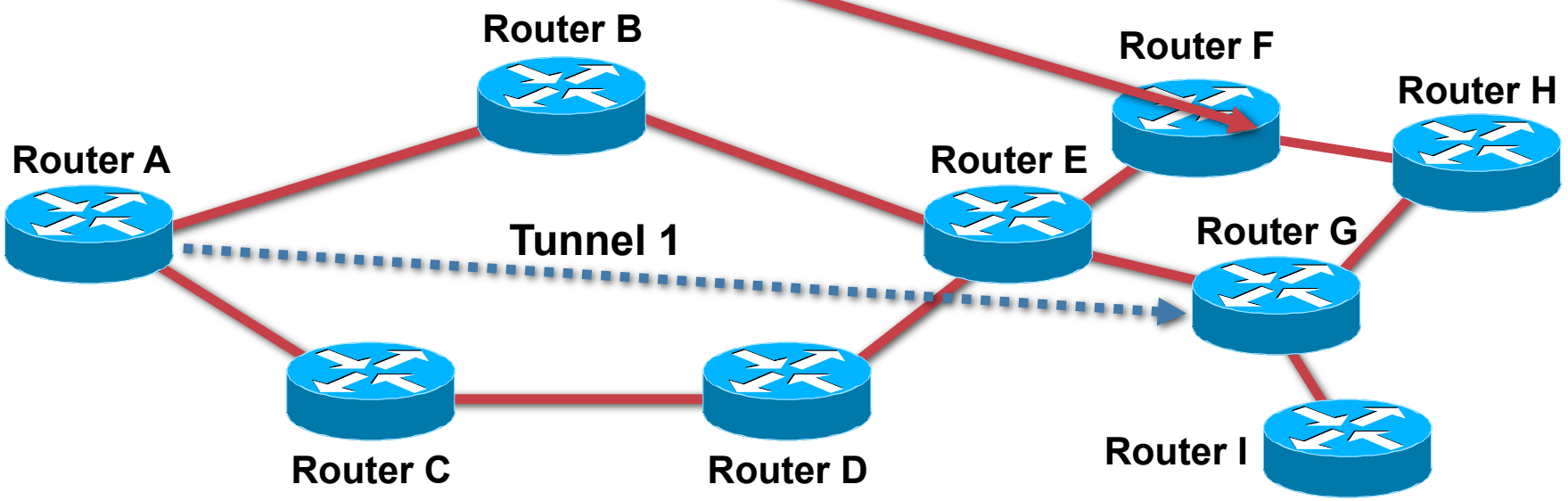


© 2006 Cisco Systems, Inc. All rights reserved.

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

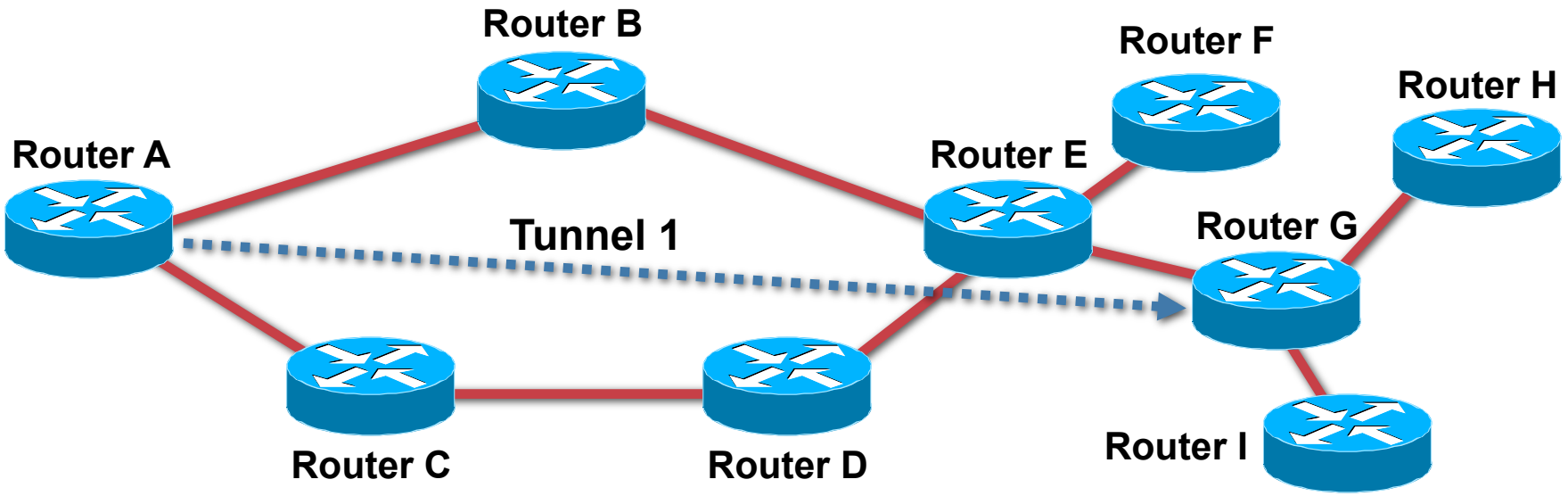


© 2006 Cisco Systems, Inc. All rights reserved.

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

```
interface Tunnel1
 tunnel mpls traffic-eng autoroute announce
```



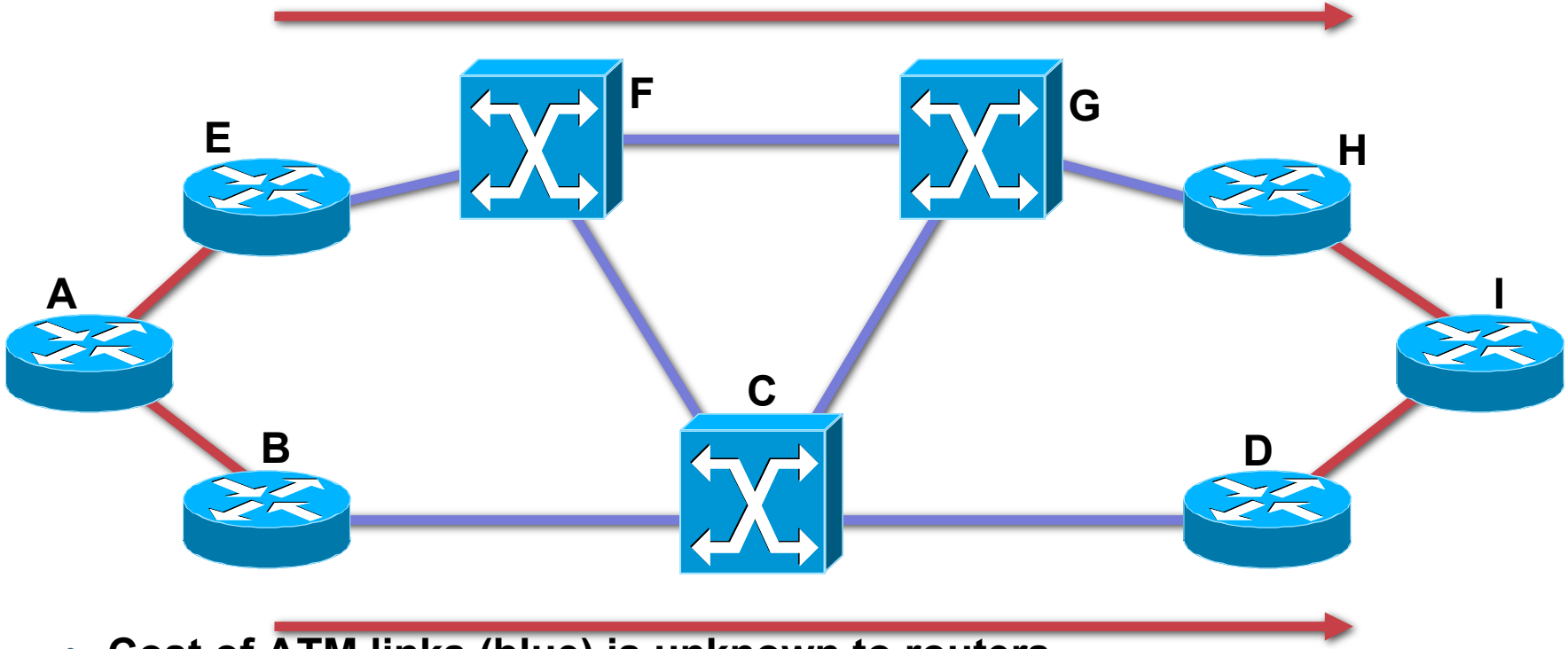
© 2006 Cisco Systems, Inc. All rights reserved.

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

© 2006 Cisco Systems, Inc. All rights reserved.

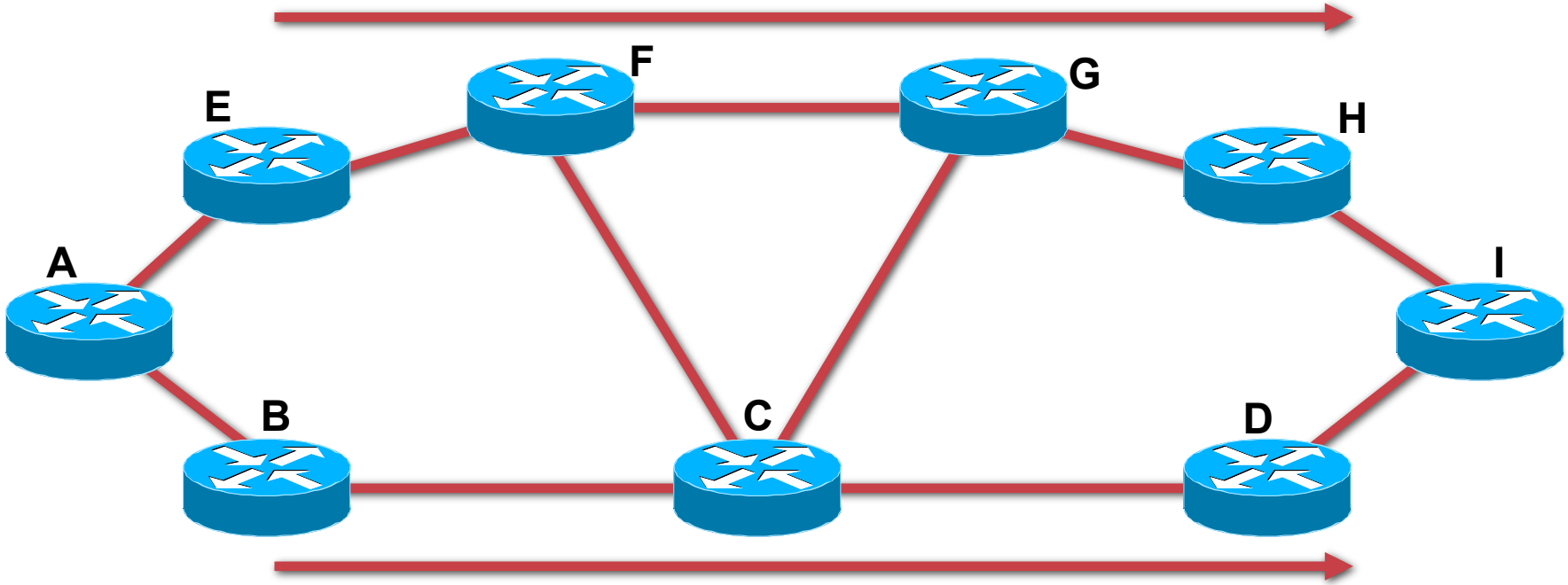
ATM Model



- 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

© 2006 Cisco Systems, Inc. All rights reserved.

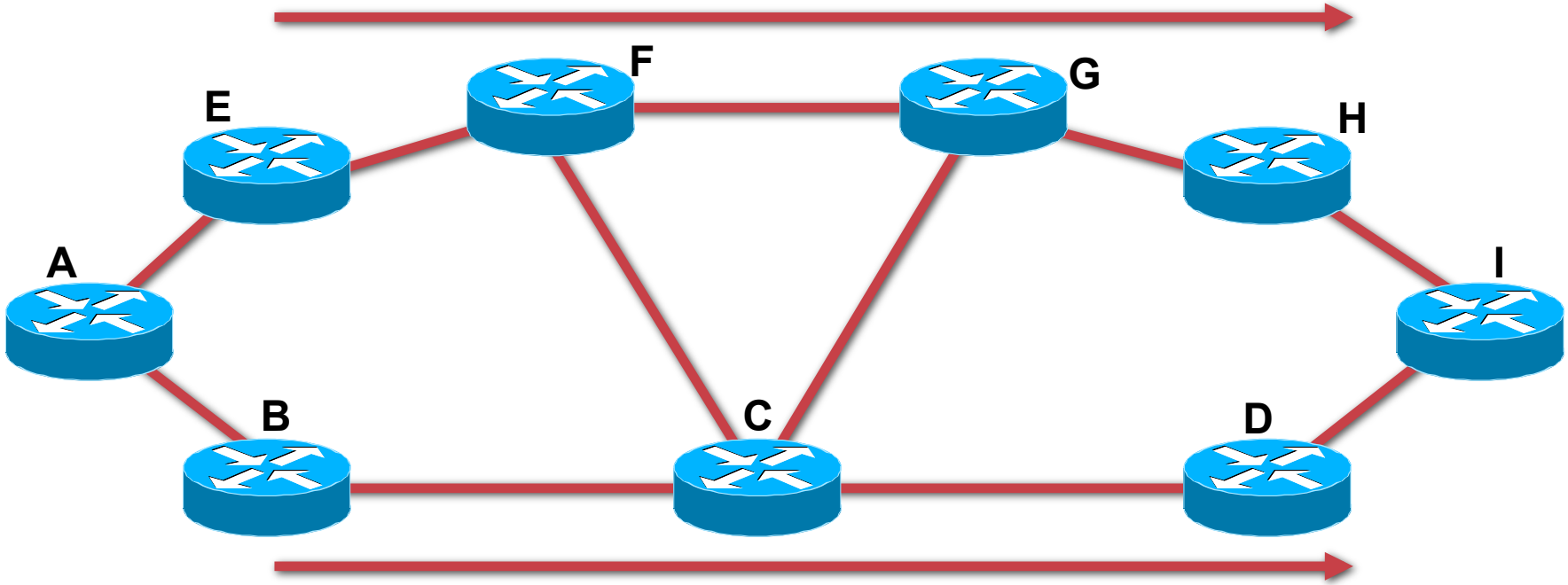
Before FA



- 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

© 2006 Cisco Systems, Inc. All rights reserved.

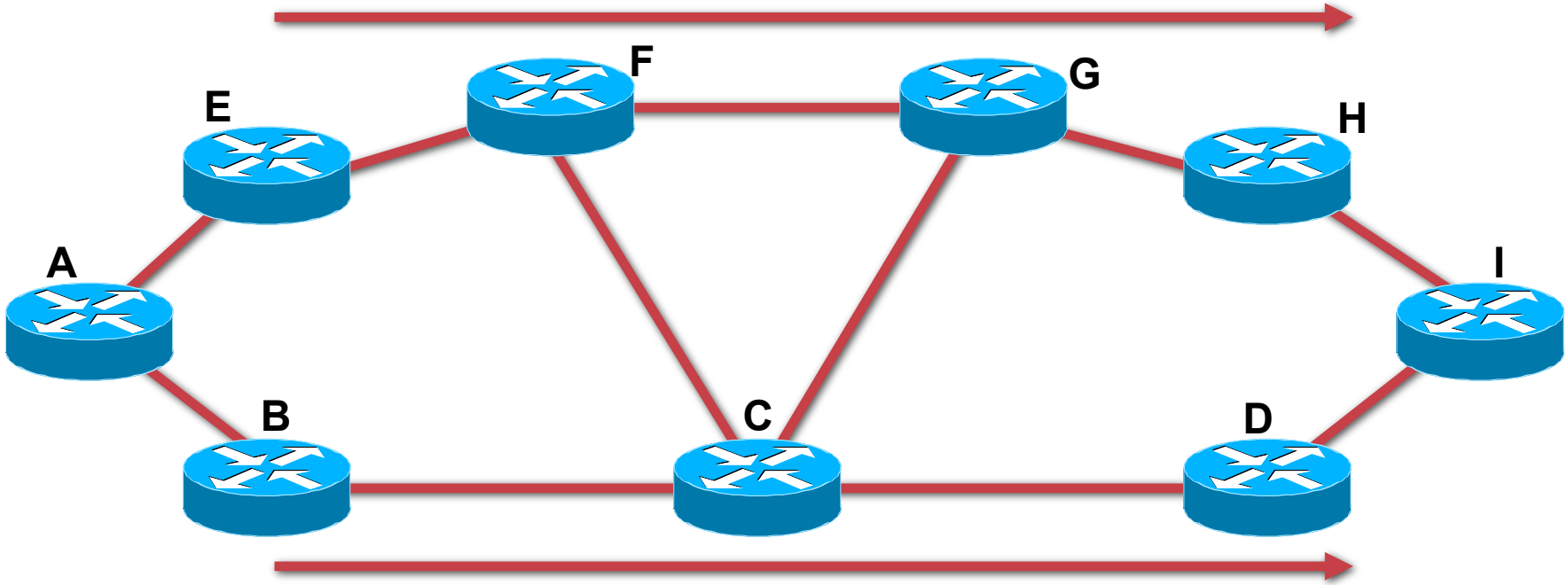
FA Advertises TE Tunnels in the IGP



- 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)

© 2006 Cisco Systems, Inc. All rights reserved.

FA Advertises TE Tunnels in the IGP

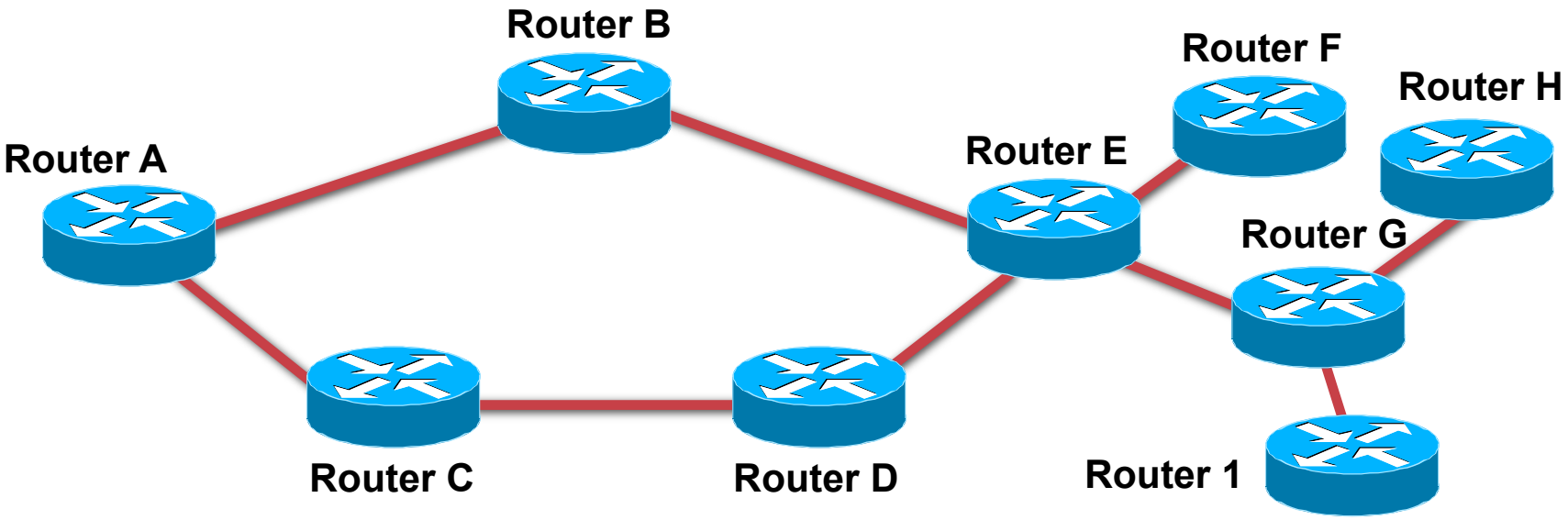



```
tunnel mpls traffic-eng forwarding-adjacency
isis metric <x> level-<y>
        OR
ip ospf cost <x>
```

© 2006 Cisco Systems, Inc. All rights reserved.

Static Routing

```
RtrA(config)#ip route H.H.H.H 255.255.255.255 Tunnel1
```

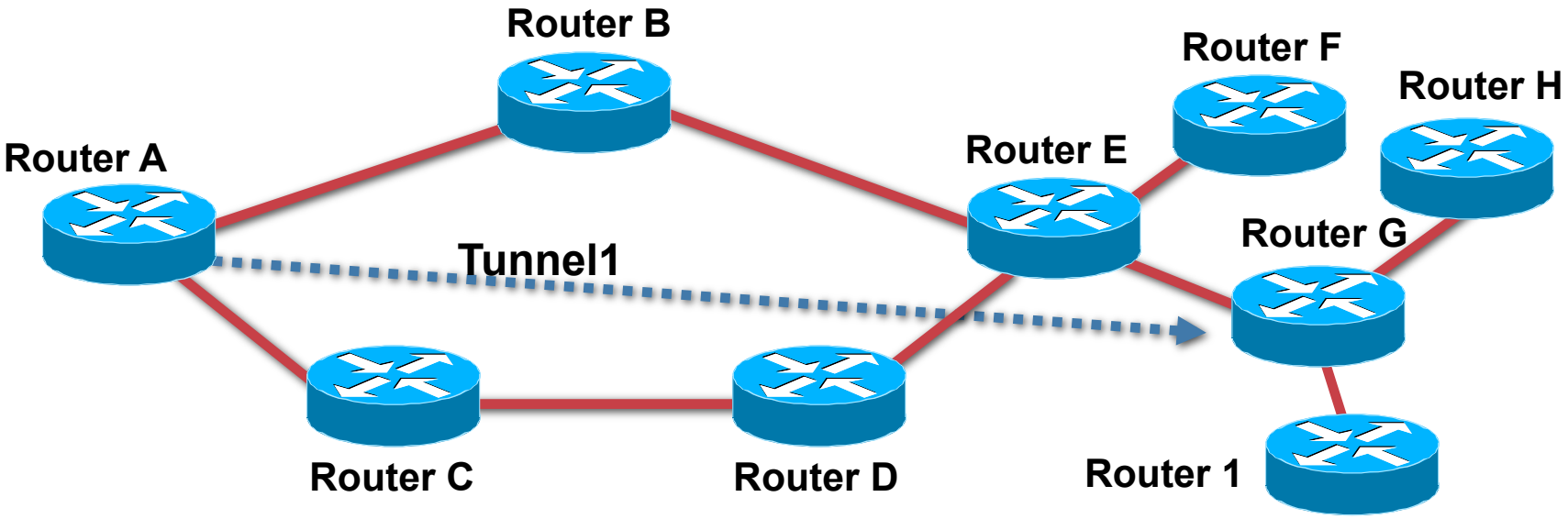


© 2006 Cisco Systems, Inc. All rights reserved.

Static Routing

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

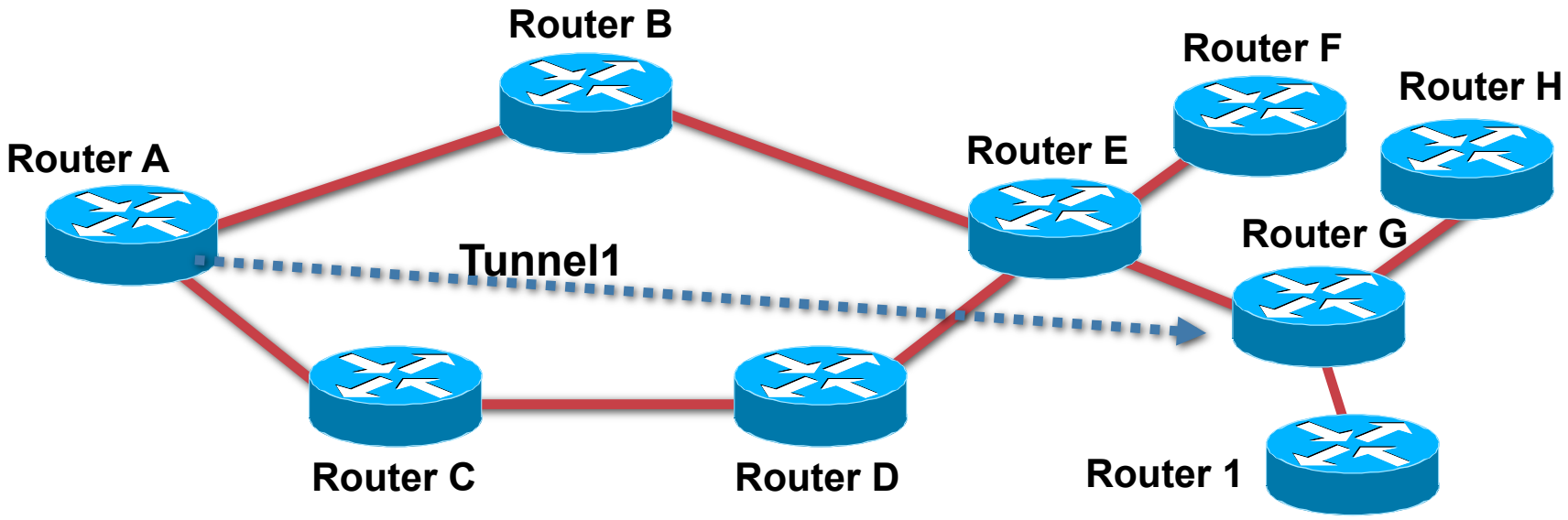
- Router H is known via the tunnel
- Router G is not routed to over the tunnel, even though it's the tunnel tail!



© 2006 Cisco Systems, Inc. All rights reserved.

Policy Routing

```
RtrA(config-if)#ip policy route-map set-tunnel
RtrA(config)#route-map set-tunnel
RtrA(config-route-map)#match ip address 101
RtrA(config-route-map)#set interface Tunnel1
```

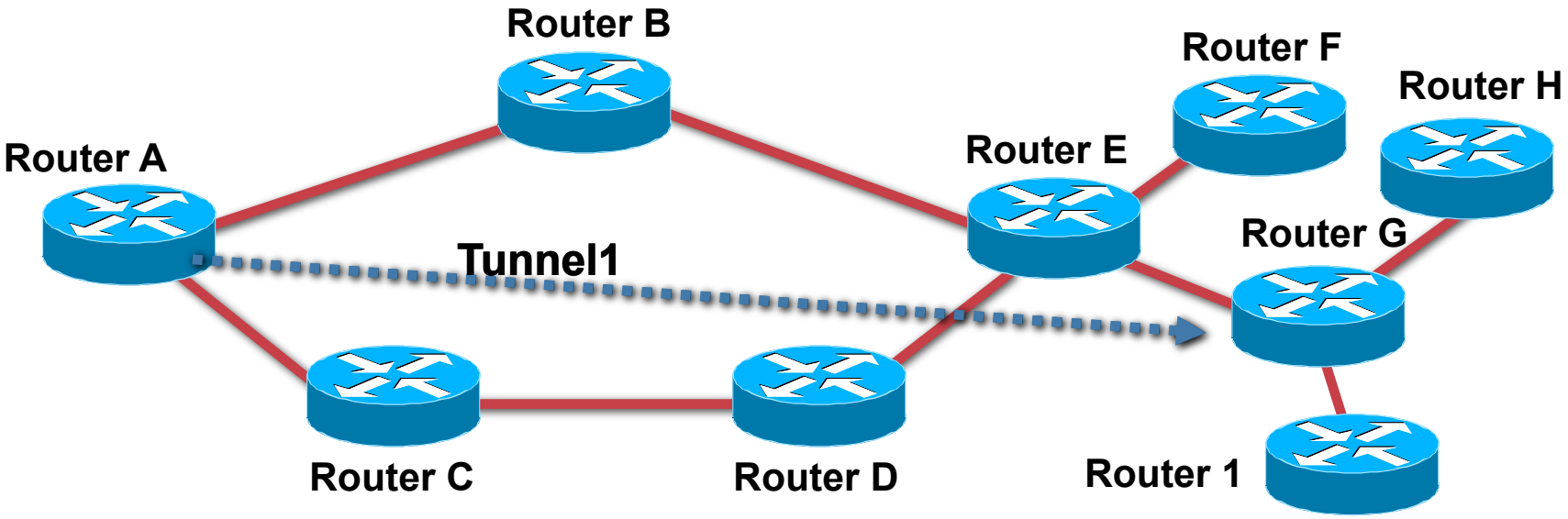


© 2006 Cisco Systems, Inc. All rights reserved.

Policy Routing

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

- Routing table isn't affected by policy routing



© 2006 Cisco Systems, Inc. All rights reserved.

Enhancement to SPF - metric check

Tunnel metric:

A. **Relative +/- X**
B. **Absolute Y (only for ISIS)**
C. **Fixed Z**

Example:

Metric of native IP path to the found node = 50

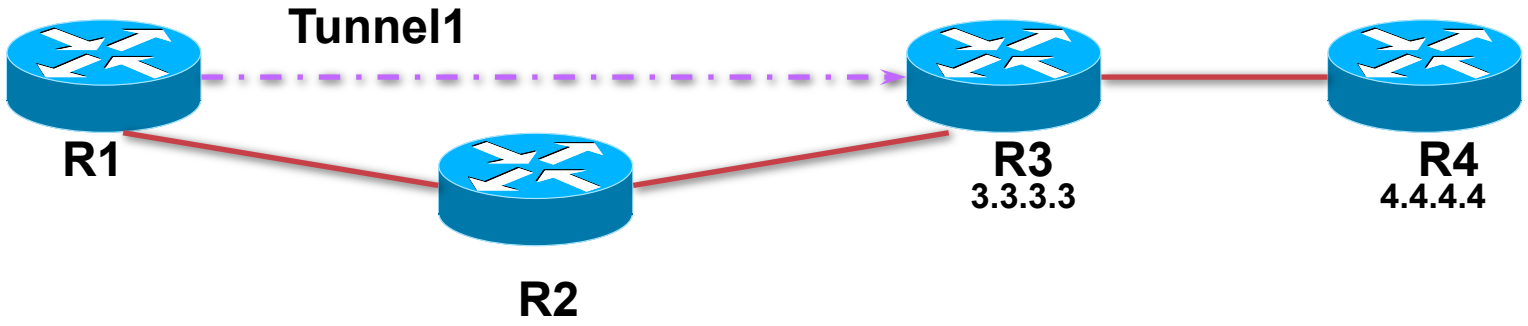
1. Tunnel with relative metric of -10 => 40

2. Tunnel with relative metric of +10 => 60

3. Tunnel with absolute metric of 10 => 10

© 2006 Cisco Systems, Inc. All rights reserved.

Absolute/Relative/Fixed Metric in action



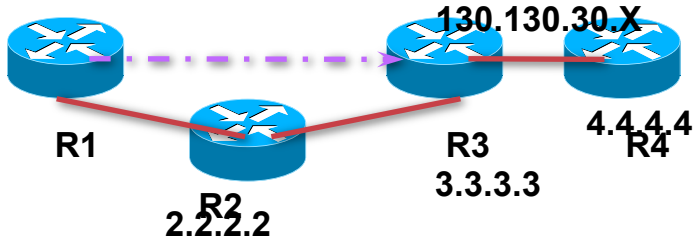
Routing Table on R1 (with all link metrics=10)

IP Addr	Cost	Next-Hop	Interface
4.4.4.4	30	3.3.3.3	Tunnel1
3.3.3.3	20	3.3.3.3	Tunnel1

© 2006 Cisco Systems, Inc. All rights reserved.

Relative Metric in action

Metric to the tunnel tailend is the same **“Relative metric”.** Anything downstream to the tunnel tail is added to the relative metric




```
R1(config-if)#interface tunnel1
R1(config-if)#tunnel mpls traffic-eng autoroute metric relative -5
```

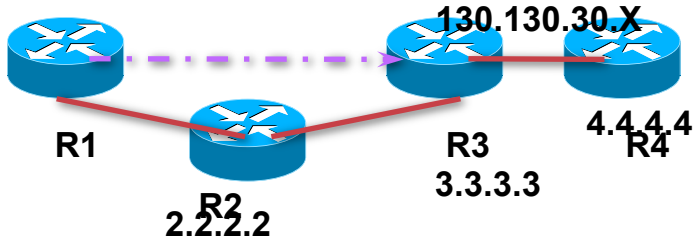
Routing Table on R1

IP Addr	Cost	Next-Hop	Interface
4.4.4.4	25	3.3.3.3	Tunnel1
3.3.3.3	15	3.3.3.3	Tunnel1

© 2006 Cisco Systems, Inc. All rights reserved.

Fixed Metric in action

Metric to the tunnel tailend is the same **“Fixed metric”.** Anything downstream to the tunnel tail is added to the fixed metric




```
R1(config-if)#interface tunnel1
R1(config-if)#tunnel mpls traffic-eng autoroute metric 5
```

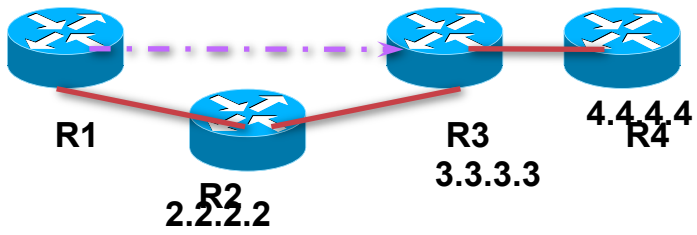
Routing Table on R1

IP Addr	Cost	Next-Hop	Interface
4.4.4.4	15	3.3.3.3	Tunnel1
3.3.3.3	5	3.3.3.3	Tunnel1

© 2006 Cisco Systems, Inc. All rights reserved.

Absolute Metric in action

Metric to the tunnel tailend and downstream destinations is the same **"Absolute metric"** value




```
R1(config-if)#interface tunnel1
R1(config-if)#tunnel mpls traffic-eng autoroute metric absolute
   2
```

Routing Table on R1

IP Addr	Cost	Next-Hop	Interface
4.4.4.4	2	3.3.3.3	Tunnel1
3.3.3.3	2	3.3.3.3	Tunnel1

© 2006 Cisco Systems, Inc. All rights reserved.

Load Sharing with TE tunnels

© 2005 Cisco Systems, Inc. All rights reserved.

Unequal Cost Load Balancing

- **IP routing has equal-cost load balancing, but not unequal cost***

***EIGRP Has 'Variance', but That's Not as Flexible**

© 2006 Cisco Systems, Inc. All rights reserved.

Unequal Cost Load Balancing

- A TE tunnel does not load share traffic between itself and the native IP path it takes
- Multiple parallel tunnels can load share traffic based on bandwidth. This can be equal or unequal cost load balancing
- TE tunnels and native IP links can load share traffic, provided the destination is downstream to the tunnel destination. In this case load sharing is equal cost

© 2006 Cisco Systems, Inc. All rights reserved.

Unequal Cost Example

```
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
```

© 2006 Cisco Systems, Inc. All rights reserved.

Monitoring Traffic in TE tunnels

© 2005 Cisco Systems, Inc. All rights reserved.

Monitoring Traffic in TE tunnels

- TE tunnels do not police traffic. This means that we could send 10 Gbps of traffic via a 10 Mbps tunnel.
- No automatic correlation between tunnel bandwidth and real traffic thru tunnel
- Auto Bandwidth enables a tunnel to adjust bandwidth based on traffic flow

© 2006 Cisco Systems, Inc. All rights reserved.

Auto Bandwidth

- **Tunnel monitors traffic say every 5 minutes and records the largest sample. At the end of 24 hour period, the tunnel applies the largest sample to its bandwidth statement in the configuration**
- **We can also define a floor and ceiling to bandwidth beyond which no change will be applied to bandwidth statement**

© 2006 Cisco Systems, Inc. All rights reserved.

Enabling Auto-Bandwidth

```
mpls traffic-eng auto-bw timers frequency
<0-604800>
```

- **Global command**
- **Enables tunnels to sample load at the configured frequency**
- **Should not be less than the “load interval” on the interface**

© 2006 Cisco Systems, Inc. All rights reserved.

Enabling Auto-Bandwidth

```
tunnel mpls traffic-eng auto-bw ?
  collect-bw Just collect Bandwidth info on this tunnel
  frequency  Frequency to change tunnel BW
  max-bw     Set the Maximum Bandwidth for auto-bw on this tunnel
  min-bw     Set the Minimum Bandwidth for auto-bw on this tunnel
  <cr>
```

- **Per-tunnel command**
- **Periodically changes tunnel BW reservation based on traffic out tunnel**
- **Timers are tunable to make auto-bandwidth more or less sensitive**

  Tradeoff: Quicker reaction versus more churn

© 2006 Cisco Systems, Inc. All rights reserved.

MPLS-TE: Advanced TE topics

© 2005 Cisco Systems, Inc. All rights reserved.

Agenda

- **MPLS-TE Rerouting**
- **Fast Reroute (Link, Node and Path)**
- **Inter-area/Inter-AS TE**

© 2006 Cisco Systems, Inc. All rights reserved.

MPLS TE rerouting

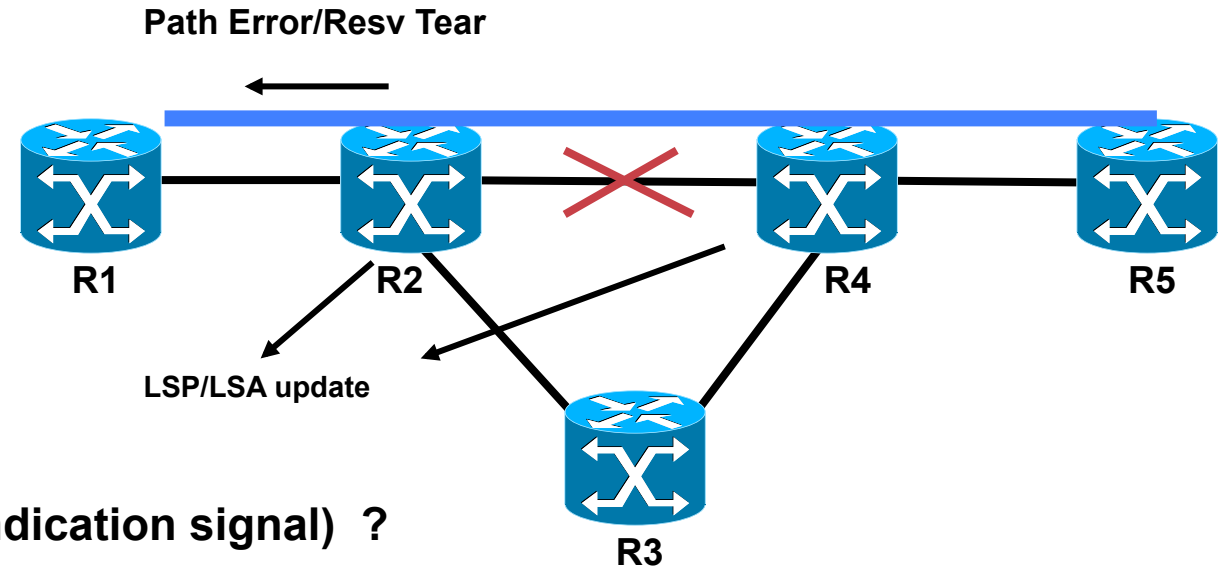
LSP rerouting

- **Controlled by the head-end of a trunk via the resilience attribute of the trunk**
- **Fallback to either (pre)configured or dynamically computed path. Preferably last path option should be dynamic**

```
interface Tunnel0
 ip unnumbered Loopback0
 no ip directed-broadcast
 tunnel destination 10.0.1.102
 tunnel mode mpls traffic-eng
 tunnel mpls traffic-eng autoroute announce
 tunnel mpls traffic-eng priority 3 3
 tunnel mpls traffic-eng bandwidth 10000
 tunnel mpls traffic-eng path-option 1 explicit name prim_path
 tunnel mpls traffic-eng path-option 2 dynamic
```

```
ip explicit-path name prim_path enable
 next-address 10.0.1.123
 next-address 10.0.1.100
```

MPLS TE rerouting

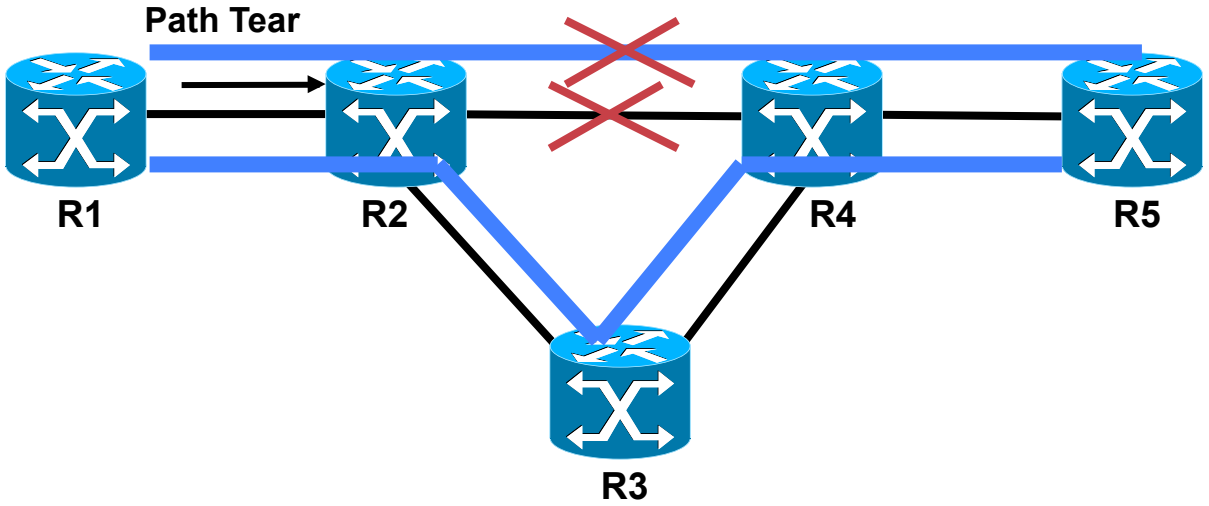


- **The FIS (failure indication signal) ?**
  - *** R1 may receive a Path Error from R2 and a Resv Tear OR**
  - *** R1 will receive a new LSA/LSP indicating the R2-R4 is down and will conclude the LSP has failed**
  - **Which one on those two events will happen first ? It depends of the failure type and IGP tuning**
- **Receipt of Path Error allows to remove the failed link from the TE database to prevent to retry the same failed link (if the IGP update has not been received yet)**

© 2006 Cisco Systems, Inc. All rights reserved.

MPLS TE rerouting

- R1 is now informed that the LSP has suffered a failure



- R1 clear the Path state with an RSVP Path Tear message
- R1 recalculates a new Path for the Tunnel and will signal the new tunnel. If no Path available, R1 will continuously retry to find a new path (local process)

Convergence = O(secs)

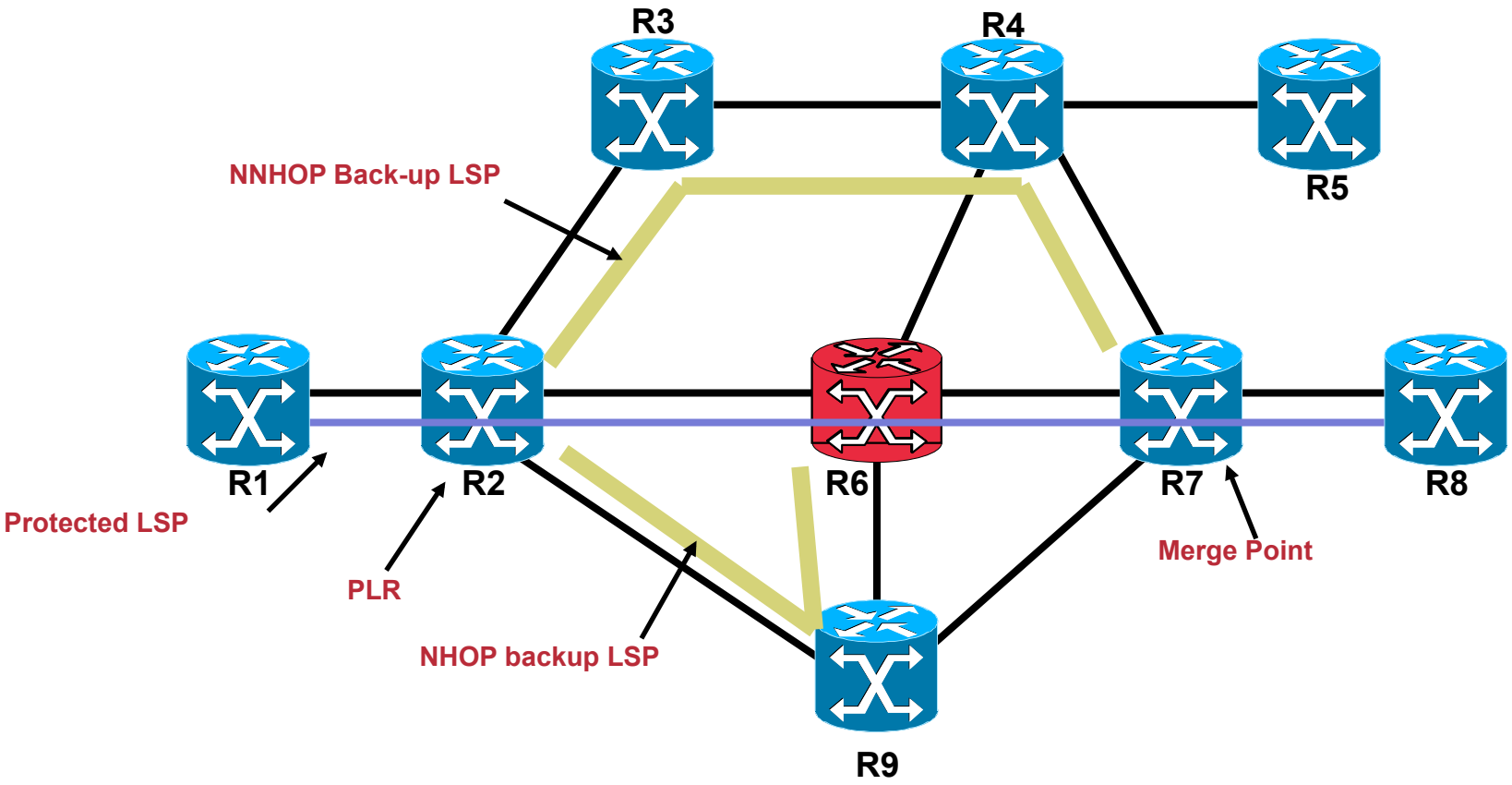
© 2006 Cisco Systems, Inc. All rights reserved.

Fast ReRoute

- FRR builds a path to be used in case of a failure in the network
- Minimize packet loss by taking a quick local decision to reroute at the failure point

© 2006 Cisco Systems, Inc. All rights reserved.

Terminology

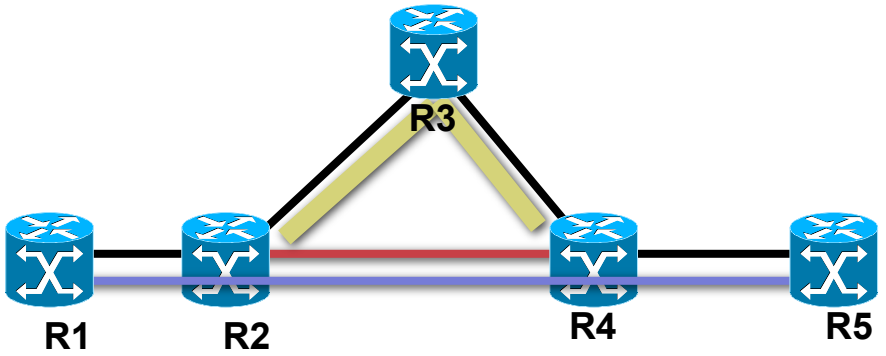


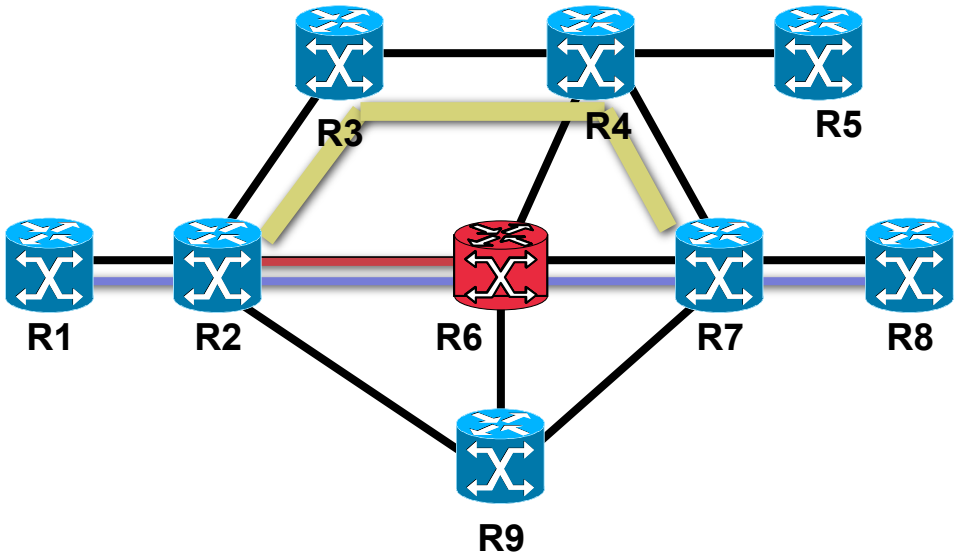
© 2006 Cisco Systems, Inc. All rights reserved.

Fast ReRoute

MPLS Fast Reroute Local Repair

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





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

© 2006 Cisco Systems, Inc. All rights reserved.

IP Failure Recovery

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

© 2006 Cisco Systems, Inc. All rights reserved.

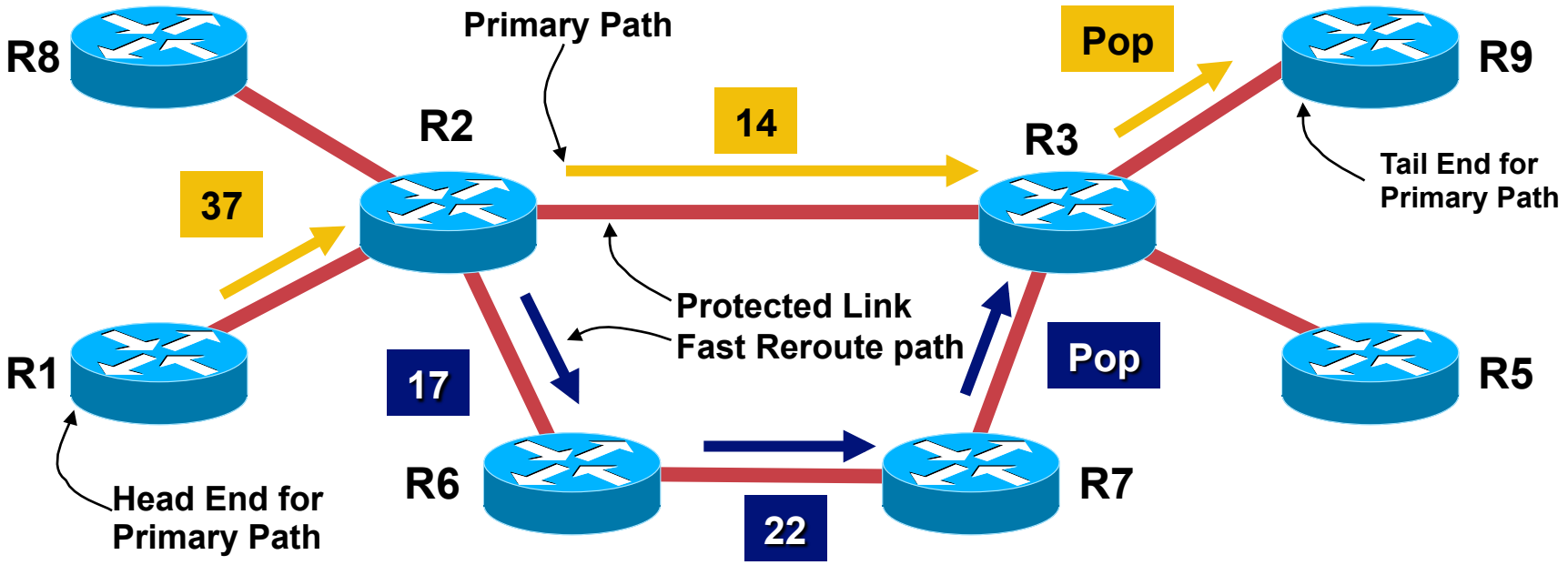
FRR Failure Recovery

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)

© 2006 Cisco Systems, Inc. All rights reserved.

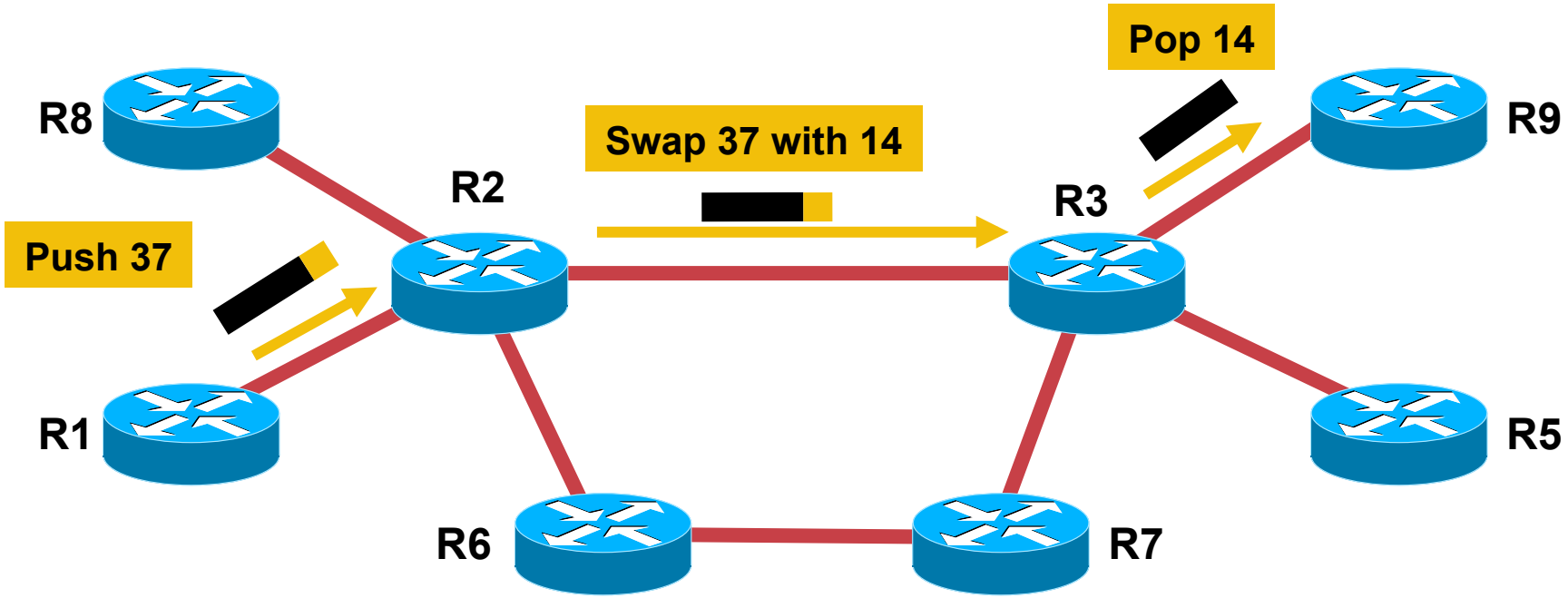
Link Protection Example



Primary path: R1 → R2 → R3 → R9

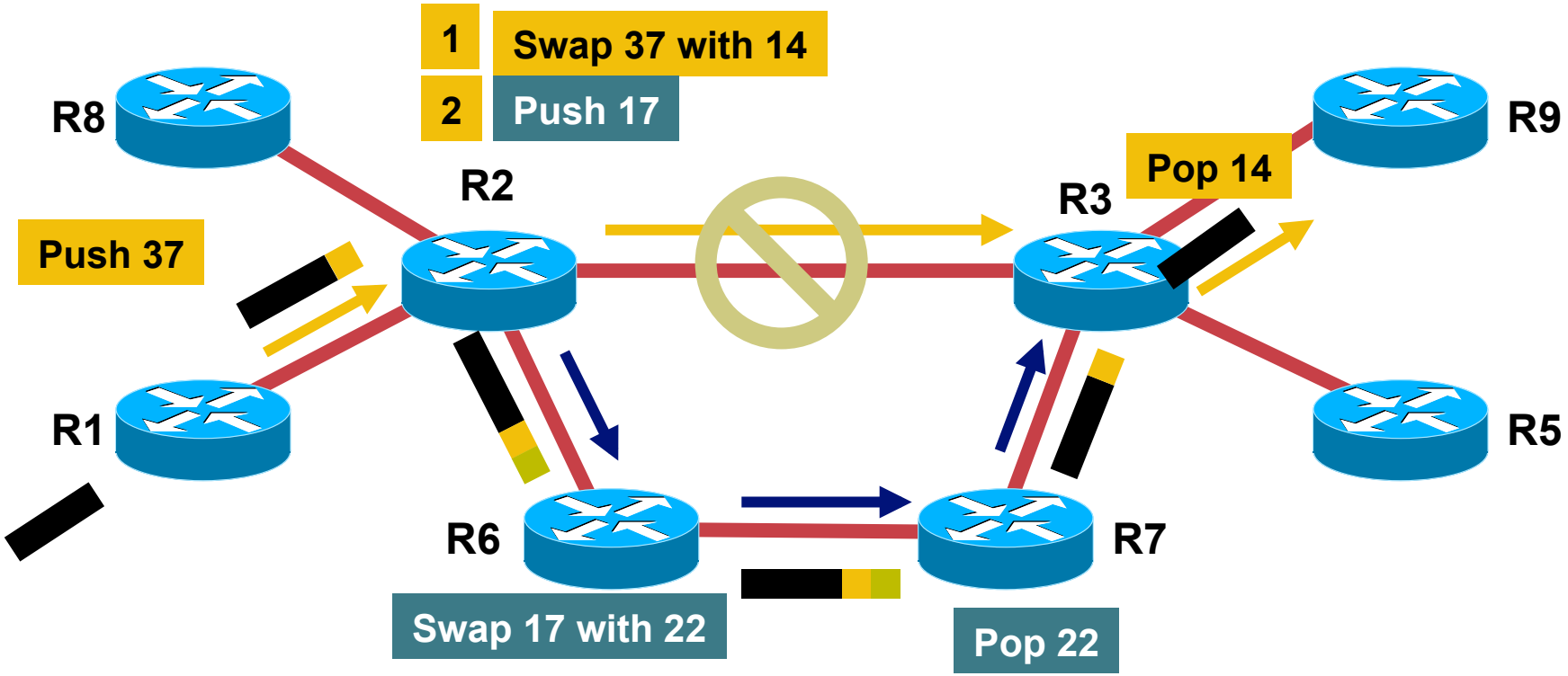
Fast Reroute path: R2 → R6 → R7 → R3

Normal TE Operation



© 2006 Cisco Systems, Inc. All rights reserved.

Fast Reroute Link Failure



© 2006 Cisco Systems, Inc. All rights reserved.

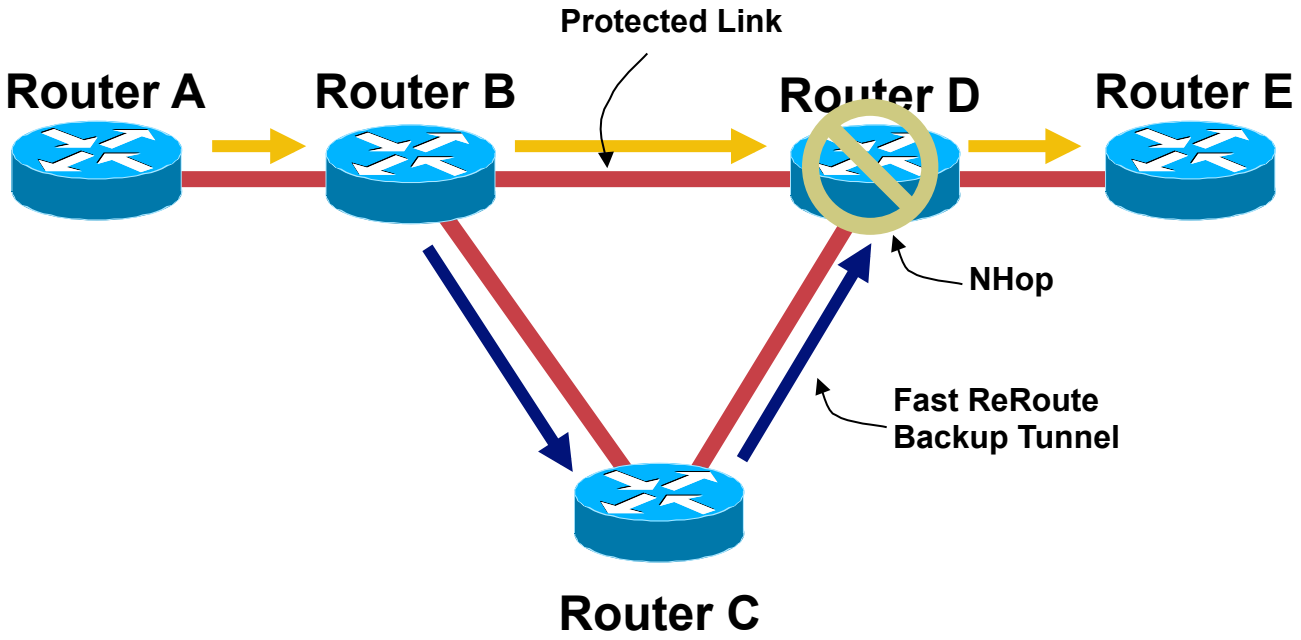
FRR Procedures

1. Pre-establish backup paths
2. Failure happens, protected traffic is switched onto backup paths
3. After local repair, tunnel headends are signaled 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

© 2006 Cisco Systems, Inc. All rights reserved.

Node Protection

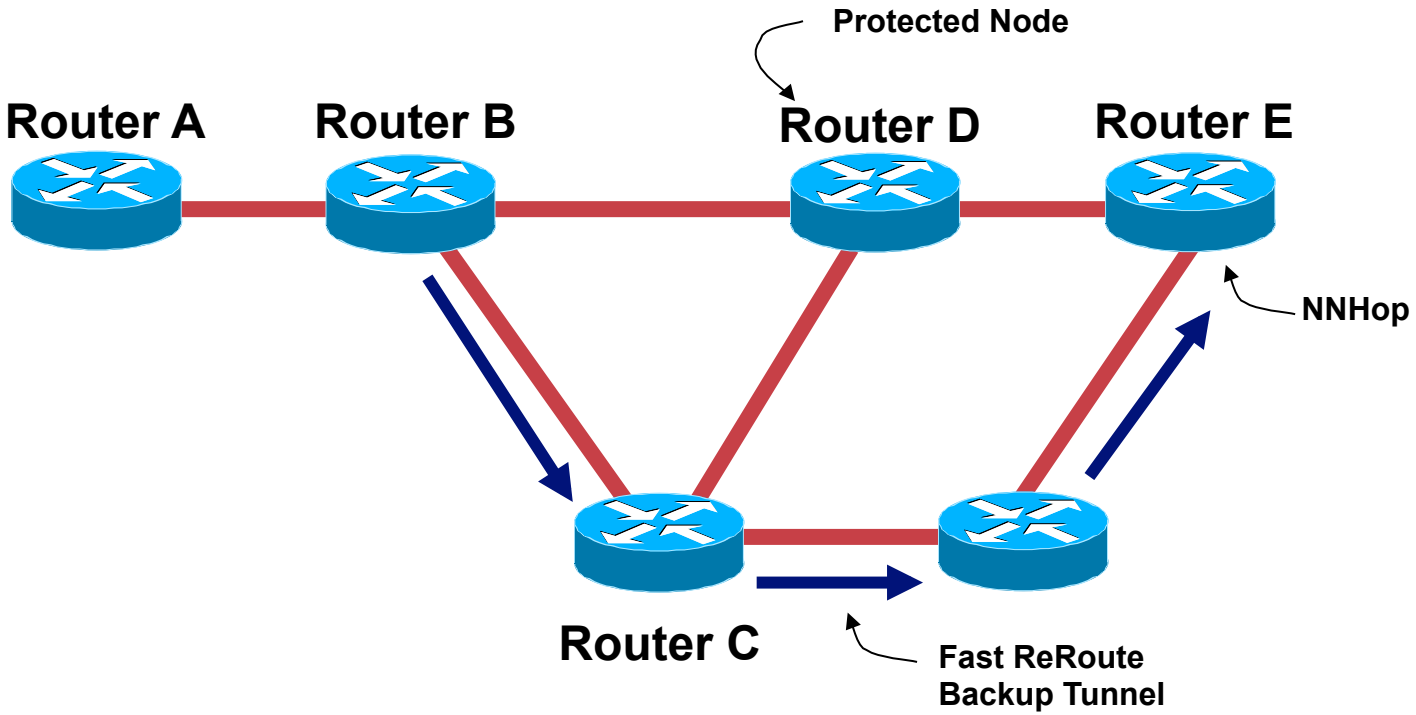
- What if Router D failed?
- Link protection would not help as the backup tunnel terminates on Router D (which is the NHop of the protected link)



© 2006 Cisco Systems, Inc. All rights reserved.

Node Protection

- **SOLUTION: NODE PROTECTION (If network topology allows)**
- **Protect tunnel to the next hop PAST the protected link (NNhop)**



© 2006 Cisco Systems, Inc. All rights reserved.

Node Protection

- Node protection still has the same convergence as link protection
- Deciding where to place your backup tunnels is a much harder problem to solve on a large-scale
- For small-scale protection, link may be better
- Configuration is identical to link protection, except where you terminate the backup tunnel (NNHop vs. NHop)

```
RouterB(config)# ip explicit-path name avoid-node
RouterB(cfg-ip-expl-path)# exclude-address <Router_D>

RouterB(config)# interface Tunnel2
RouterB(config-if)# tunnel mpls traffic-eng path-option explicit
avoid-node
```

© 2006 Cisco Systems, Inc. All rights reserved.

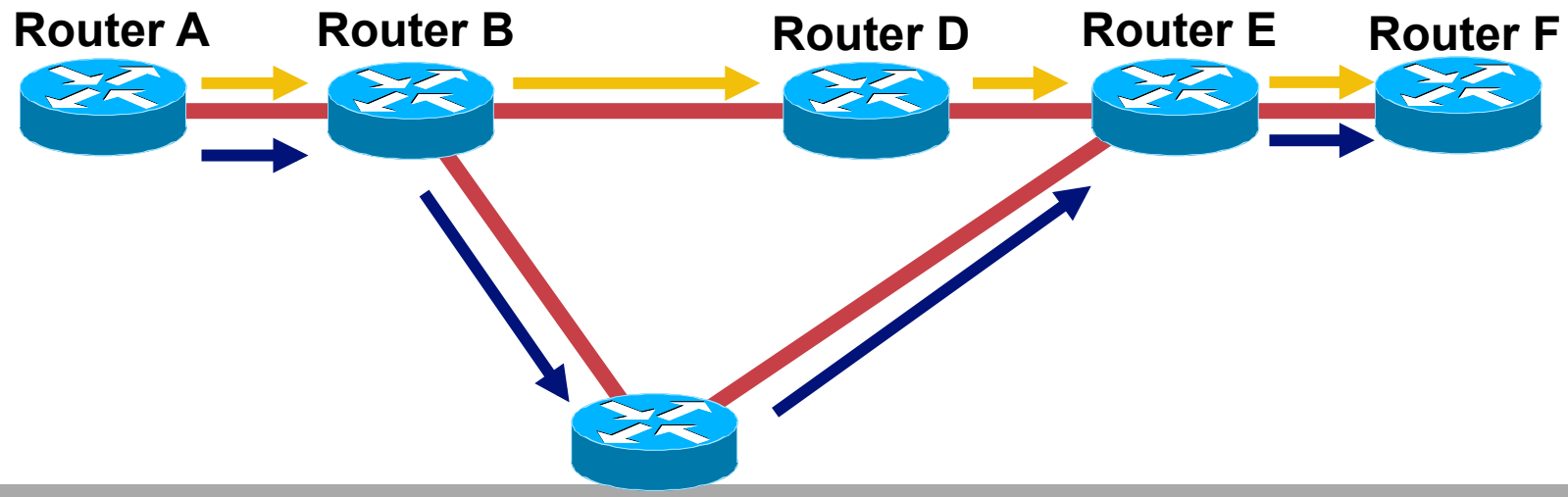
Link and Node Protection Times

- **Link and Node protection are very similar**
- **Protection times are commonly linear to number of protected items**
- **One provider gets ~35ms of loss**

© 2006 Cisco Systems, Inc. All rights reserved.

Path Protection

- **Path protection: Multiple tunnels from TE head to tail, across diverse paths**
- **Backup tunnel pre-signalled. If primary tunnel goes down, tell headend to use backup**



© 2006 Cisco Systems, Inc. All rights reserved.

Path Protection

- Least scalable, most resource-consuming, slowest convergence of all 3 protection schemes
- With no protection, worst-case packet loss is 3x path delay
- With path protection, worst-case packet loss is 1x path delay
- With link or node protection, packet loss is easily engineered to be subsecond (<100ms, <50ms, 4ms, all possible)

© 2006 Cisco Systems, Inc. All rights reserved.

Inter-area TE

- Build LSPs across different OSPF areas
- OSPF uses Opaque LSA (type 10) within area to propagate TE information
- Use explicit path with “loose hop” option
- Each loose hop node is an ABR
- Each ABR will run CSPF to get to the next ABR in its area and inset the nodes in explicit path
- Inter-area tunnels can do reoptimization and FRR
- Autoroute is not supported for Inter-area, since you need to know the topology downstream to the tail

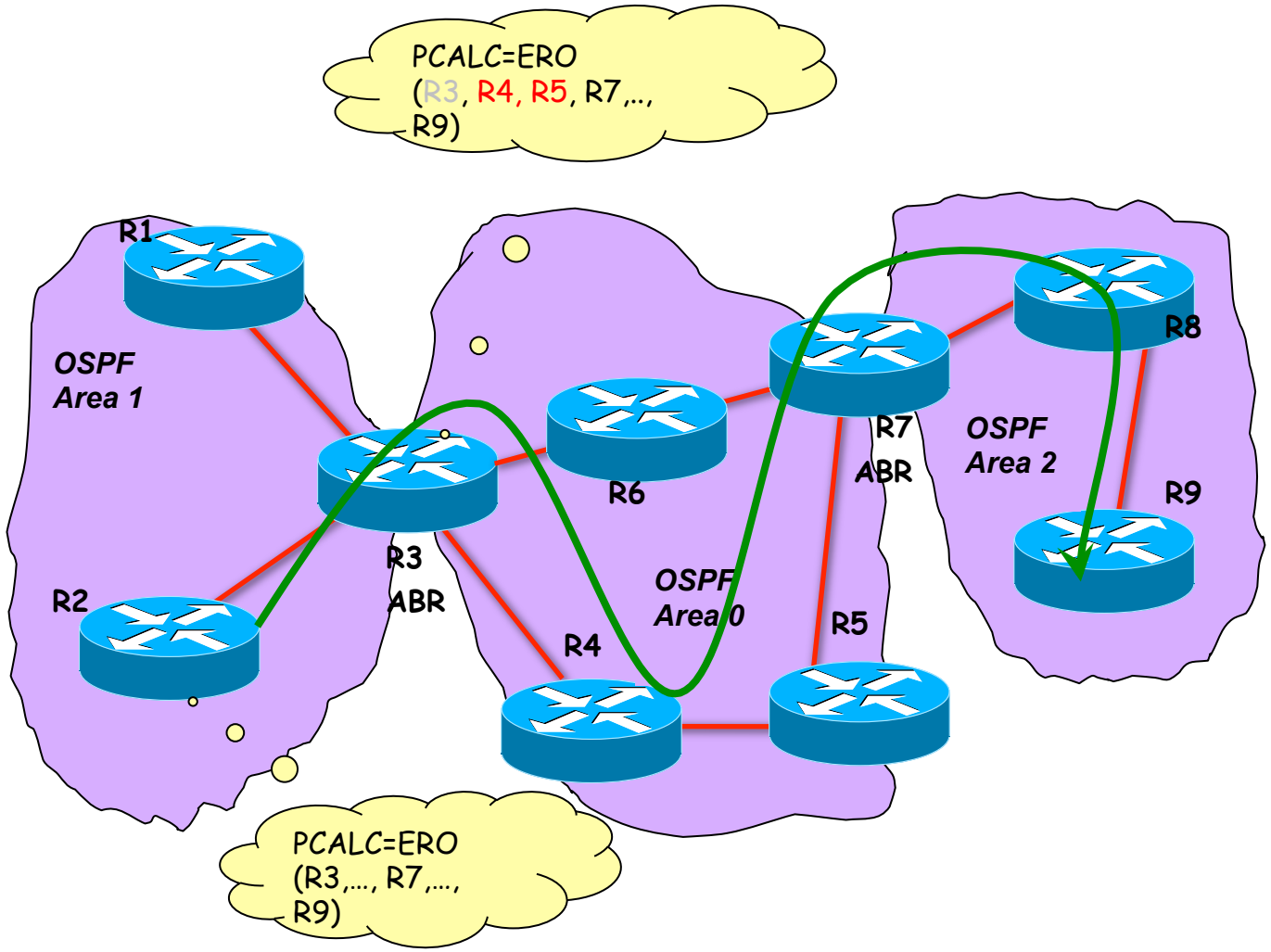
© 2006 Cisco Systems, Inc. All rights reserved.

Enabling Inter-area TE

```
interface Tunnel1
tunnel mpls traffic-eng path-option 1 explicit name
  path-tunnel1
!
ip explicit-path name path-tunnel1
next-address loose <ABR1>
next-address loose <ABR2>
next-address loose <ABR3>
```

© 2006 Cisco Systems, Inc. All rights reserved.

Inter-area TE



© 2006 Cisco Systems, Inc. All rights reserved.

Configuring MPLS-TE

Backup (if time ever permits)

© 2005 Cisco Systems, Inc. All rights reserved.

Prerequisite Configuration (Global)

```
ip cef [distributed]
mpls traffic-eng tunnels
```

© 2006 Cisco Systems, Inc. All rights reserved.

Information Distribution

- **OSPF**

```
mpls traffic-eng tunnels
mpls traffic-eng router-id loopback0
mpls traffic-eng area ospf-area
```

- **ISIS**

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

© 2006 Cisco Systems, Inc. All rights reserved.

Information Distribution

- **On each physical interface**

```
interface pos0/0
  mpls traffic-eng tunnels
  ip rsvp bandwidth Kbps (Optional)
  mpls traffic-eng attribute-flags attributes (Opt)
```

© 2006 Cisco Systems, Inc. All rights reserved.

Build a Tunnel Interface (Headend)

```
interface Tunnel0
  ip unnumbered loopback0
  tunnel destination RID-of-tail
  tunnel mode mpls traffic-eng
   tunnel mpls traffic-eng bandwidth  10
```

© 2006 Cisco Systems, Inc. All rights reserved.

Tunnel Attributes

```
interface Tunnel0
  tunnel mpls traffic-eng bandwidth Kbps
  tunnel mpls traffic-eng priority pri [hold-pri]
  tunnel mpls traffic-eng affinity properties [mask]
  tunnel mpls traffic-eng autoroute announce
```

© 2006 Cisco Systems, Inc. All rights reserved.

Path Calculation

- **Dynamic path calculation**

```
int Tunnel0
   tunnel mpls traffic-eng path-option # dynamic
```

- **Explicit path calculation**

```
int Tunnel0
  tunnel mpls traffic path-opt # explicit name foo

ip explicit-path name foo
   next-address 1.2.3.4 [loose]
   next-address 1.2.3.8 [loose]
```

© 2006 Cisco Systems, Inc. All rights reserved.

Multiple Path Calculations

- **A tunnel interface can have several path options, to be tried successively**

```
Interface Tunnel 1
…….
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
```

© 2006 Cisco Systems, Inc. All rights reserved.

Static and Policy Routing Down a Tunnel

- **Static routing**

```
ip route prefix mask Tunnel0
```

- **Policy routing (Global Table)**

```
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
```

© 2006 Cisco Systems, Inc. All rights reserved.

Autoroute and Forwarding Adjacency

```
interface Tunnel0
  tunnel mpls traffic-eng autoroute announce
                              OR
  tunnel mpls traffic-eng forwarding-adjacency
  isis metric x level-y (ISIS)
  ip ospf cost ospf-cost (OSPF)
```

© 2006 Cisco Systems, Inc. All rights reserved.

L2VPN Concepts

© 2006 Cisco Systems, Inc. All rights reserved.

Agenda

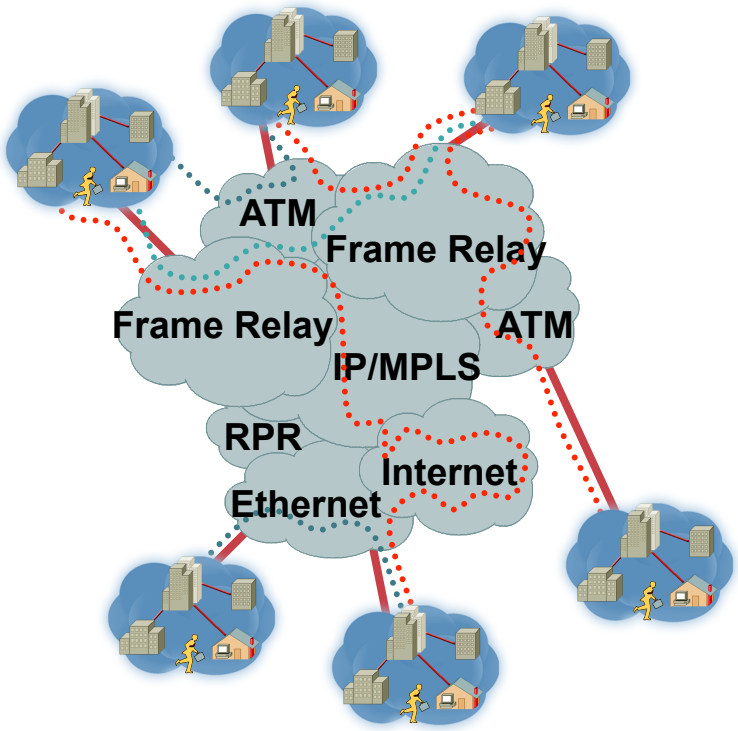
- Introduction to L2VPN
- PWE3 Signaling Concepts
- Virtual Private Wire Service (VPWS) Transports
- VPWS Service Interworking
- Virtual Private LAN Service (VPLS)

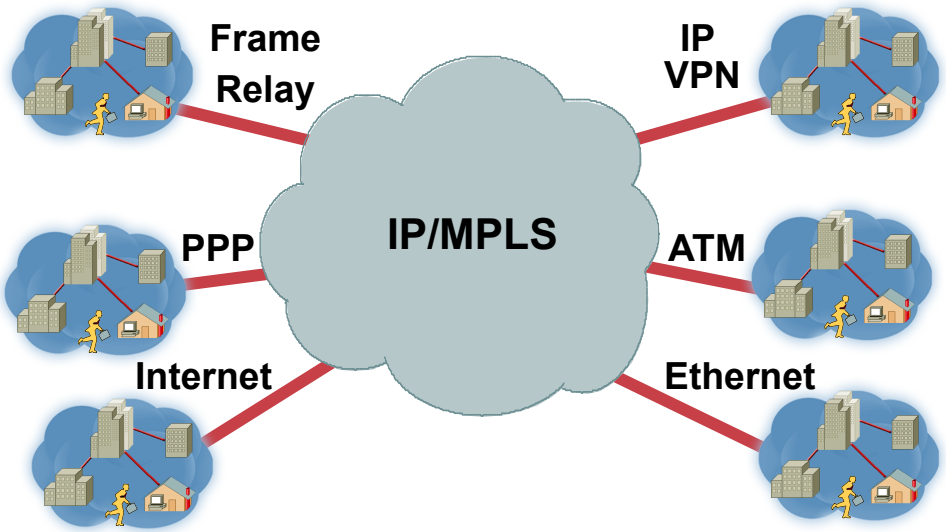
© 2006 Cisco Systems, Inc. All rights reserved.

Introduction to L2VPN

© 2006 Cisco Systems, Inc. All rights reserved.

Multiple Services over a Converged Infrastructure

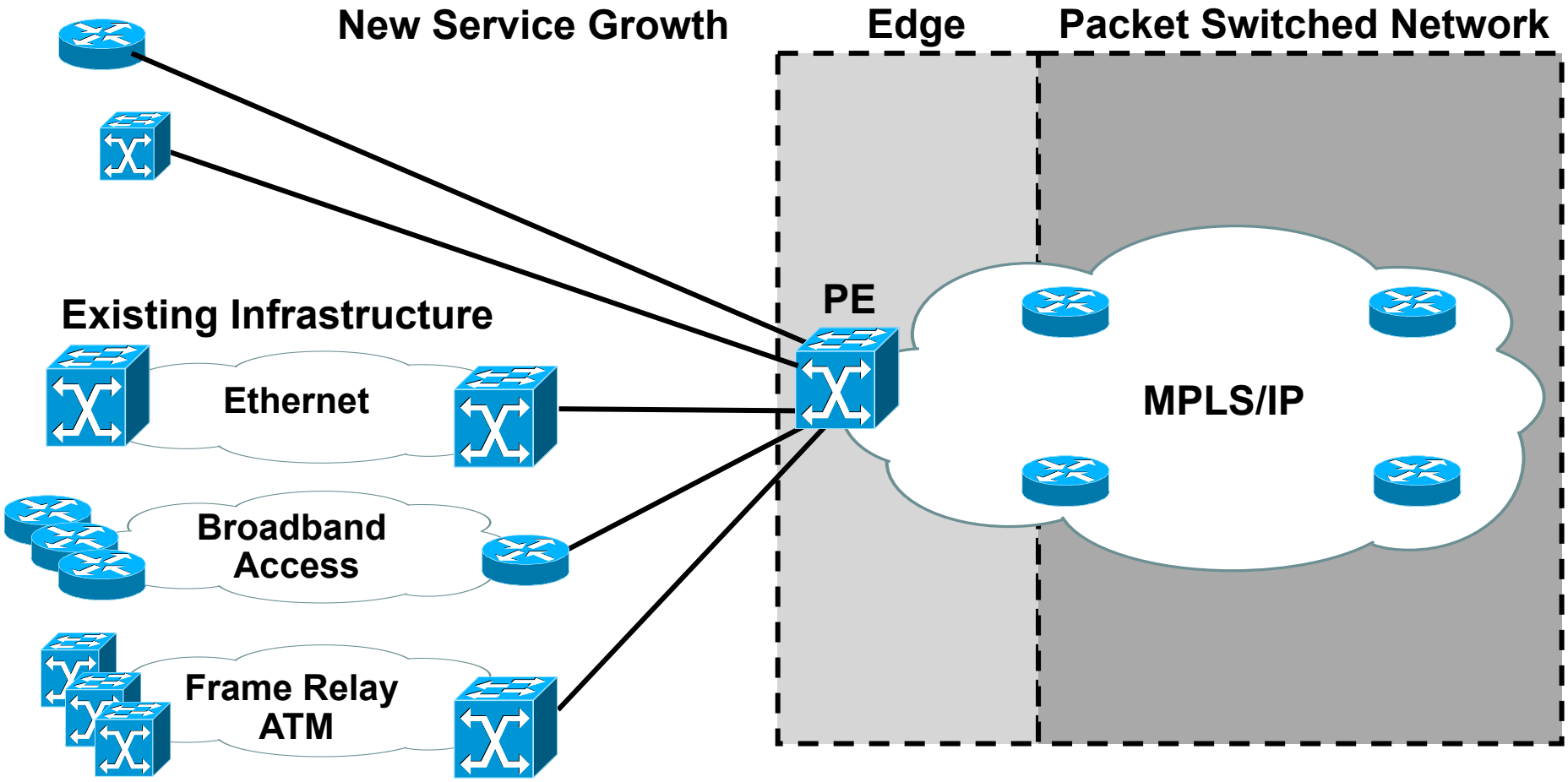




Many Services, Many Networks → **Many Services, One Network**

© 2006 Cisco Systems, Inc. All rights reserved.

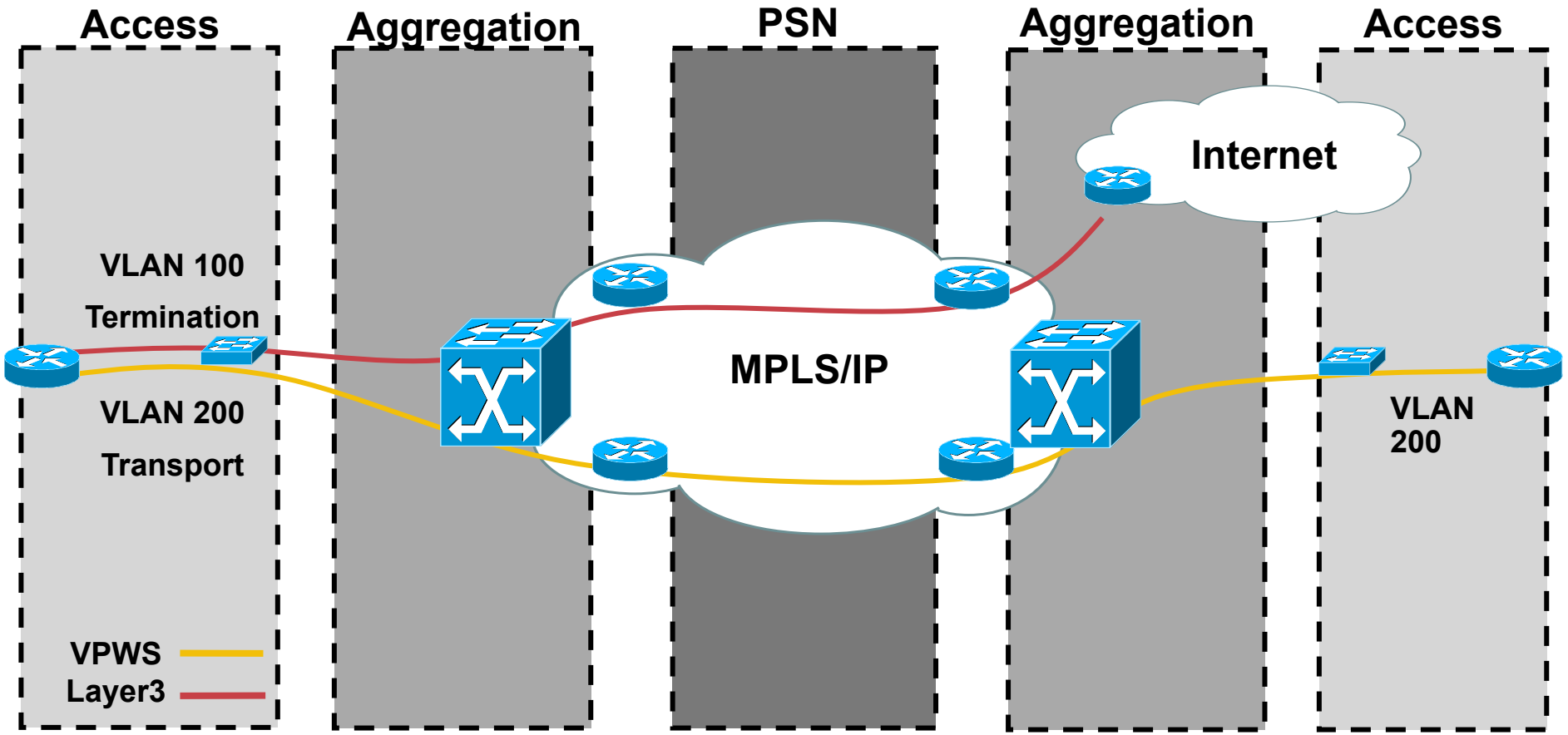
Motivation for L2VPNs:
Converged Networks Support



- Reduce overlapping core expense; consolidate trunk lines
- Offer multiservice/common interface (i.e. Ethernet MUX = L2, L3 and Internet)
- Maintain existing revenues from legacy services

© 2006 Cisco Systems, Inc. All rights reserved.

Motivation for L2VPNs:
The Ever Expanding Applications of Ethernet



- Fast becoming the access technology of choice
- Layer 2, Layer 3 and Internet Services on a common port
- Extends the reach of Metro Area Ethernet Networks

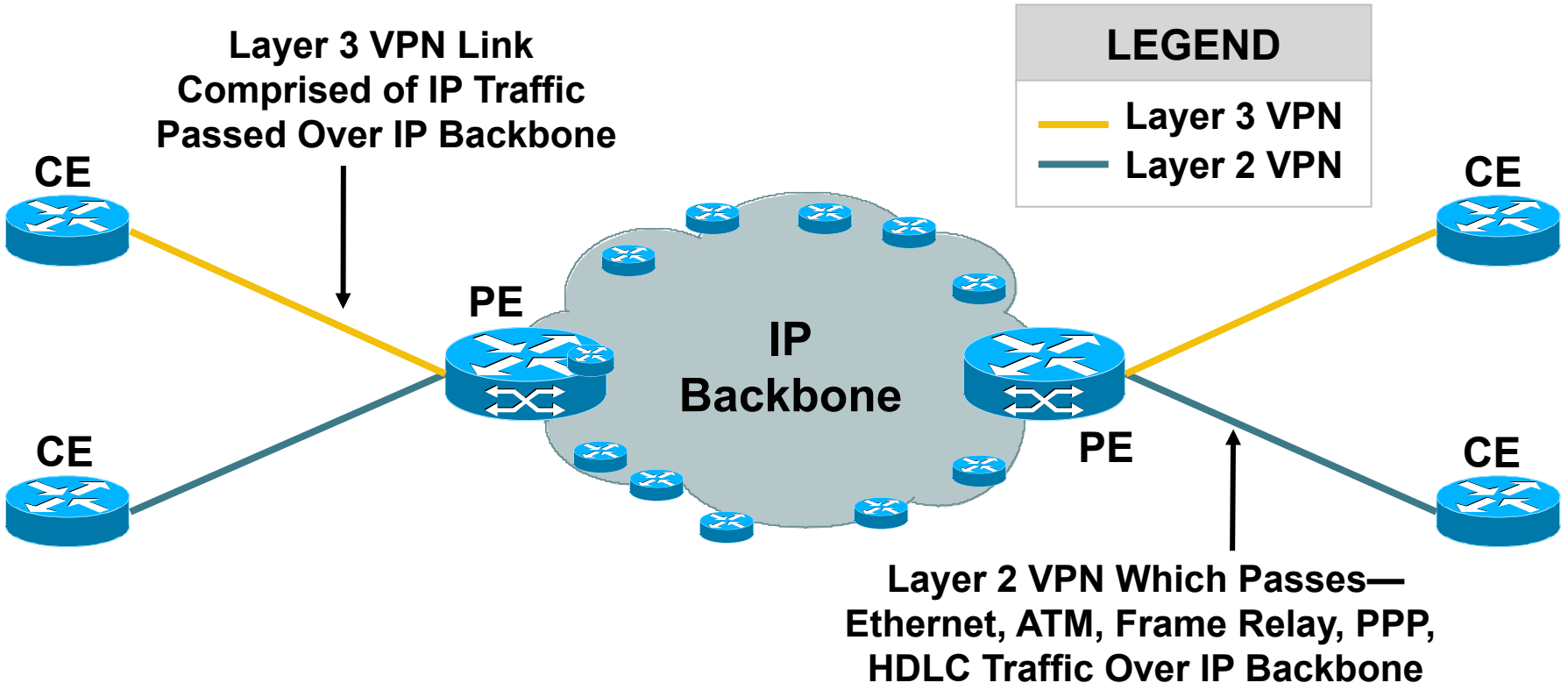
© 2006 Cisco Systems, Inc. All rights reserved.

Why is L2VPN needed?

- Allows SP to have a single infrastructure for both IP and legacy services
  - Migration
  - Provisioning is incremental
  - Network Consolidation
  - Capital and Operational savings
- Customer can have their own routing, qos policies, security mechanisms, etc
  - Layer 3 (IPv4, IPX, OSPF, BGP, etc …) on CE routers is transparent to MPLS core
  - CE1 router sees CE2 router as next-hop
  - No routing involved with MPLS core
- open architecture and vendor interoperability

© 2006 Cisco Systems, Inc. All rights reserved.

Introduction to Layer 2 and Layer 3 VPN Services



- **Layer 2 and Layer 3 VPN Services are offered from the edge of a network**

© 2006 Cisco Systems, Inc. All rights reserved.

Layer 3 and Layer 2 VPN Characteristics

Layer 3 VPNs

1. Packet-based forwarding e.g. IP
2. SP is involved
3. IP specific
4. Example: RFC 2547bis VPNs (L3 MPLS-VPN)

Layer 2 VPNs

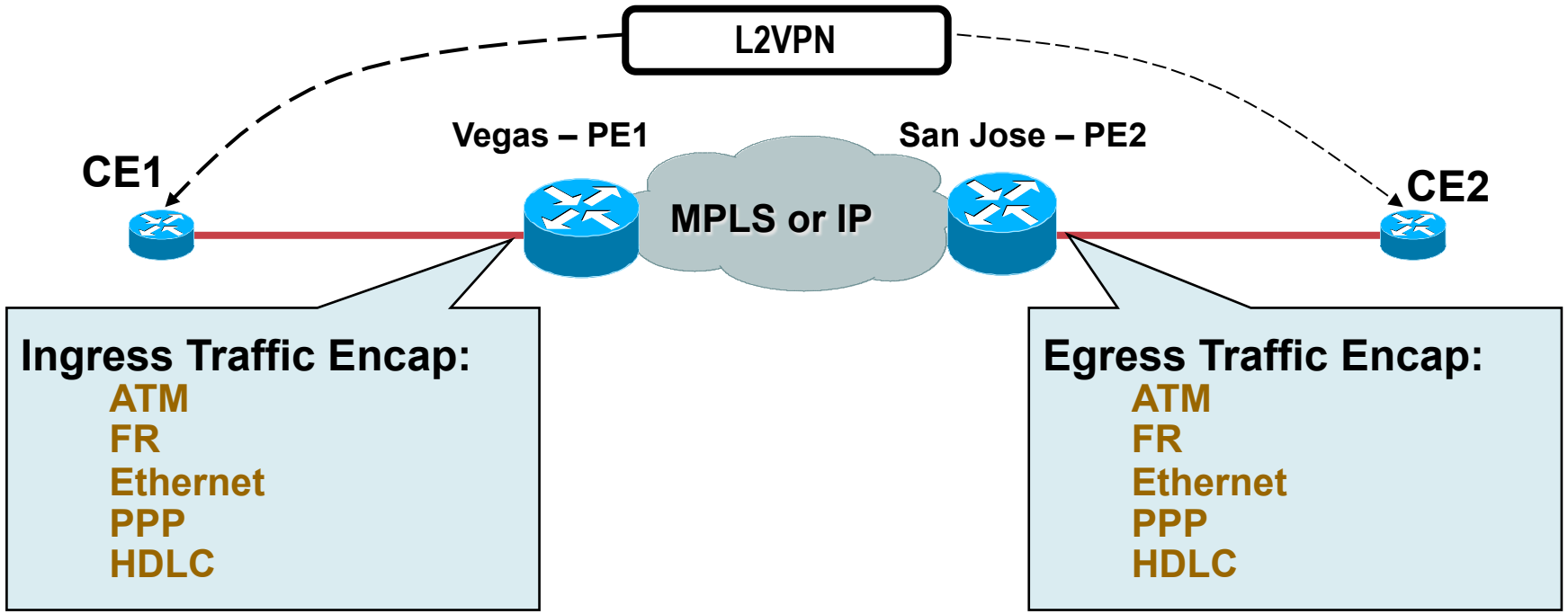
1. Frame-based forwarding e.g. DLCI, VLAN, VPI/VCI
2. No SP involvement
3. Multiprotocol support
4. Example: FR—ATM—Ethernet

The Choice of L2VPN over L3VPN Will Depend on How Much Control the Enterprise Wants to Retain

L2 VPN Services Are Complementary to L3 VPN Services

© 2006 Cisco Systems, Inc. All rights reserved.

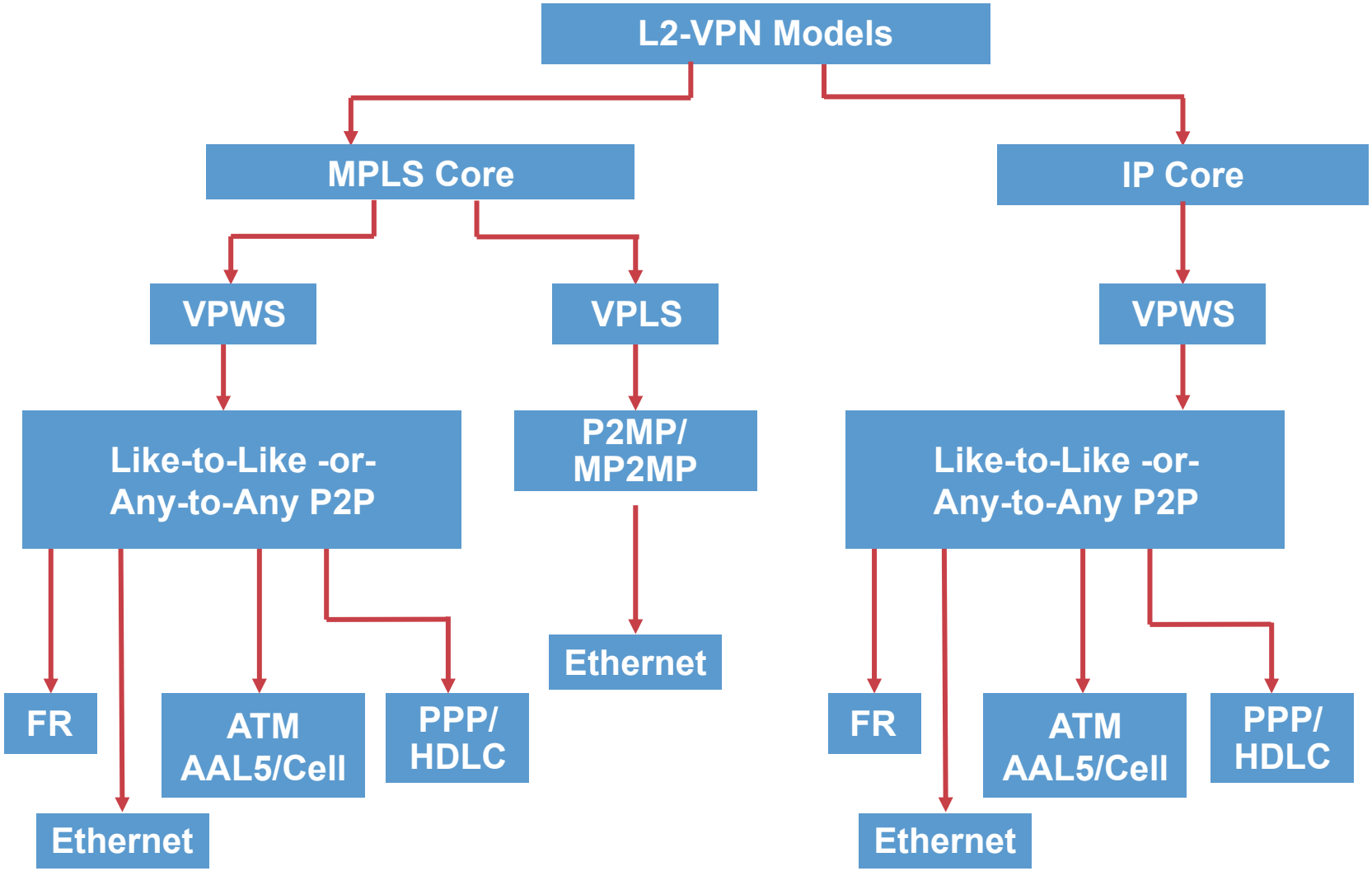
L2VPN - Simple definition



L2VPN provides an end-to-end layer 2 connection to an enterprise office in Vegas and San Jose over a SP's MPLS or IP core

© 2006 Cisco Systems, Inc. All rights reserved.

L2VPN Models



© 2006 Cisco Systems, Inc. All rights reserved.

Pseudowire—
IETF Technology Adoption

- **Virtual private wire service (VPWS) P2P**

  RFC3916 Pseudo Wire Emulation Edge-to-Edge (PWE3) Requirements

  RFC3985 Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture

  RFC 4447 Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)

  RFC4385 Pseudo wire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN

  RFC 4448 Encapsulation Methods for Transport of Ethernet over MPLS Networks

  draft-ietf-pwe3-[atm, frame-relay etc.]

- **Virtual private LAN services (VPLS) P2M**

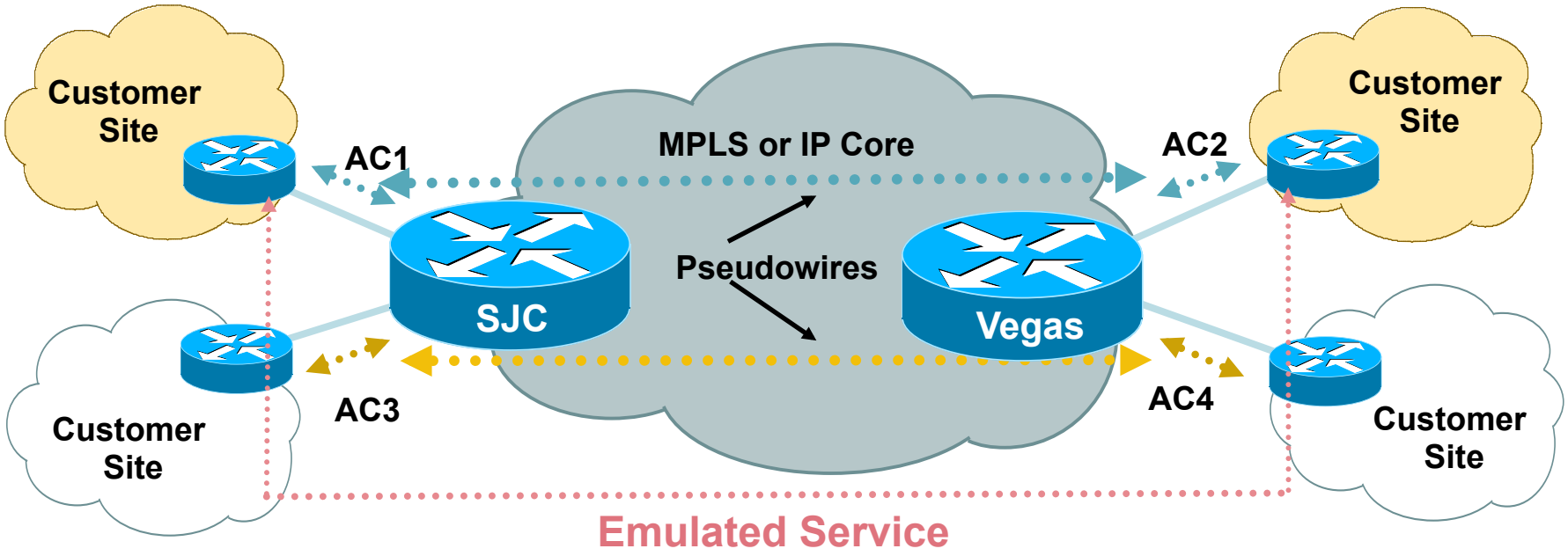
  draft-ietf-l2vpn-vpls-ldp-xx
  draft-ietf-l2vpn-vpls-bgp-xx

© 2006 Cisco Systems, Inc. All rights reserved.

- **Layer 2 Transport (VPWS)**
  - **L2TPv3**
    - **draft-ietf-l2tpext-l2tp-base-xx**
    - **draft-ietf-l2tpext-l2tpmib-base-xx**

© 2006 Cisco Systems, Inc. All rights reserved.

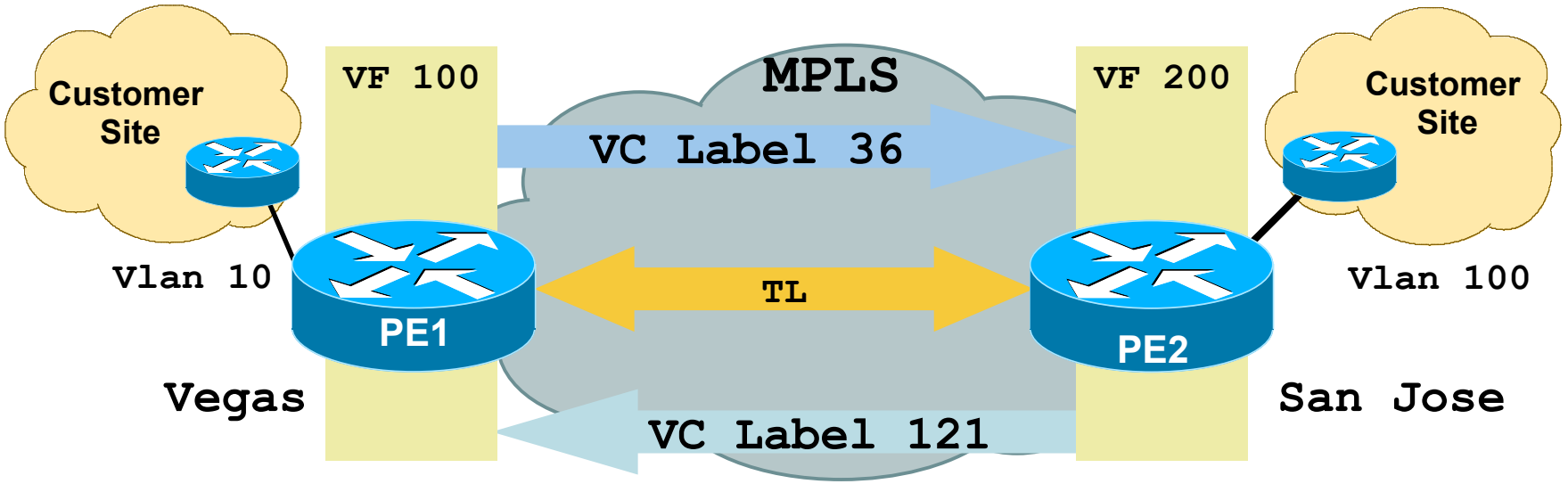
VPWS—Pseudowire Reference Model



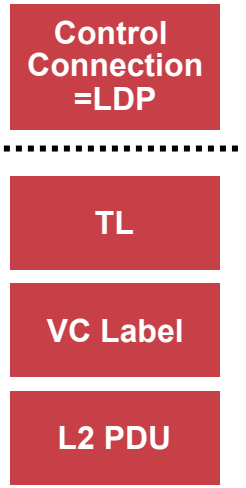
A Pseudowire (PW) Is a Connection Between Two Provider Edge (PE) Devices Which Connects Two Attachment Circuits (ACs)

© 2006 Cisco Systems, Inc. All rights reserved.

Building Blocks for L2VPNs—Data Plan Components—MPLS Core

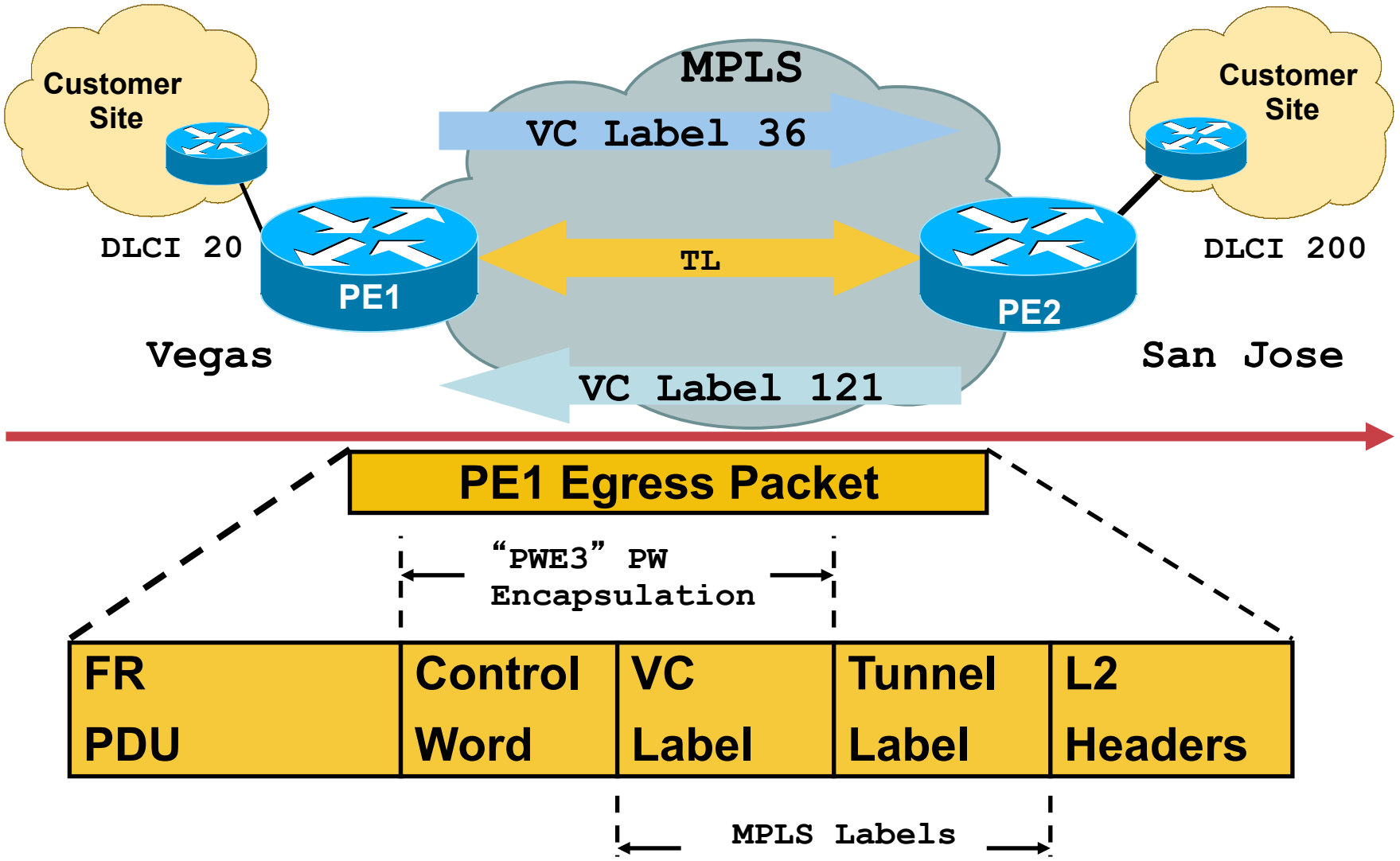


- **Virtual Forwarders (VF)—Subsystem that associates AC to PW**
- **Tunnel Label (TL)—Path between PE1 and PE2**
- **Pseudowire (PW)—Paths between VFs, a pair of unidirectional LSPs—VC label**
- **Attachment Circuits (AC)—L2 connection between CE and PE, i.e. VLAN, DLCI, ATM, etc.**



© 2006 Cisco Systems, Inc. All rights reserved.

Building Blocks for L2VPNs—Data Plan Components—FR Example

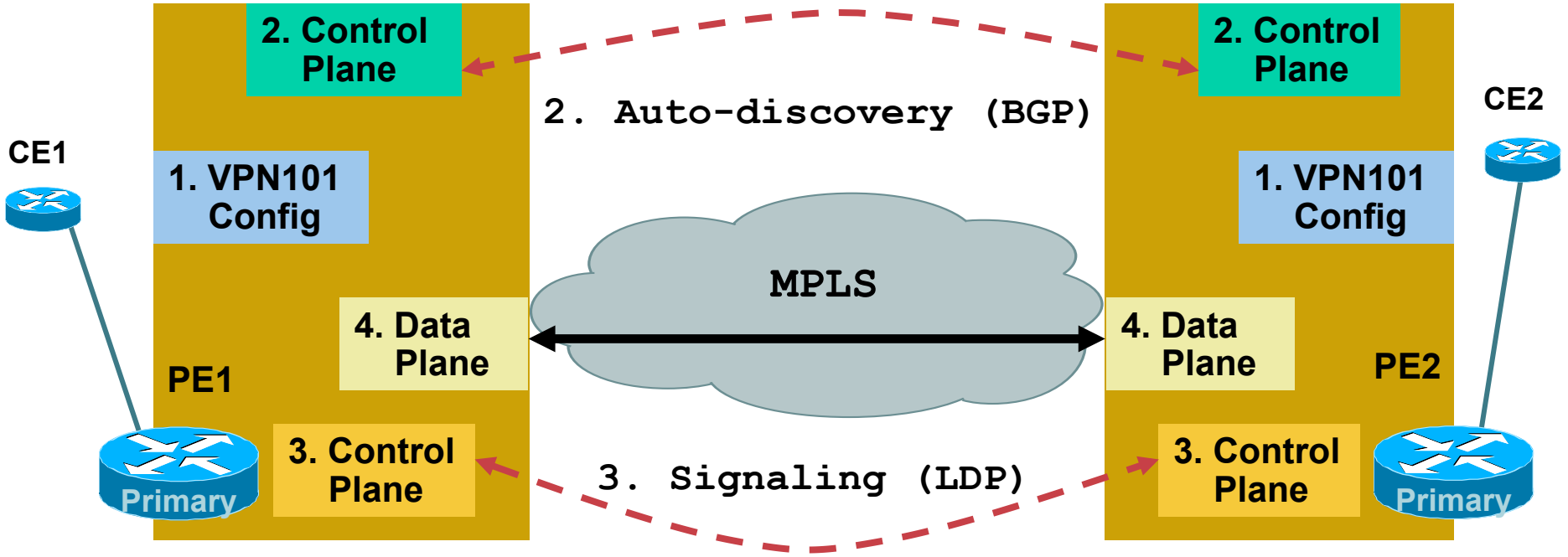


© 2006 Cisco Systems, Inc. All rights reserved.

PWE3 Signaling Concepts

© 2006 Cisco Systems, Inc. All rights reserved.

Building Blocks for L2VPNs—Control Plane



1. **Provision** — Config VPN
2. **Auto-discovery** — Advertise loopback and VPN members
3. **Signaling** — Setup pseudowire
4. **Data Plane** — Packet forwarding

© 2006 Cisco Systems, Inc. All rights reserved.

LDP Signaling Overview

Four Classes of LDP Messages:

1. **Peer discovery**
   - LDP link hello message
   - Targeted hello message

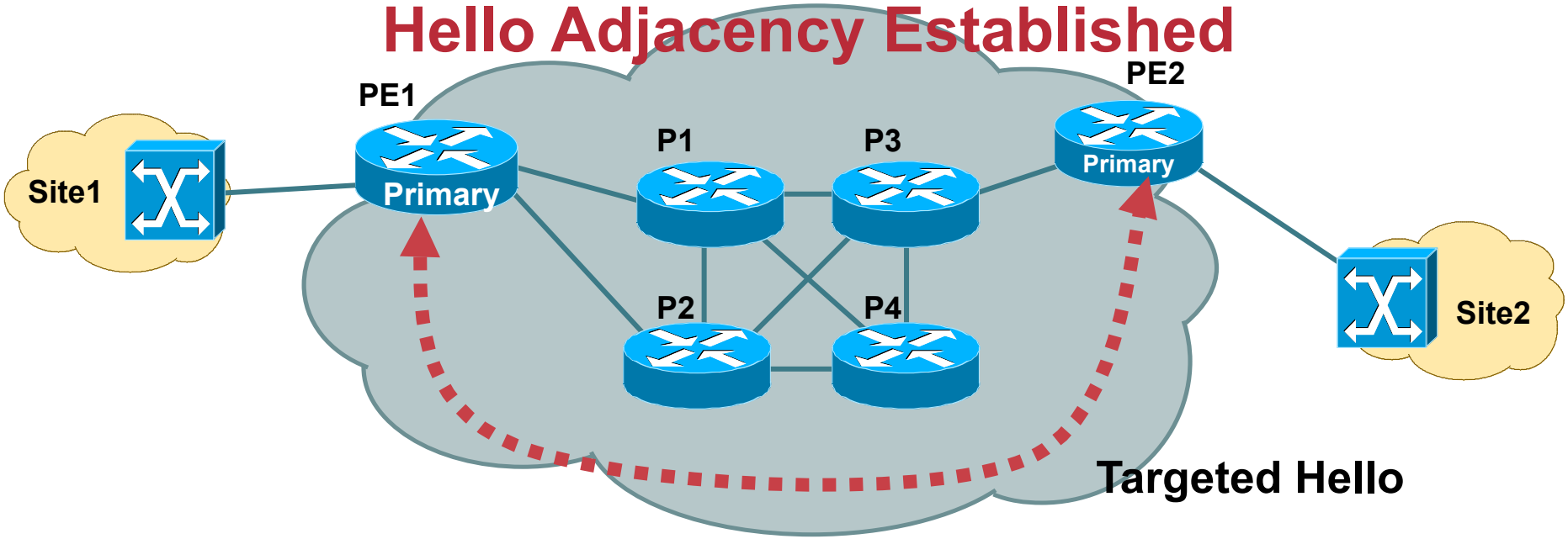
   } UDP

2. **LDP session**
   - LDP initialization and keepalive
   - Setup, maintain and disconnect LDP session
3. **Label advertisement**
   - Create, update and delete label mappings
4. **LDP notification**
   - Signal error or status info

} TCP

© 2006 Cisco Systems, Inc. All rights reserved.

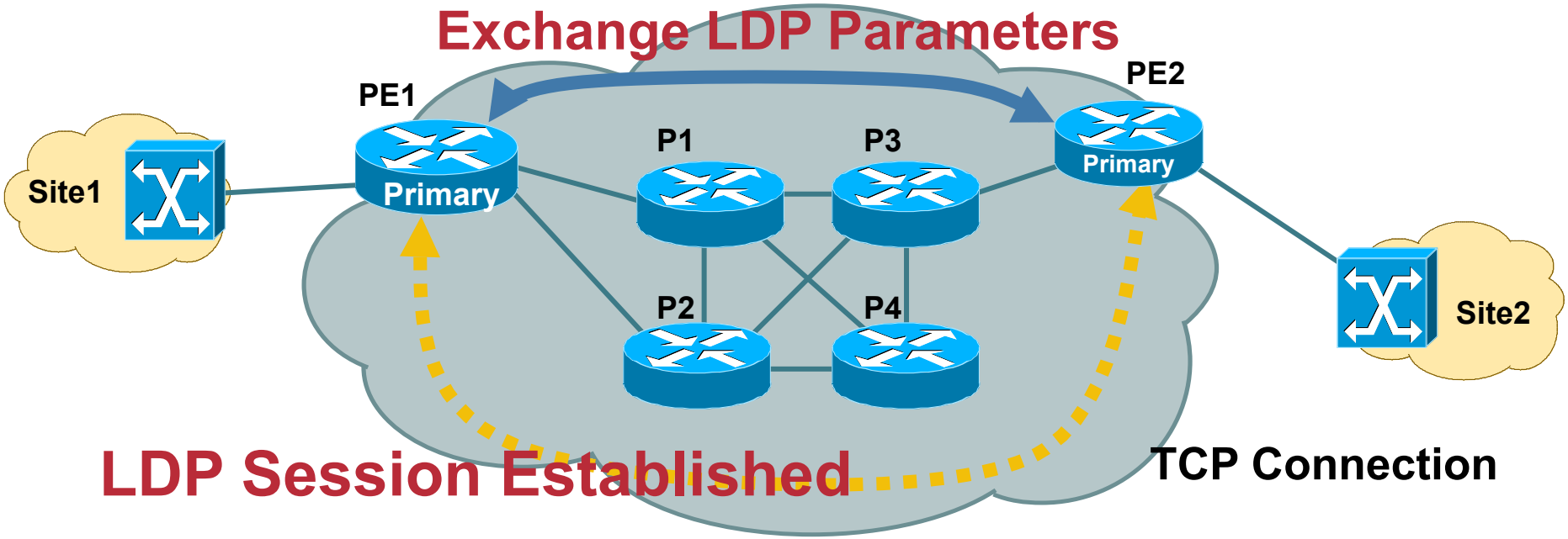
L2VPN LDP Extended Discovery



- Targeted Hello Messages Are Exchanged as UDP Packets on Port 646 Consisting of router-id and label space

© 2006 Cisco Systems, Inc. All rights reserved.

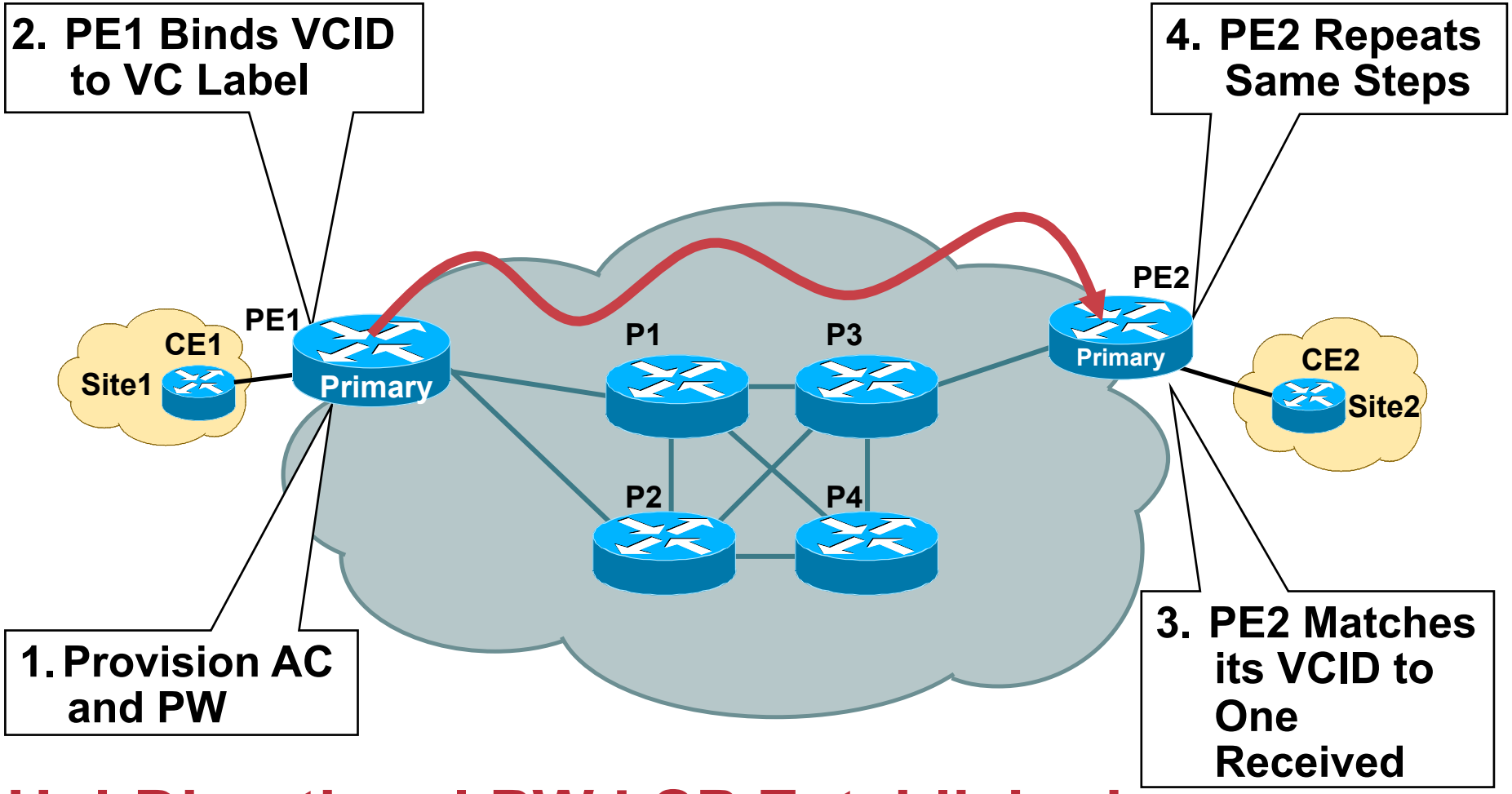
L2VPN LDP Session Establishment



- Active role PE—establishes TCP connection using port 646
- LDP peers exchange and negotiate session parameters such as the protocol version, label distribution methods, timer values, label ranges, and so on
- LDP session is operational

© 2006 Cisco Systems, Inc. All rights reserved.

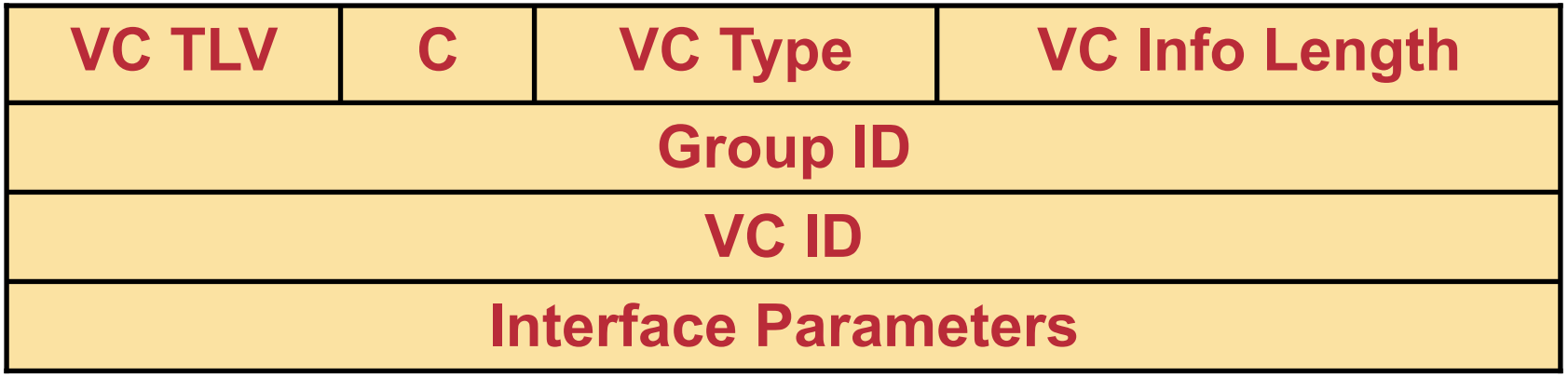
L2VPN—Pseudowire Label Binding



Uni-Directional PW LSP Established

© 2006 Cisco Systems, Inc. All rights reserved.

New VC FEC Element



Virtual Circuit FEC Element

- C—control word present
- VC Type—ATM, FR, Ethernet, HDLC, PPP, etc.
- VC Info Length—length of VCID
- Group ID—group of VCs referenced by index (user configured)
- VC ID—used to identify Virtual Circuit
- Interface Parameters—MTU, etc.

© 2006 Cisco Systems, Inc. All rights reserved.

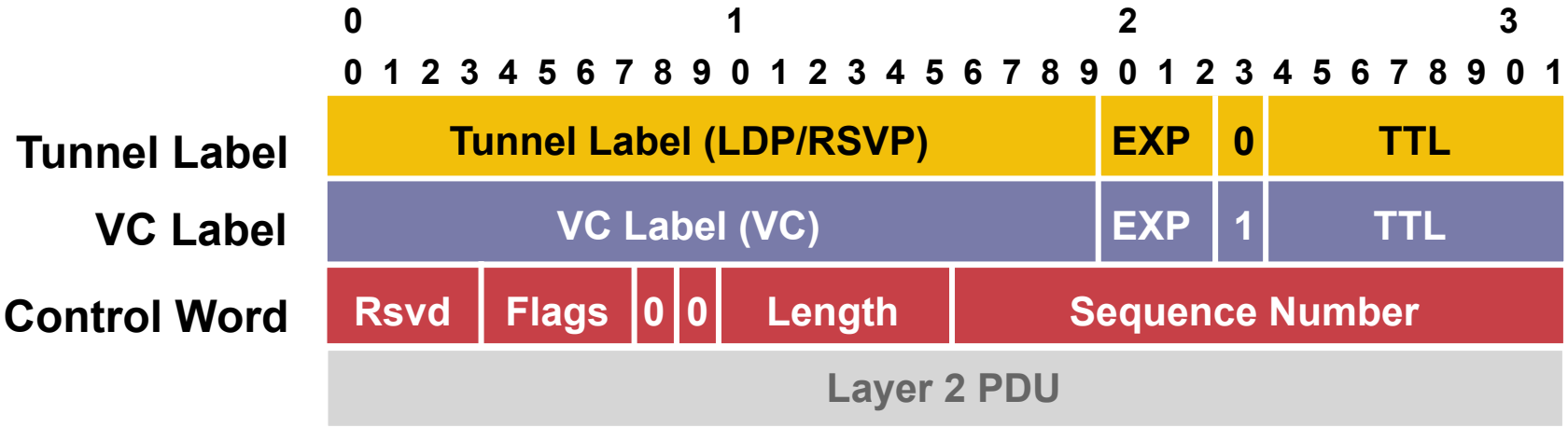
Pseudowire VC Type

Some Widely Deployed VC Types

PW Type	Description
0x0001	Frame Relay DLCI
0x0002	ATM AAL5 SDU VCC transport
0x0003	ATM transparent cell transport
0x0004	Ethernet Tagged Mode (VLAN)
0x0005	Ethernet
0x0006	HDLC
0x0007	PPP

© 2006 Cisco Systems, Inc. All rights reserved.

L2VPNs—Label Stacking



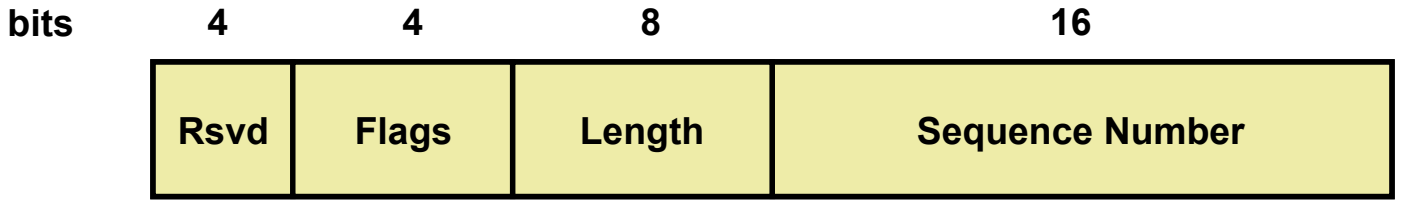
Three Layers of Encapsulation

- Tunnel label—determines path through network
- VC label—identifies VC at endpoint
- Control word—contains attributes of L2 payload (optional)

© 2006 Cisco Systems, Inc. All rights reserved.

Generic Control Word—
VC Information Fields

Control Word



- Use of control word is optional
- Flags—carries "flag" bits depending on encapsulation

  (FR;FECN, BECN, C/R, DE, ATM;CLP, EFCI, C/R, etc)
- Length—required for padding small frames when < interface MTU
- Sequence number—used to detect out of order delivery of frames

Control Word	
Encap.	Required
CR	No
AAL5	Yes
Eth	No
FR	Yes
HDLC	No
PPP	No

© 2006 Cisco Systems, Inc. All rights reserved.

VPWS Transport

© 2006 Cisco Systems, Inc. All rights reserved.

VPWS Transports—Encapsulations

- **Ethernet/802.1Q VLAN (EoMPLS)**

  RFC 4448 Encapsulation Methods for Transport of Ethernet over MPLS Networks

- **Frame Relay (FRoMPLS)**

  draft-ietf-pwe3-frame-relay-encap-xx.txt

- **ATM AAL5 and ATM Cell (ATMoMPLS)**

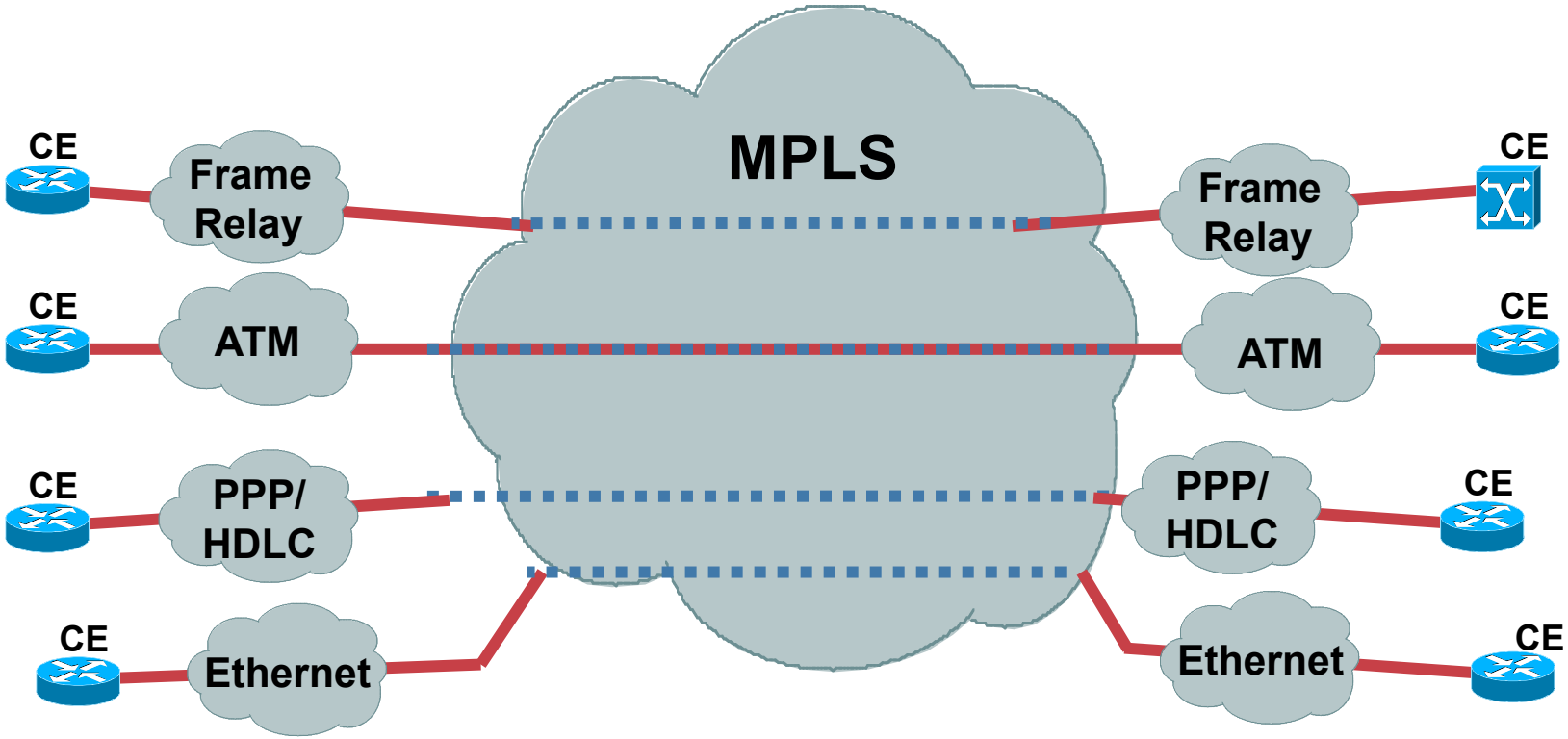
  draft-ietf-pwe3-atm-encap-xx.txt

- **PPP/HDLC (PPPoMPLS/HDLCoMPLS)**

  draft-ietf-pwe3-hdlc-ppp-encap-mpls-xx.txt

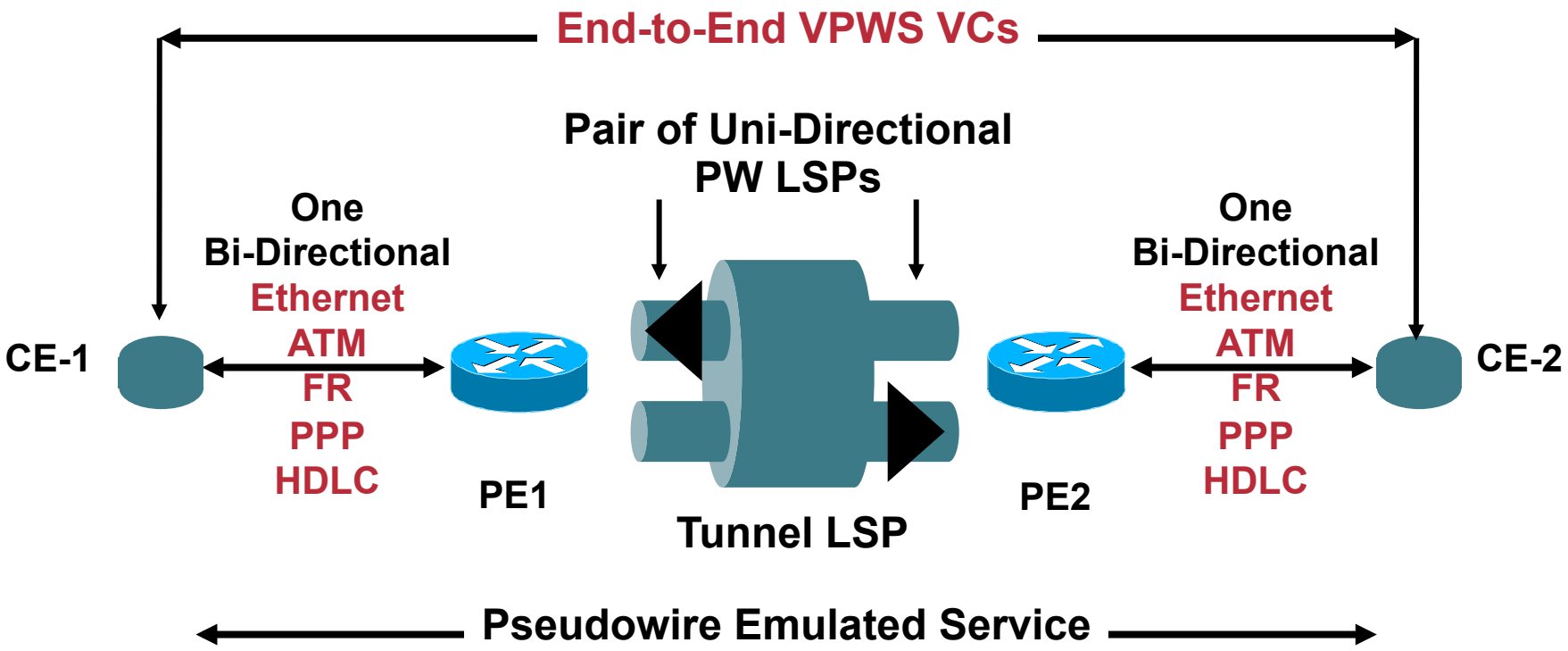
© 2006 Cisco Systems, Inc. All rights reserved.

VPWS Transports



© 2006 Cisco Systems, Inc. All rights reserved.

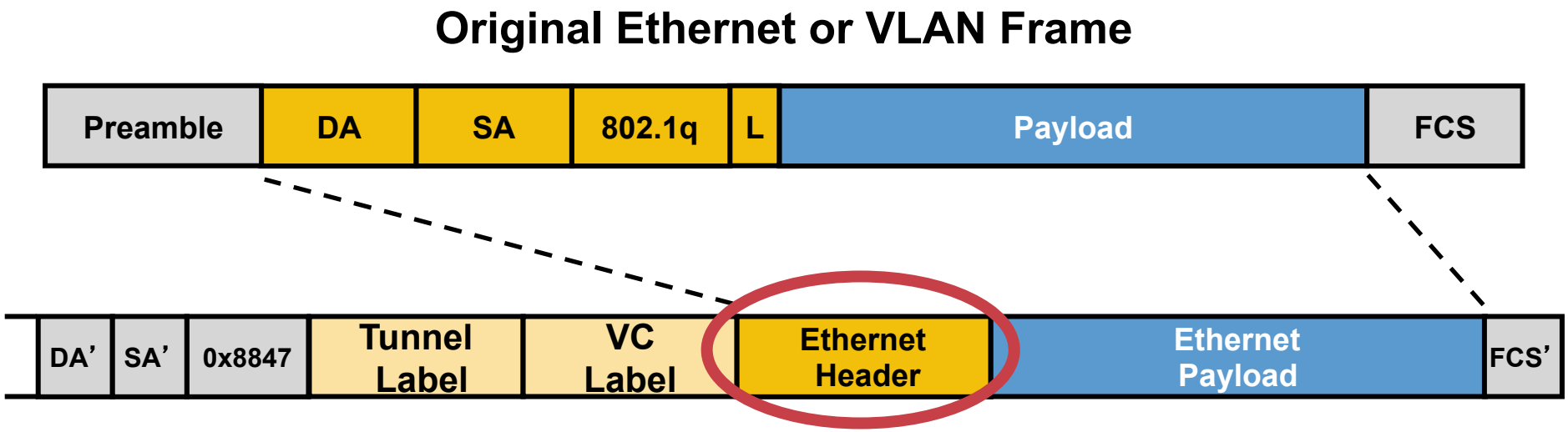
VPWS Transports Service—Reference Model



- Pseudowire transport (across PEs) applications
- Local switching (within a PE) applications

© 2006 Cisco Systems, Inc. All rights reserved.

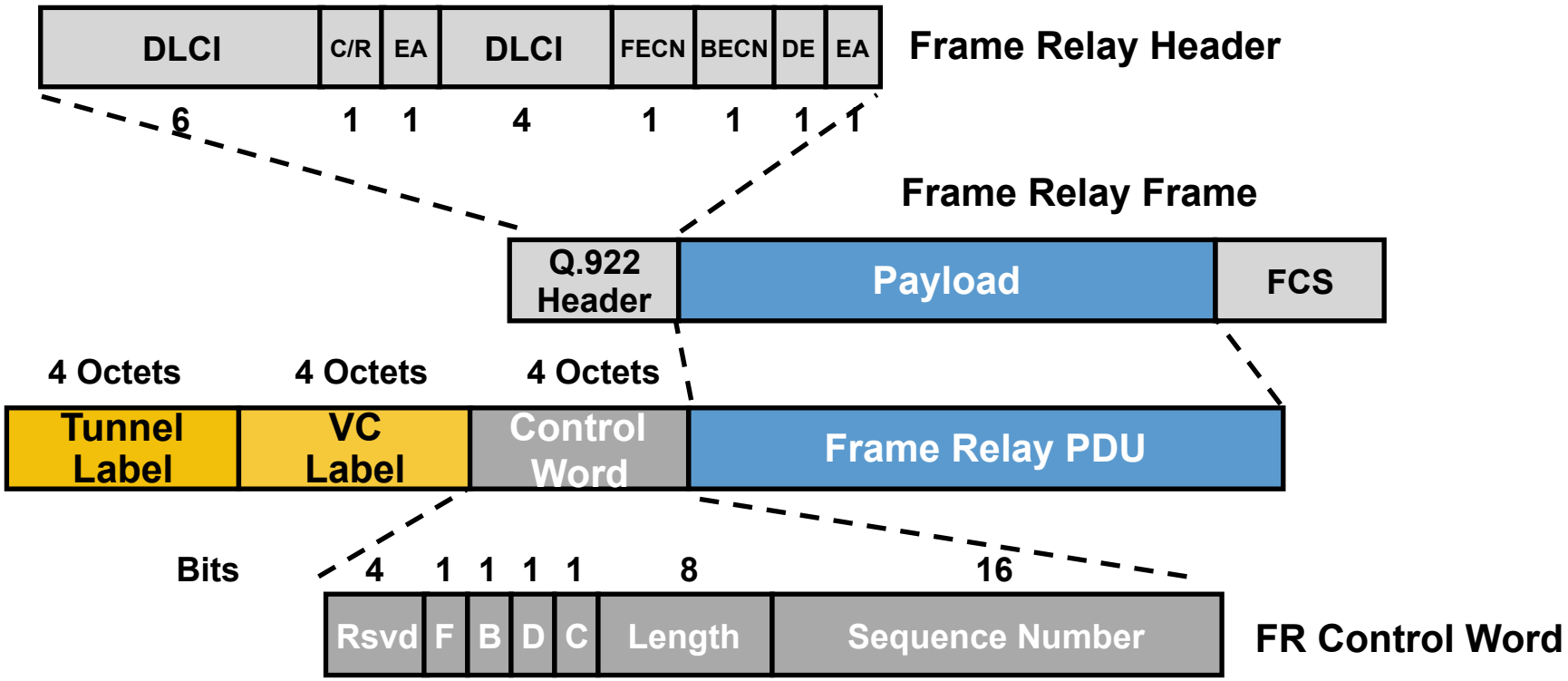
VPWS EoMPLS—
RFC 4448



- VC type-0x0004 is used for VLAN over MPLS application
- VC type-0x0005 is used for Ethernet port tunneling application (port transparency)

© 2006 Cisco Systems, Inc. All rights reserved.

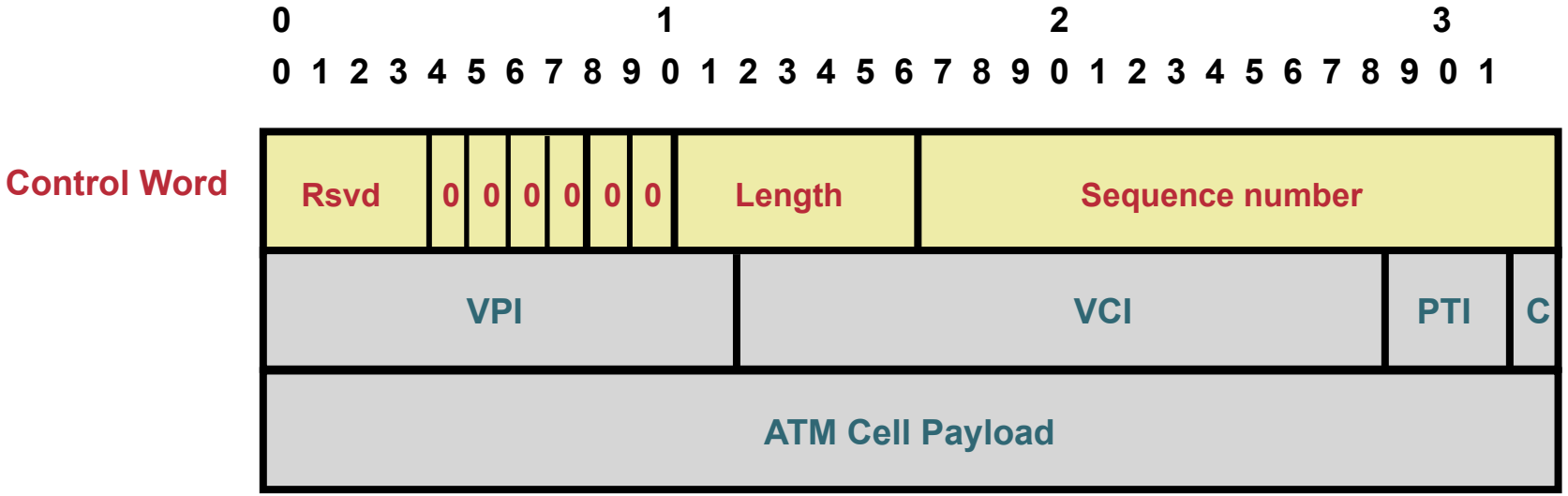
VPWS FRoMPLS—
draft-ietf-pwe3-frame-relay-encap-xx.txt



- F = FECN (Forward Explicit Congestion Notification)
- B = BECN (Backward Explicit Congestion Notification)
- D = DE (Discard Eligibility Indicator)
- C = C/R (Command/Response Field)

© 2006 Cisco Systems, Inc. All rights reserved.

VPWS CRoMPLS—
draft-ietf-pwe3-atm-encap-xx.txt

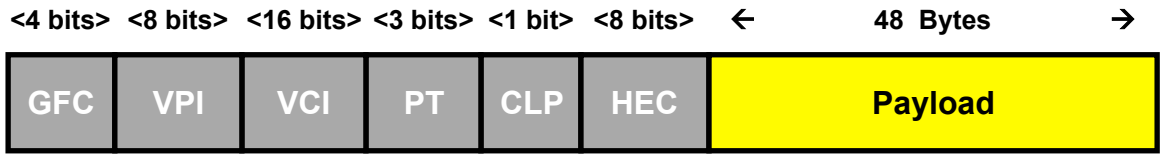


- This is cell relay over MPLS (VC/VP/port mode)
- Single cell is encapsulated; no HEC (52 bytes only)
- Control word is optional
- Control word flags should be set to zero and ignored

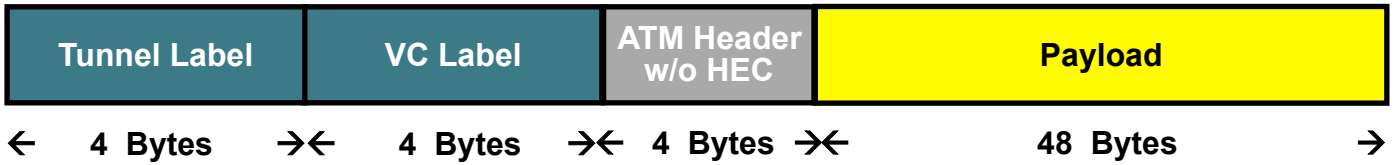
© 2006 Cisco Systems, Inc. All rights reserved.

VPWS CPKoMPLS—Encapsulation
draft-ietf-pwe3-atm-encap-xx.txt

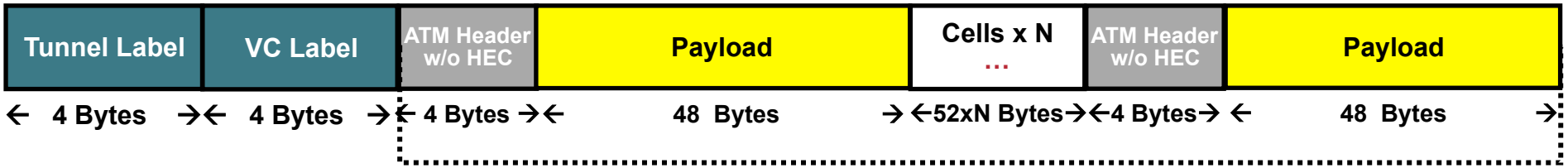
ATM Cell



Single Cell Relay



Packed Cell Relay



Packed Cells Max 28

28*52=1456 Bytes

© 2006 Cisco Systems, Inc. All rights reserved.

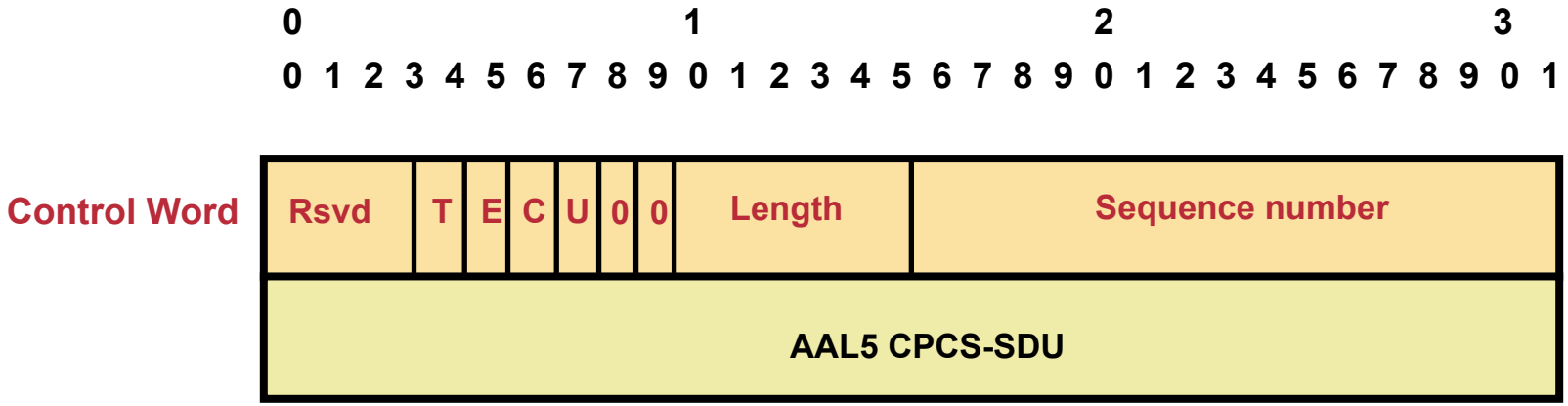
VPWS CPKoMPLS—
draft-ietf-pwe3-atm-encap-xx.txt

CPKoMPLS = Cell Packing over MPLS

- **Used to mitigate cell to MPLS packet MTU inefficiencies**
- **Concatenated ATM cell (52 bytes); no HEC**
- **Maximum 28 cells per MPLS frame (<1500 byte MTU)**
- **VC/VP/port mode support**
- **Cell Packing operation:**
  - **-Maximum Number of Cells to Pack (MNCP)**
  - **-Minimum Cell Packing Timer (MCPT)**

© 2006 Cisco Systems, Inc. All rights reserved.

VPWS AAL5oMPLS—
draft-ietf-pwe3-atm-encap-xx.txt

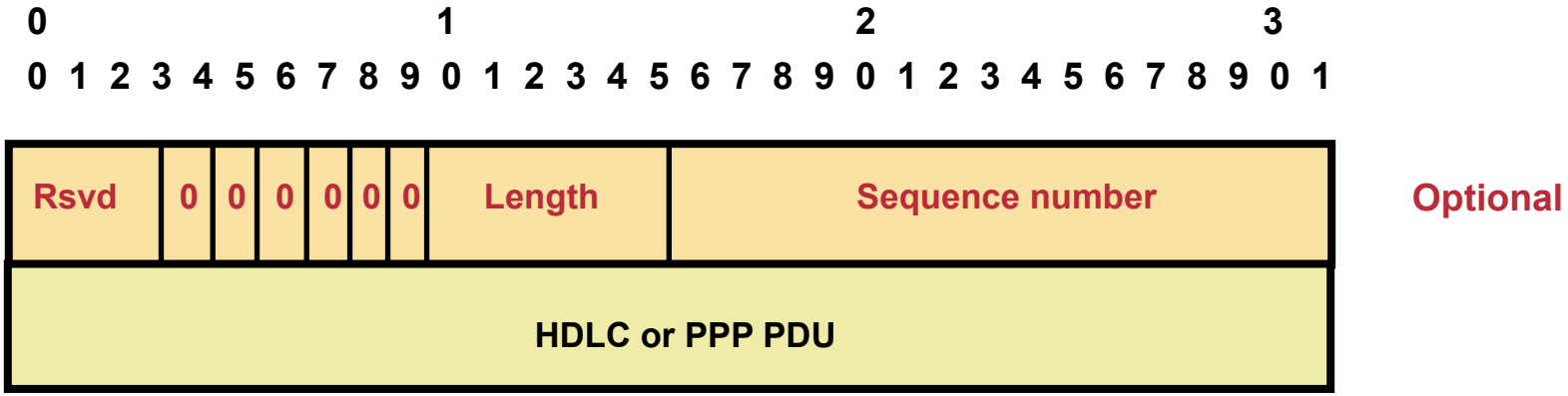


- AAL5 SDU is encapsulated
- Control word is required
- Service allows transport of OAM and Resource Management cells
- Control word flags encapsulate transport type, EFCI, CLP, C/R bit

© 2006 Cisco Systems, Inc. All rights reserved.

VPWS PPPoMPLS/HDLCoMPLS—
draft-ietf-pwe3-hdlc-ppp-encap-xx.txt

- **Cisco HDLC and PPP PDUs are transported without flags or FCS**
  - **PPP frames also do not carry HDLC address and control information**
- **The control word is optional**



© 2006 Cisco Systems, Inc. All rights reserved.

Frame Format CE — LER

Original Ethernet Frame

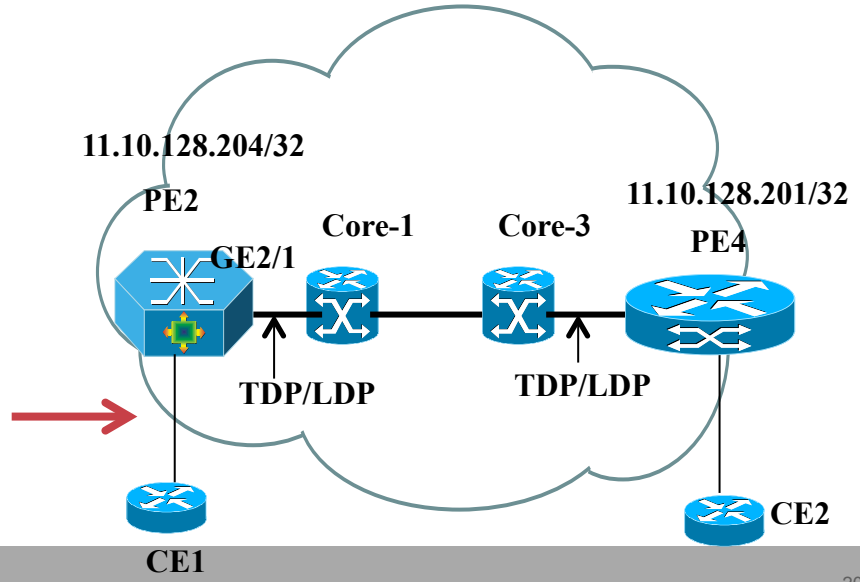
DA SA	8000 V HL TOS …

VLAN Encapsulated Frame

DA SA	8100 Pbits Cbit VLAN ID	Ethernet Frame

4 Byte 802.1q Header

- **2 Byte EtherType Field (8100)**
- **3 P bits**
- **C bit**
- **12 bit VID**



© 2006 Cisco Systems, Inc. All rights reserved.

Frame Format LER—LSR

VLAN Encapsulated Frame

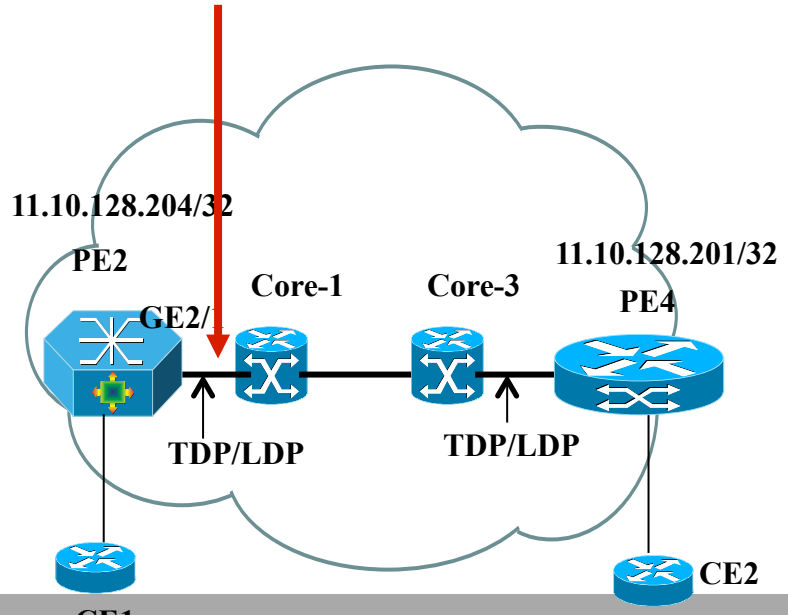
DA SA	8100 Pbits Cbit VLAN ID	Ethernet Frame

MPLS Labeled Packet

DA SA 8847 MPLS LSEs	DA SA	8100 Pbits Cbit VLAN ID	Ethernet Frame

LSE (Label Stack Entries)

- **20 Bit Label**
- **3 Bit Experimental Field (Exp)**
- **1 Bit Bottom of Stack Indicator (S)**
- **1 Byte TTL**



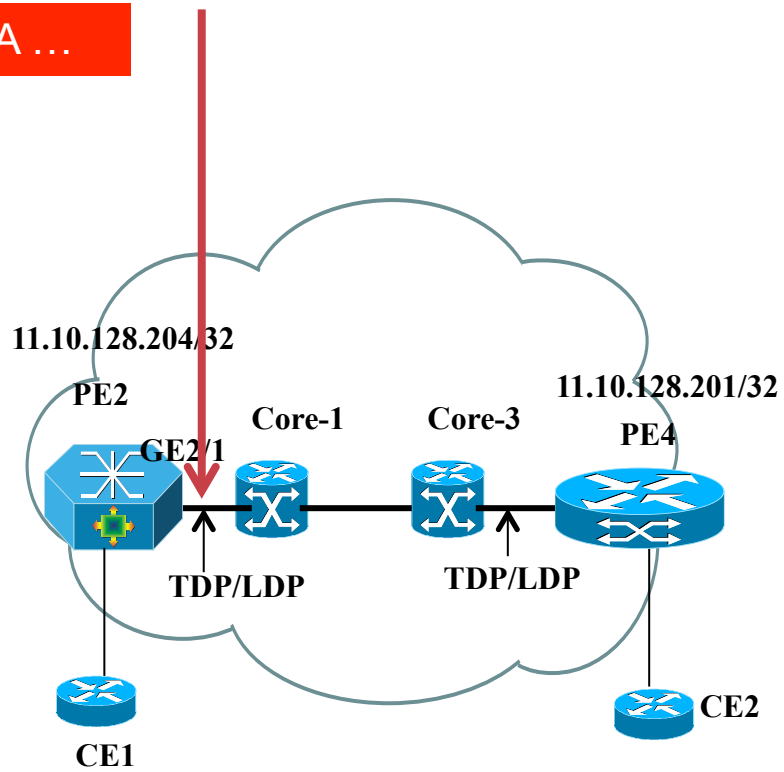
© 2006 Cisco Systems, Inc. All rights reserved.

Frame Format LER—LSR (Cont.)

MPLS Labeled Packet

DA SA 8847	00037 0 FE	00012 1 02	DA SA …

- **Tunnel Label Entry**
  - **Label 55 (37)**
  - **Exp = 0**
  - **S = 0**
  - **TTL = FE**
- **VC Label**
  - **Label 18 (12)**
  - **Exp = 0**
  - **S=1**
  - **TTL = 02**



Detaled packet header explanation at:
http://www-tac.cisco.com/Teams/NSA/MPLS/EOMPLS/pac1.htm

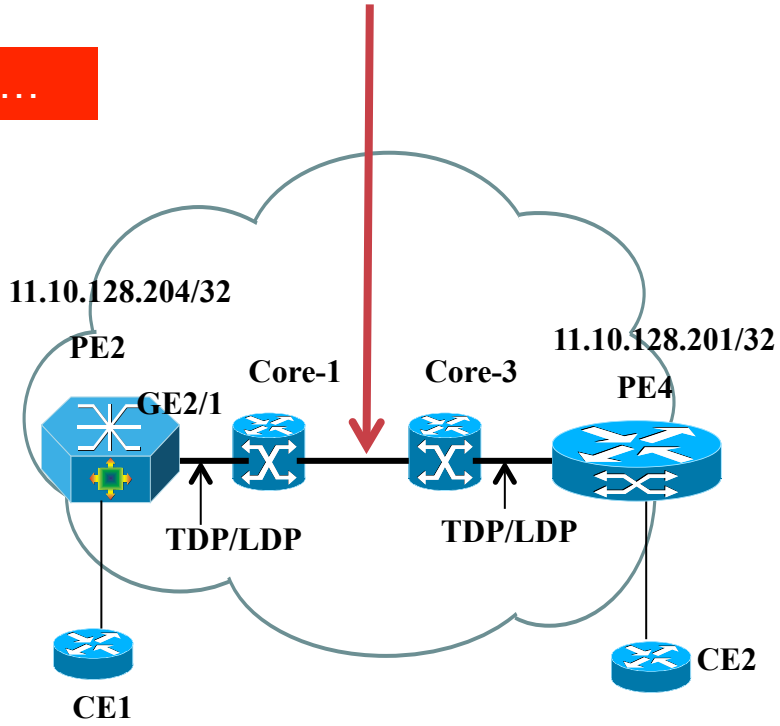
© 2006 Cisco Systems, Inc. All rights reserved.

Frame Format LSR—LSR

MPLS Labeled Packet

DA SA 8847	00088 0 FD	00012 1 02	DA SA …

- **Tunnel Label Entry**
  - **Label 136 (88)**
  - **Exp/S = 0**
  - **TTL = FD**
- **VC Label**
  - **Label 18 (12)**
  - **Exp/S = 1**
  - **TTL = 02**



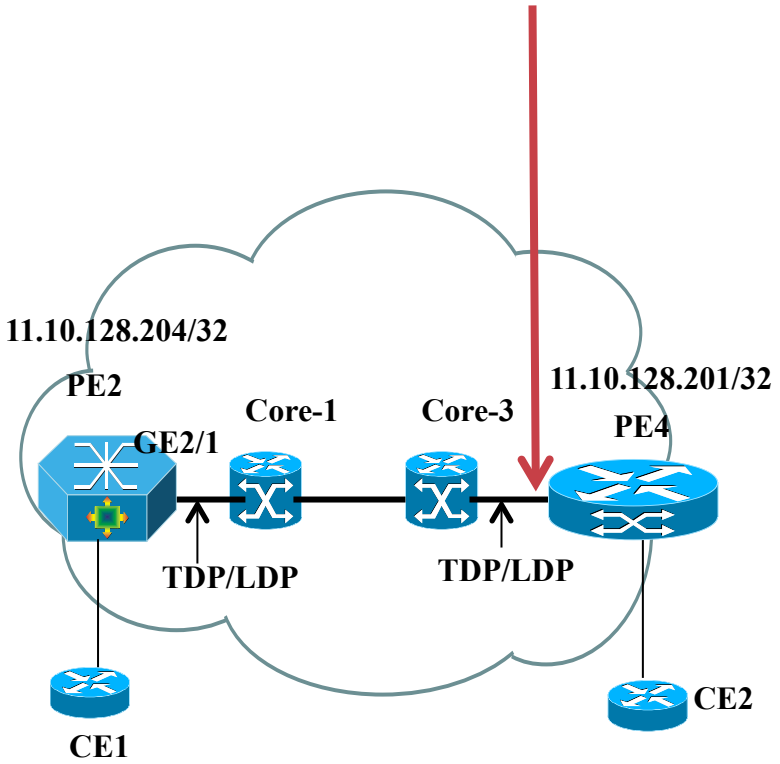
© 2006 Cisco Systems, Inc. All rights reserved.

Frame Format LSR—LER

MPLS Labeled Packet

DA SA 8847	00012 1 01	DA SA …

- **VC Label**
  - **Label 18 (12)**
  - **Exp/S = 1**
  - **TTL = 01**

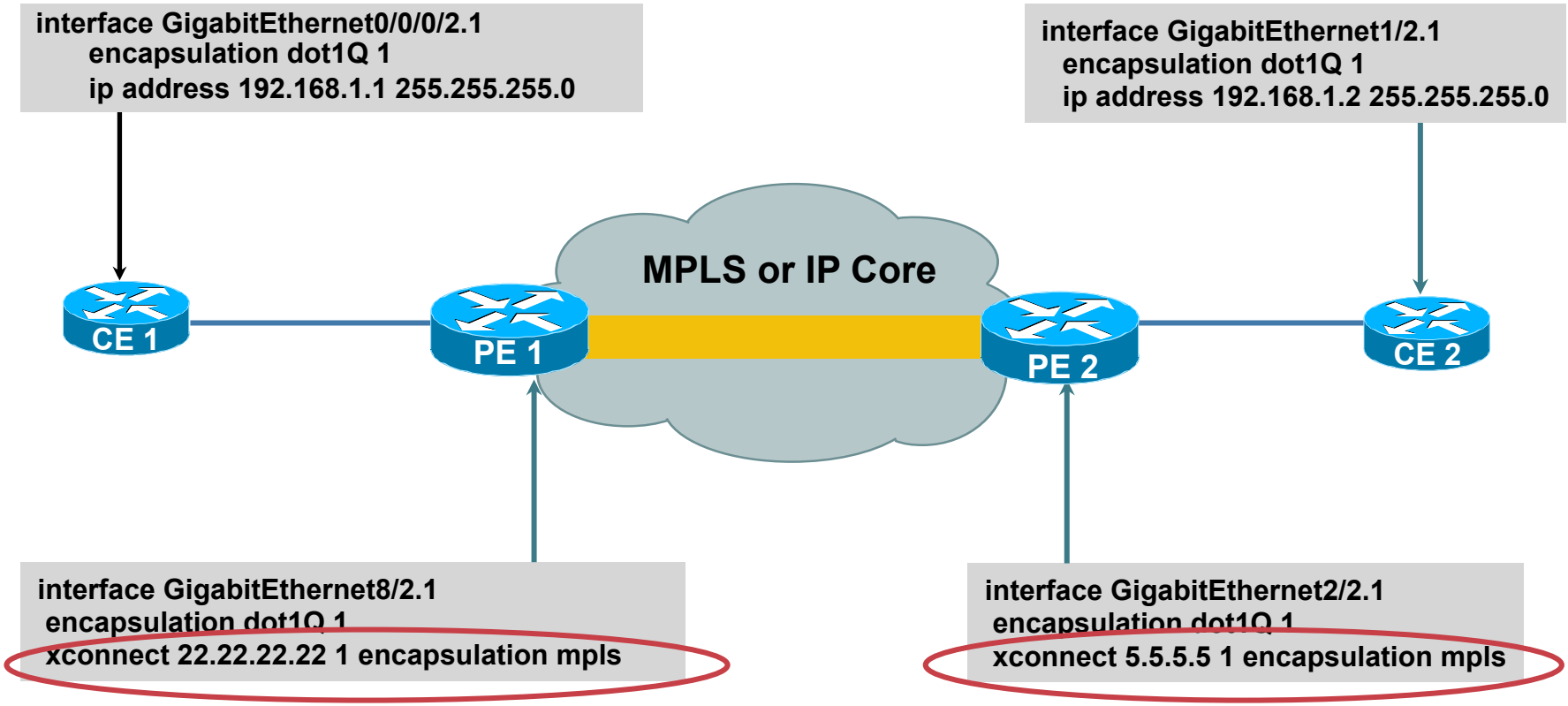


© 2006 Cisco Systems, Inc. All rights reserved.

Example: VPWS

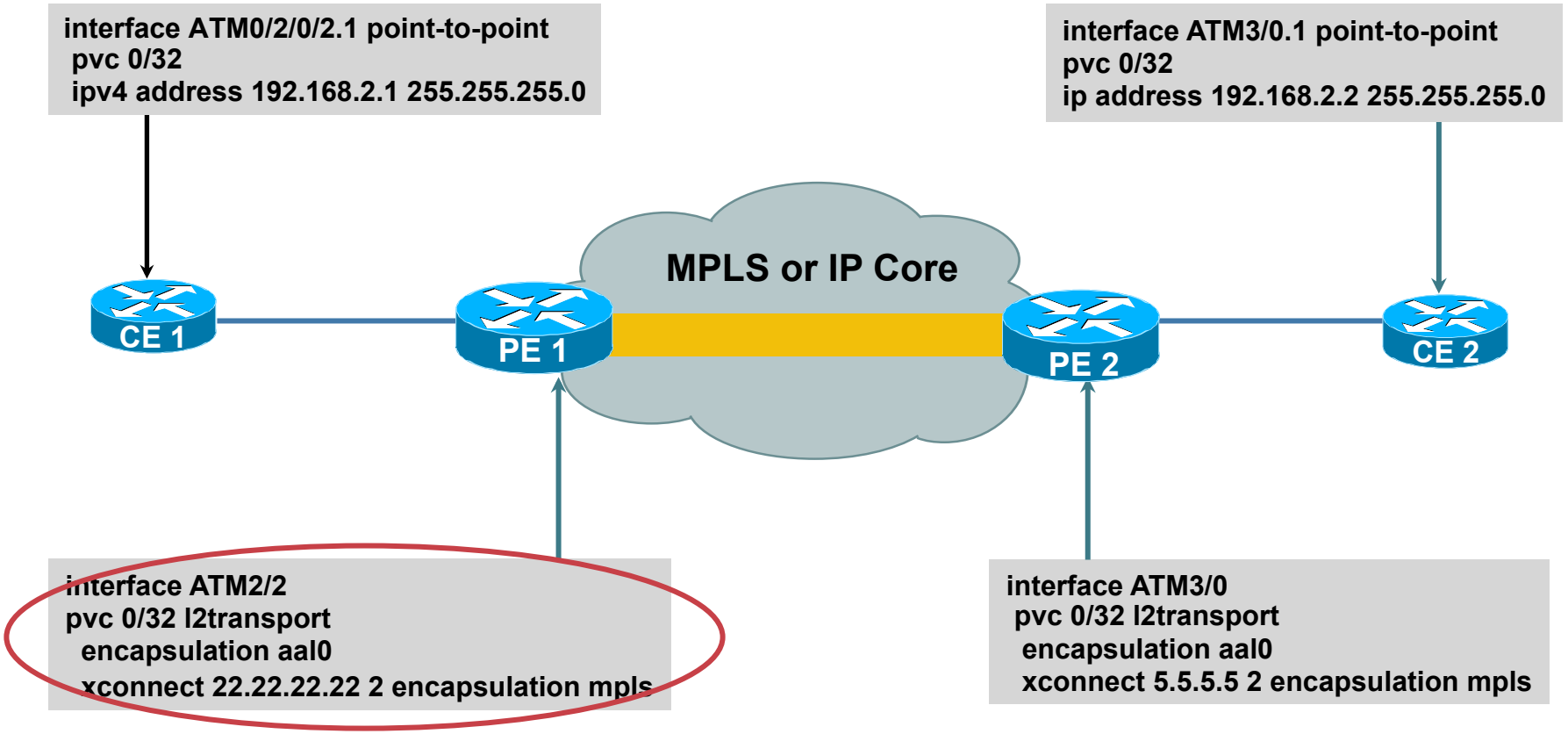
© 2006 Cisco Systems, Inc. All rights reserved.

Point-to-Point VLAN over MPLS



© 2006 Cisco Systems, Inc. All rights reserved.

Point-to-Point Cell Relay over MPLS

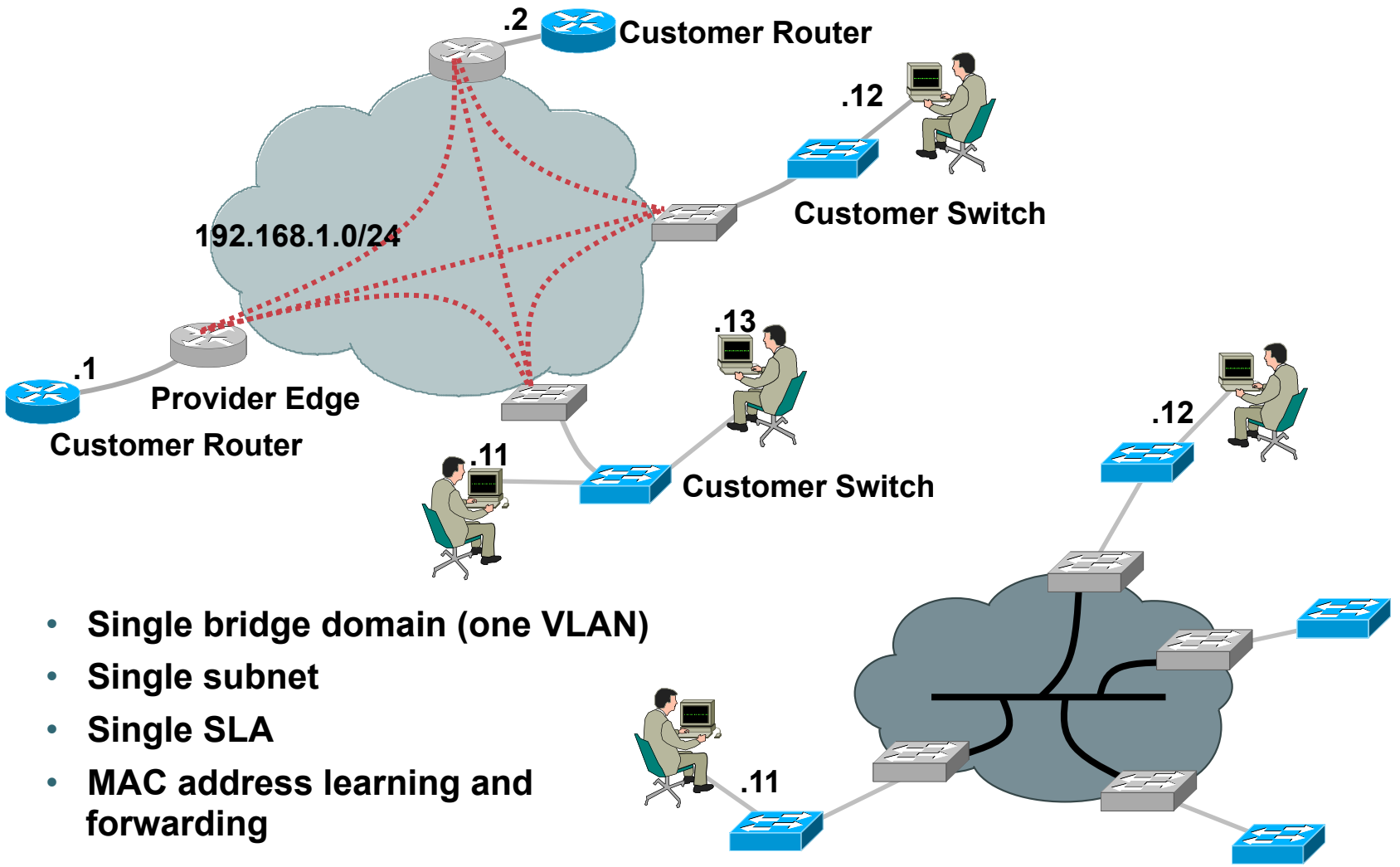


© 2006 Cisco Systems, Inc. All rights reserved.

Virtual Private LAN Service (VPLS)

© 2006 Cisco Systems, Inc. All rights reserved.

VPLS: Customer View



- Single bridge domain (one VLAN)
- Single subnet
- Single SLA
- MAC address learning and forwarding

© 2006 Cisco Systems, Inc. All rights reserved.

VPLS—Overview

- **Architecture**

  It is an end-to-end architecture that allows IP/MPLS networks to provide Layer 2 multipoint Ethernet services while using LDP as signaling protocol

- **Bridge emulation**

  Emulates an Ethernet bridge

- **Bridge functions**

  Operation is the same as for an Ethernet bridge, i.e. forwards using the destination MAC address, learns source addresses and floods broad-/multicast and unknown frames

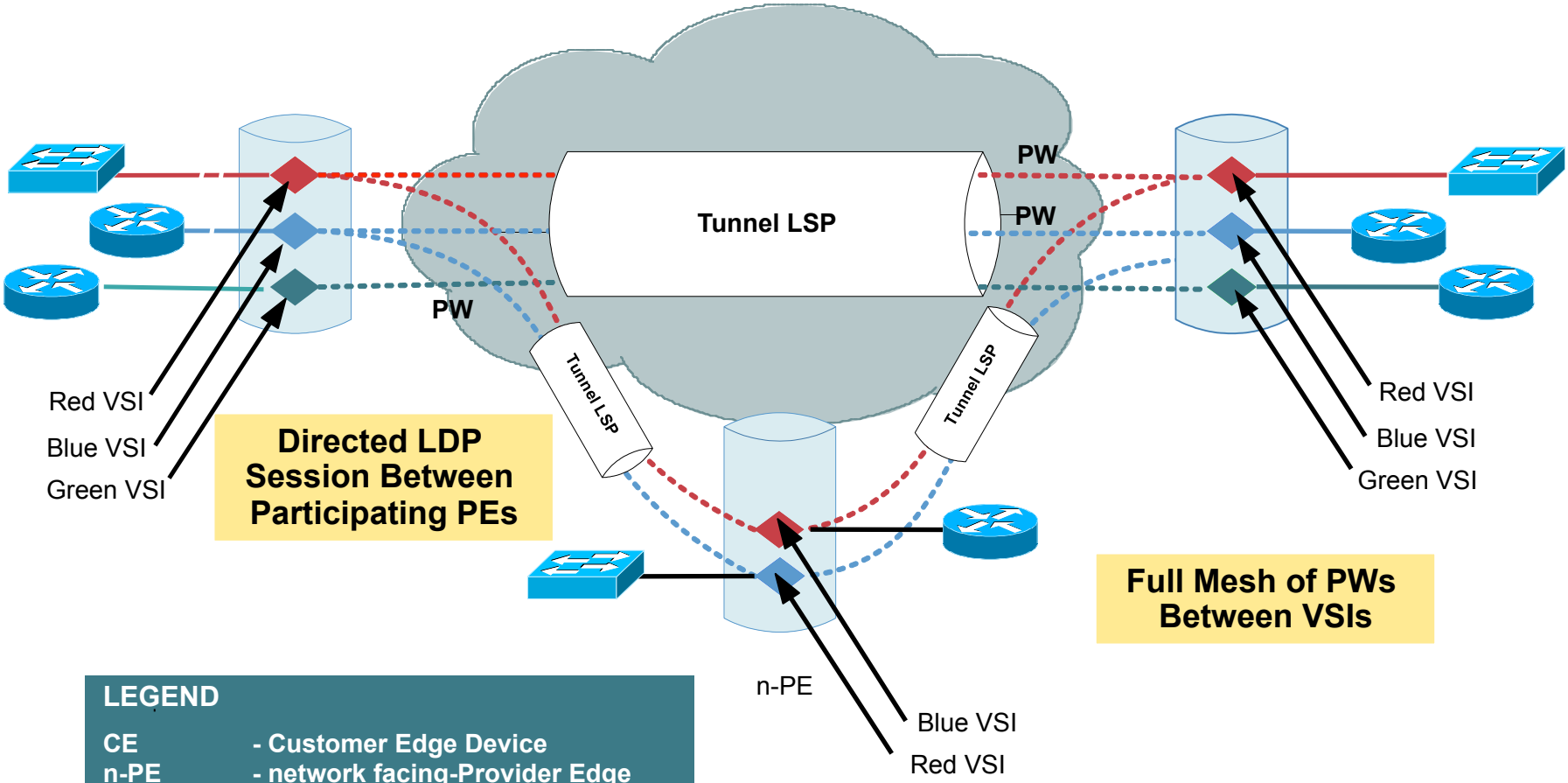
- **Several drafts in existence**

  draft-ietf-l2vpn-vpls-ldp-xx.txt

  draft-ietf-l2vpn-vpls-bgp-xx-txt

© 2006 Cisco Systems, Inc. All rights reserved.

VPLS Components

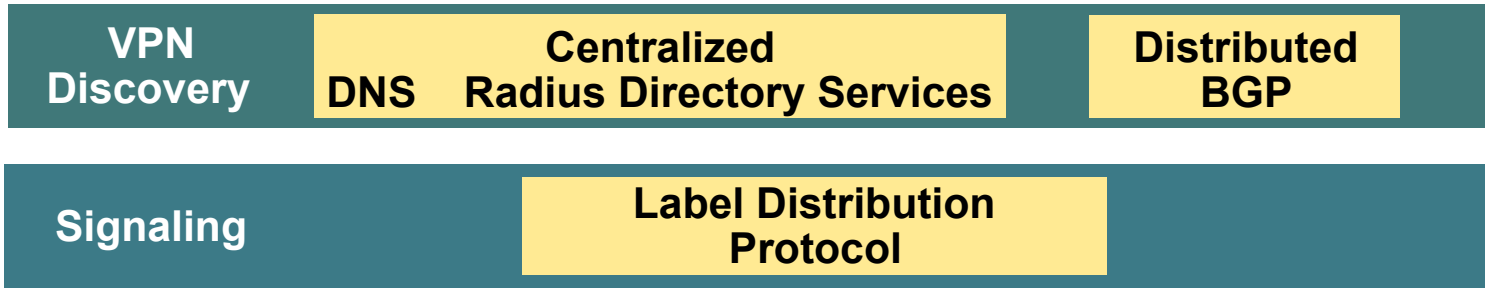


LEGEND

CE	- Customer Edge Device
n-PE	- network facing-Provider Edge
VSI	- Virtual Switch Instance
PW	- Pseudo-Wire
Tunnel LSP	- Tunnel Label Switch Path that provides PW transport

© 2006 Cisco Systems, Inc. All rights reserved.

VPLS Auto-Discovery and Signaling



- **Draft-ietf-l2vpn-vpls-ldp-01 does not mandate an auto-discovery protocol**
  - **Can be BGP, RADIUS, DNS based**
- **Draft-ietf-l2vpn-vpls-ldp-01 describes using Targeted LDP for Label exchange and PW signaling**
  - **PWs signal other information such as attachment circuit state, sequencing information, etc.**

© 2006 Cisco Systems, Inc. All rights reserved.

VPLS: Layer 2 Forwarding Instance Requirements

A Virtual Switch Must Operate Like a Conventional L2 Switch!

Flooding/Forwarding:

- MAC table instances per customer and per customer VLAN (L2-VRF idea) for each PE
- VSI will participate in learning, forwarding process
- Uses Ethernet VC-Type defined in pwe3-control-protocol-xx

Address Learning/Aging:

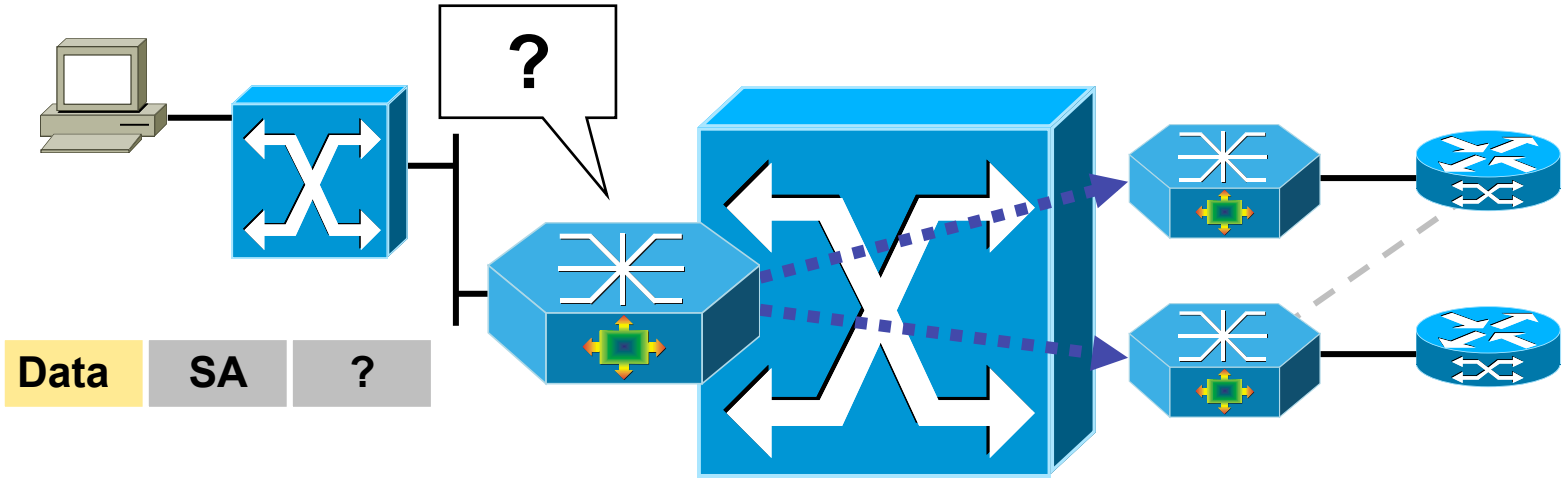
- Self-learn source MAC to port associations
- Refresh MAC timers with incoming frames
- New additional MAC TLV to LDP

Loop Prevention:

- Create partial or full-mesh of EoMPLS VCs per VPLS
- Use “split horizon” concepts to prevent loops

© 2006 Cisco Systems, Inc. All rights reserved.

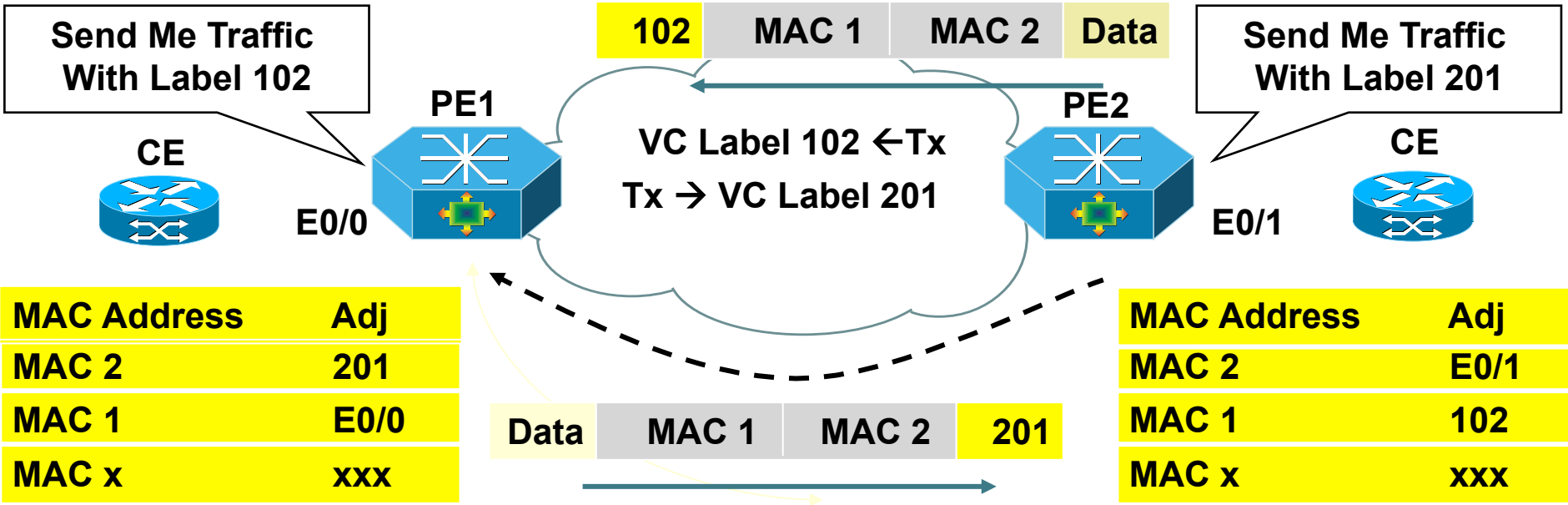
VPLS Overview:
Flooding and Forwarding



- Flooding (Broadcast, Multicast, Unknown Unicast)
- Dynamic learning of MAC addresses on PHY and VCs
- Forwarding
  - Physical port
  - Virtual circuit

© 2006 Cisco Systems, Inc. All rights reserved.

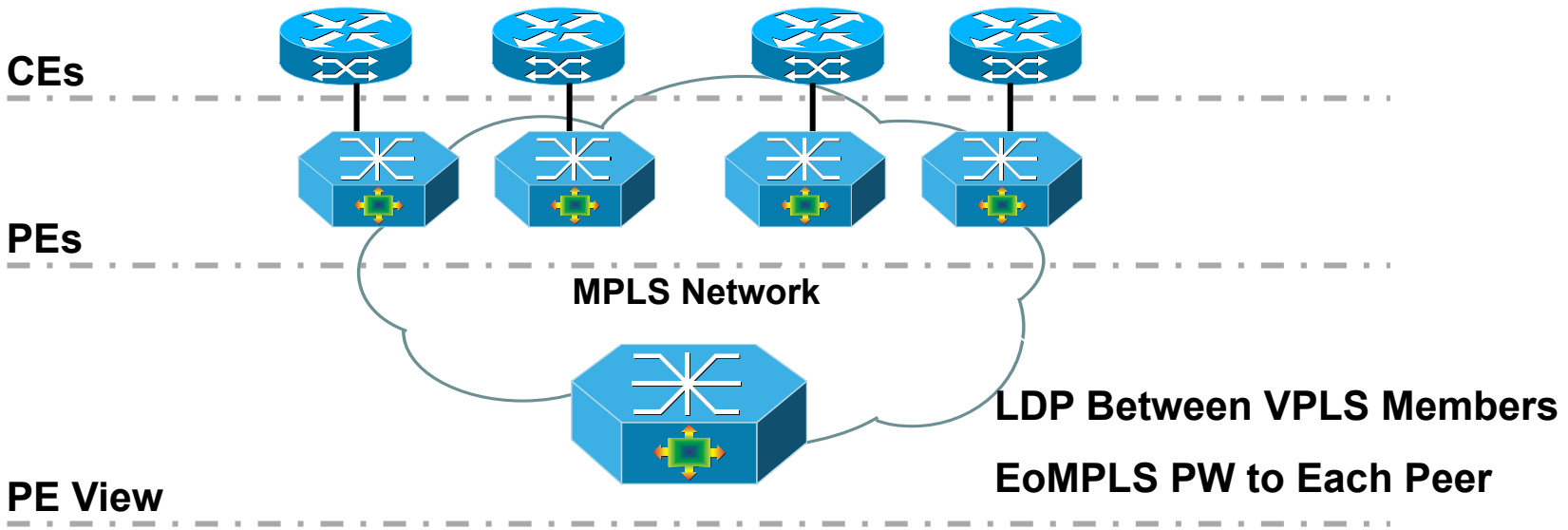
VPLS Overview:
MAC Address Learning



- Broadcast, multicast, and unknown unicast are learned via the received label associations
- Two LSPs associated with an VC (Tx and Rx)
- If inbound or outbound LSP is down, then the entire circuit is considered down

© 2006 Cisco Systems, Inc. All rights reserved.

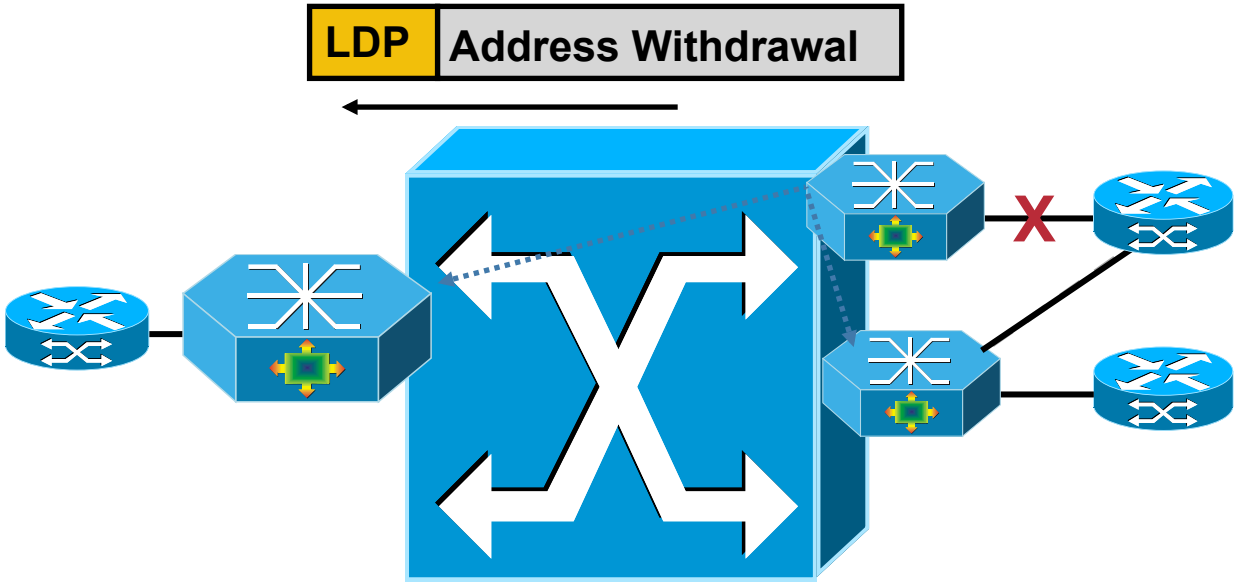
VPLS Overview:
VPLS Loop Prevention



- Each PE has a P2MP view of all other PEs it sees it self as a root bridge, split horizon loop protection
- Full mesh topology obviates STP requirements in the service provider network
- Customer STP is transparent to the SP/customer BPDUs are forwarded transparently
- Traffic received from the network will not be forwarded back to the network

© 2006 Cisco Systems, Inc. All rights reserved.

VPLS Overview: MAC Address Withdrawal

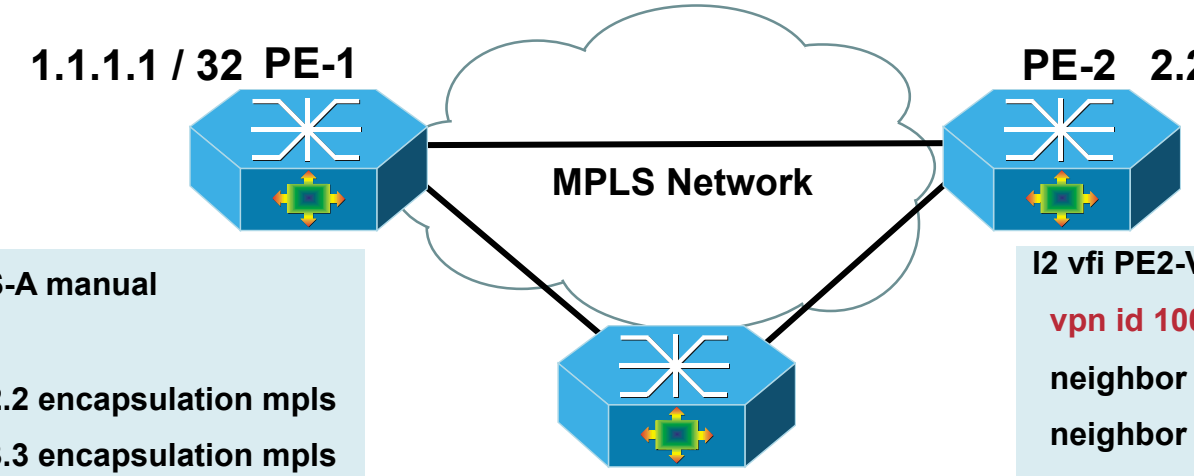


- Primary link failure triggers notification message
- PE removes any locally learned MAC addresses and sends LDP address withdrawal (RFC3036) to remote PEs in VPLS
- New MAC TLV is used

© 2006 Cisco Systems, Inc. All rights reserved.

VPLS: Configuration Example
PE → PE

Create a L2 VFI with a Full Mesh of Participating VPLS PE Nodes



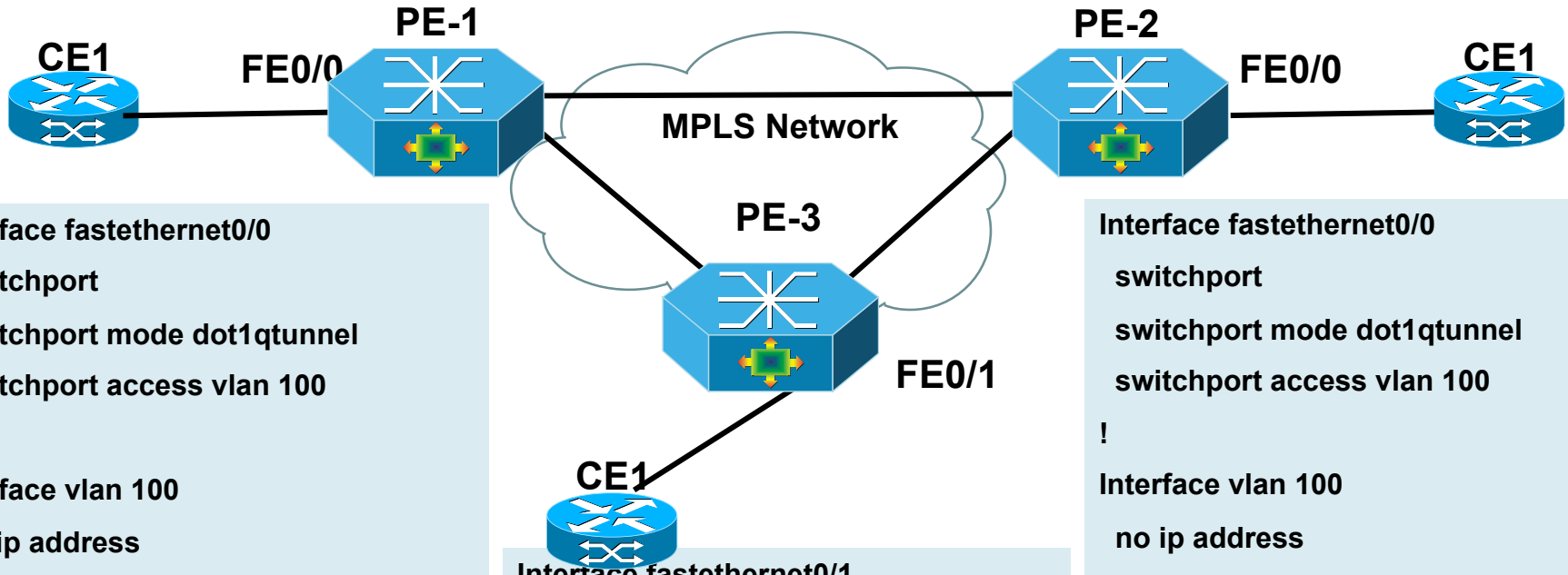

```
l2 vfi PE1-VPLS-A manual
  vpn id 100
  neighbor 2.2.2.2 encapsulation mpls
  neighbor 3.3.3.3 encapsulation mpls
!
Interface loopback 0
  ip address 1.1.1.1 255.255.255.255
```

```
l2 vfi PE2-VPLS-A manual
  vpn id 100
  neighbor 1.1.1.1 encapsulation mpls
  neighbor 3.3.3.3 encapsulation mpls
!
Interface loopback 0
  ip address 2.2.2.2 255.255.255.255
```

```
l2 vfi PE3-VPLS-A manual
  vpn id 100
  neighbor 1.1.1.1 encapsulation mpls
  neighbor 2.2.2.2 encapsulation mpls
!
Interface loopback 0
  ip address 3.3.3.3 255.255.255.255
```

© 2006 Cisco Systems, Inc. All rights reserved.

VPLS: Configuration Example
PE → CE



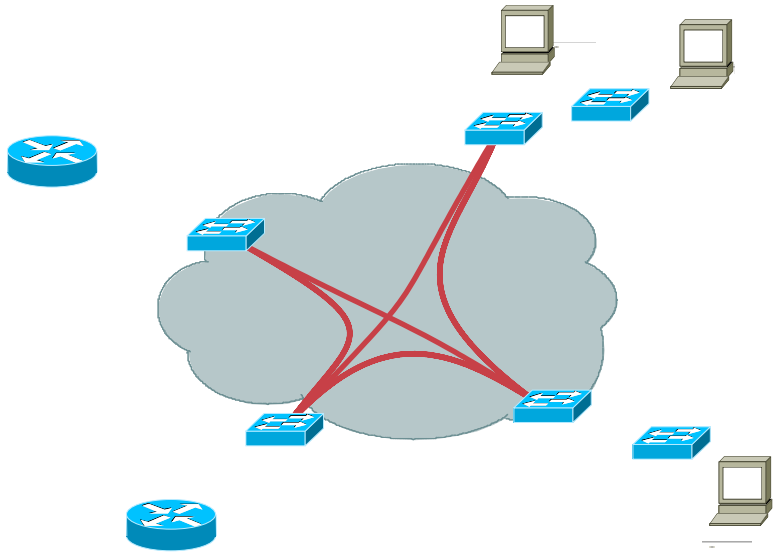

```
Interface fastethernet0/0
  switchport
  switchport mode dot1qtunnel
  switchport access vlan 100
!
Interface vlan 100
  no ip address
 xconnect vfi PE1-VPLS-A
!
vlan 100
  state active
```

```
Interface fastethernet0/1
  switchport
  switchport mode dot1qtunnel
  switchport access vlan 100
!
Interface vlan 100
  no ip address
 xconnect vfi PE3-VPLS-A …etc.
```

```
Interface fastethernet0/0
  switchport
  switchport mode dot1qtunnel
  switchport access vlan 100
!
Interface vlan 100
  no ip address
 xconnect vfi PE2-VPLS-A
!
vlan 100
  state active
```

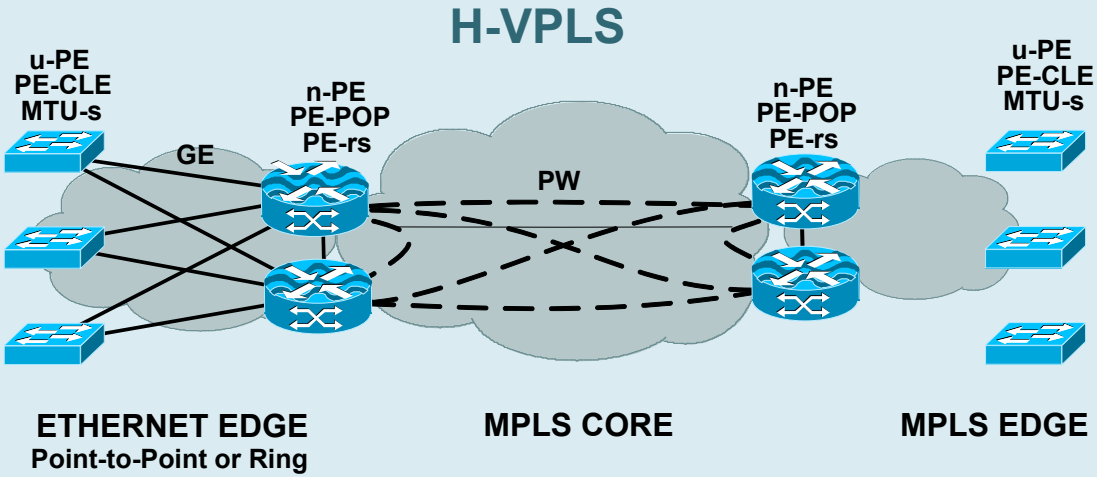
© 2006 Cisco Systems, Inc. All rights reserved.

VPLS and H-VPLS



- **VPLS**
  - **Single flat hierarchy**
  - **MPLS to the EDGE**

- **H-VPLS**
  - **Two Tier Hierarchy**
  - **MPLS or Ethernet Edge**
  - **MPLS Core**



© 2005 Cisco Systems, Inc. All rights reserved.