

MPLS TE Seminar

Umberto Poschi

umberto@poschi.it

uposchi@cisco.com

Traffic Engineering (RFC 2702)

- **Traffic Engineering (TE) is concerned with performance optimization of operational networks.**
- **Traffic Engineering encompasses the application of technology and scientific principles to the measurement, modeling, characterization, and control of Internet traffic, and the application of such knowledge and techniques to achieve specific performance objectives.**

Traffic Engineering (RFC 2702)

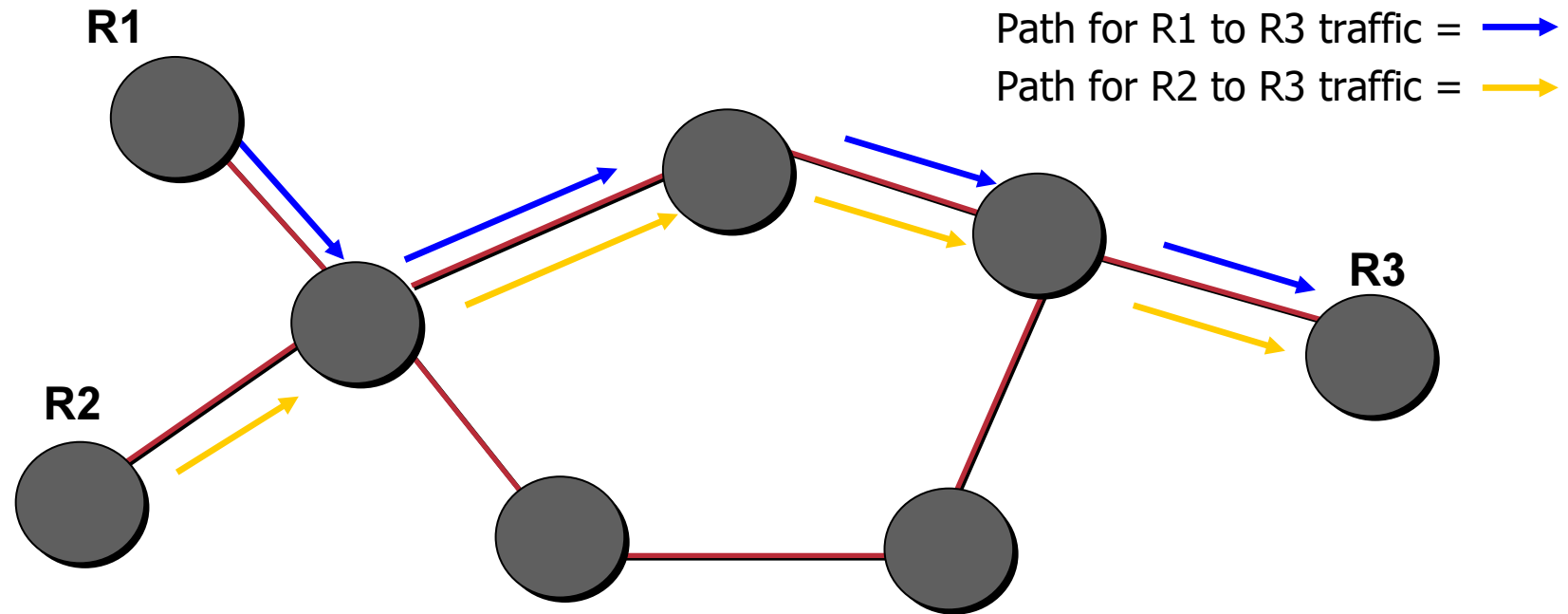
- **The key performance objectives for Traffic Engineering can be classified as follow:**
 - Traffic oriented
 - Resource oriented.

The Motivations for Traffic Engineering

Traffic Engineering: The Congestion Problem

- **Minimizing congestion is a primary traffic and resource oriented performance objective.**
 - congestion problems that are prolonged rather than on transient congestion resulting from instantaneous bursts.
 - Congestion typically manifests under two scenarios:
 - When network resources are insufficient or inadequate to accommodate offered load.
 - When traffic streams are inefficiently mapped onto available resources; causing subsets of network resources to become over-utilized while others remain underutilized.

Traffic Engineering: The Congestion Problem. Cont.

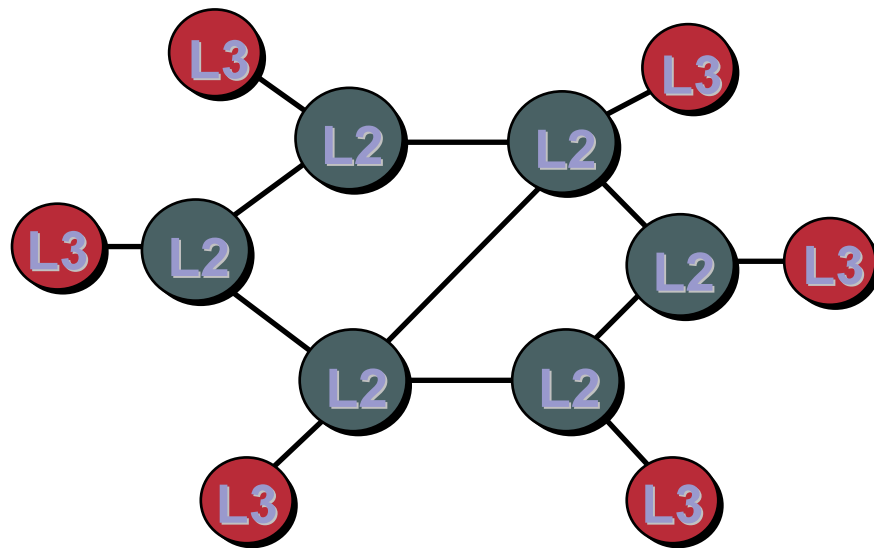


- **Conventional IGP path computation is selected based upon a simple additive metric**
 - Bandwidth availability is not taken into account
- **Some links may be underutilized while others are congested**

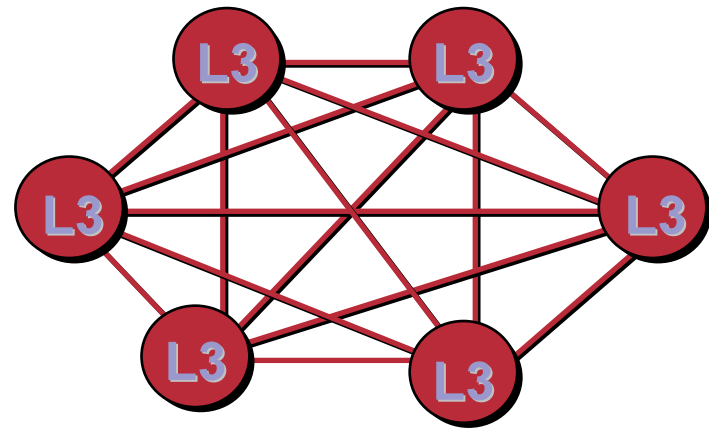
IP Routing: Tinkering With Metrics

- **Support for “explicit” (a.k.a. “source”) routing with the ability to steer traffic via the under-utilized parts of the network is not available**
 - Voice networks, Frame Relay, ATM are explicitly routed at connection setup
- **Conventional IGPs (e.g. IS-IS, OSPF) do not provide us with the capabilities needed to suitably engineer traffic**

The “Overlay” Solution: Traffic Engineering at Layer 2



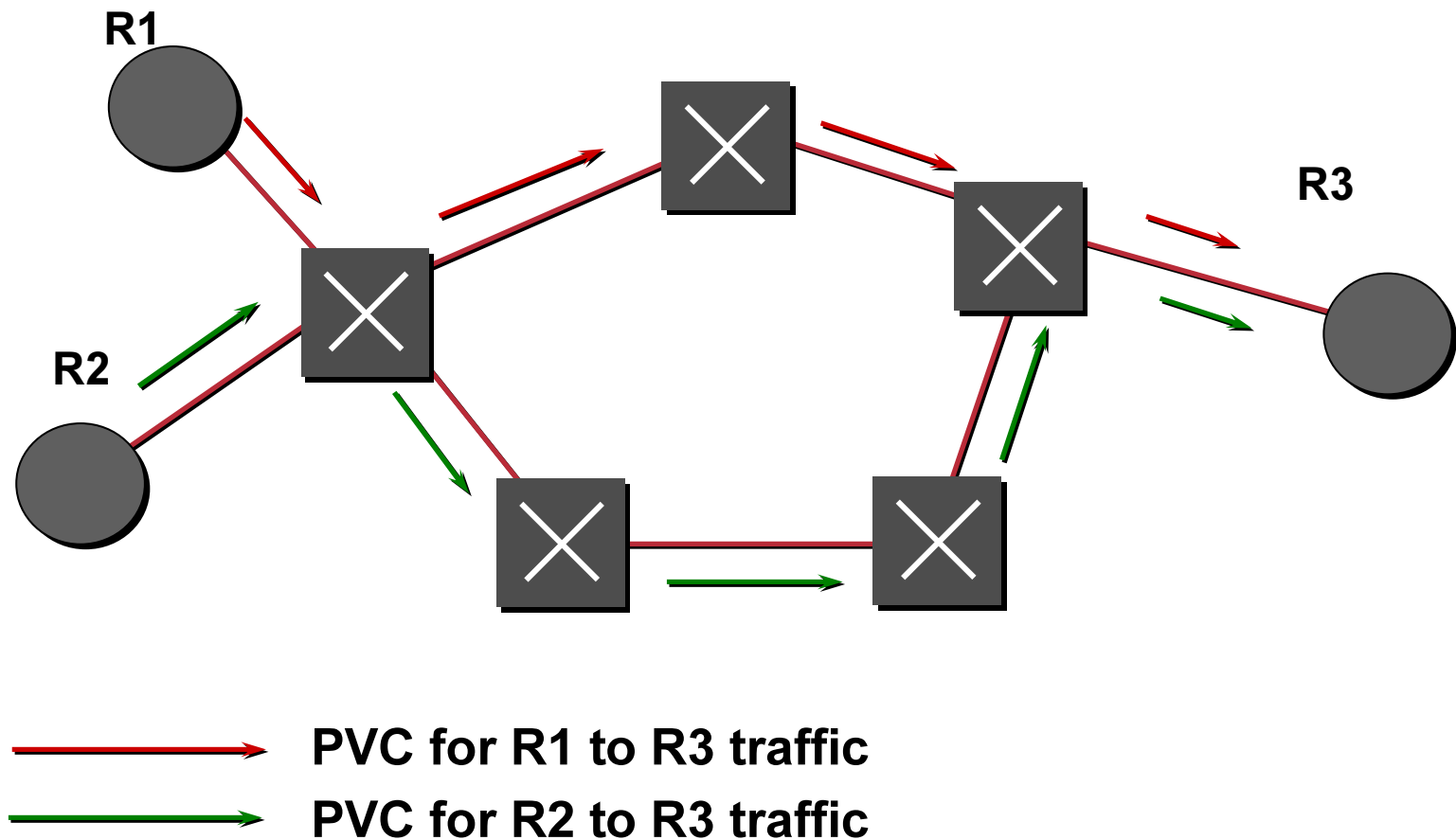
Physical



Logical

- **Routing at layer 2 (ATM or FR) is used for traffic engineering**
- **Layer 3 sees a complete mesh and routing at layer 3 is trivial**

The “Overlay” Solution: Traffic Engineering at Layer 2



The “Overlay” Solution: Traffic Engineering at Layer 2

- **Traffic Engineering at Layer 2 gives control not possible by tinkering with conventional IGP metrics**
- **Added complexity – two networks to design deploy and manage**
 - greater cost
 - longer lead times
- **IGP routing scalability issues for meshes**
- **Additional bandwidth overhead (“cell tax”)**

MPLS Traffic Engineering and its Components

MPLS Traffic Engineering

- **Traffic engineering requires an explicit routing capability**
 - IP supports only the destination-based routing not adequate for traffic engineering
- **MPLS Traffic Engineering gives us an “explicit” routing capability (a.k.a. “source routing”) at Layer 3**
 - Lets you use paths other than IGP shortest path
 - Allows unequal-cost load sharing
 - The benefits of Layer 2 traffic engineering without the disadvantages

MPLS Traffic Engineering

- **MPLS provides simple and efficient support for explicit routing**
 - separation of routing and forwarding
 - MPLS label swapping as the forwarding mechanism
 - use of explicitly routed Label Switched Paths (LSPs) to steer traffic through the network
 - RSVP as the mechanism for establishing LSPs

MPLS TE Components

- (1) Resource / policy information distribution**
- (2) Constraint based path computation**
- (3) RSVP for tunnel signaling**
- (4) Link admission control**
- (5) LSP establishment**
- (6) TE tunnel control and maintenance**
- (7) Assign traffic to tunnels**

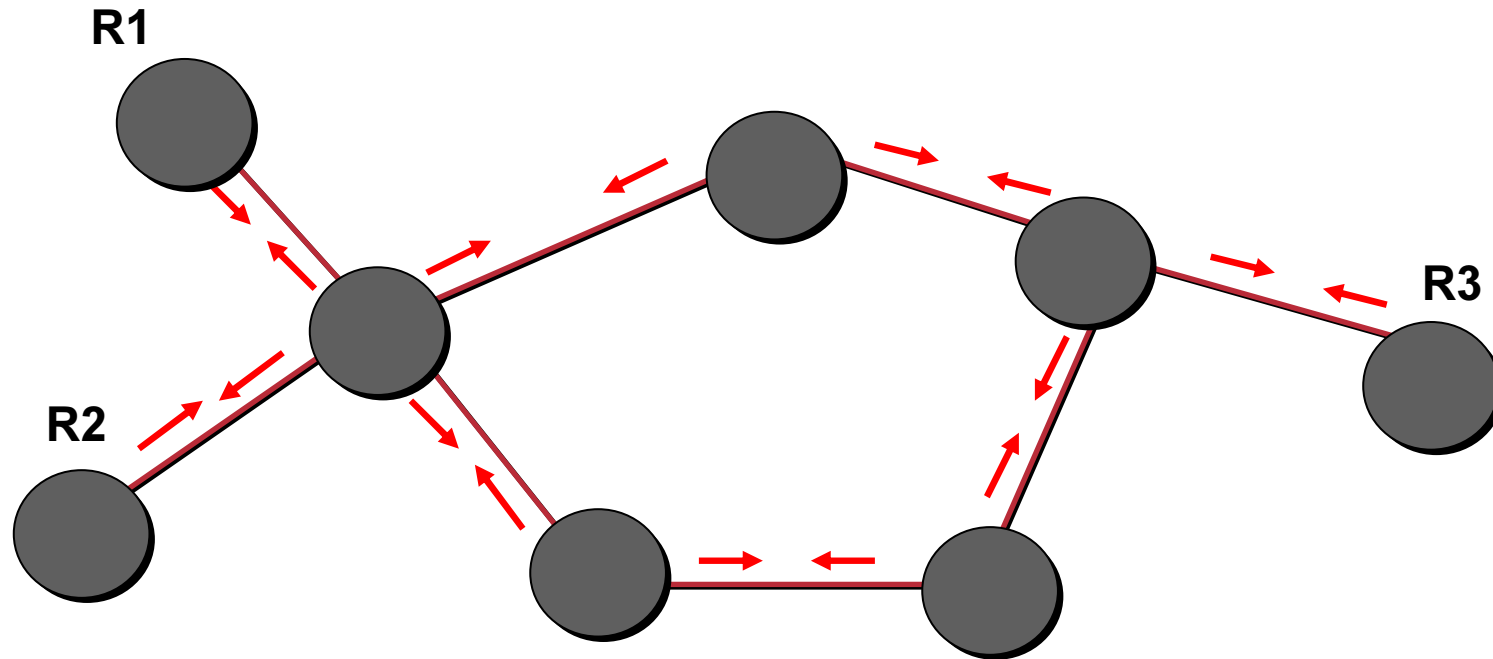
MPLS TE Components

(1) Resource / policy information distribution

- Extensions to OSPF / IS-IS are used to distribute resource or policy constraints pertaining to links**
- Available bandwidth is just one type of constraint**

MPLS TE Components:

(1) Resource / policy information distribution



- **OSPF / IS-IS extensions are used to distribute link resource or policy constraint information:**
 - Available bandwidth and different priorities levels
 - Administrative policy (Resource Class Affinity)

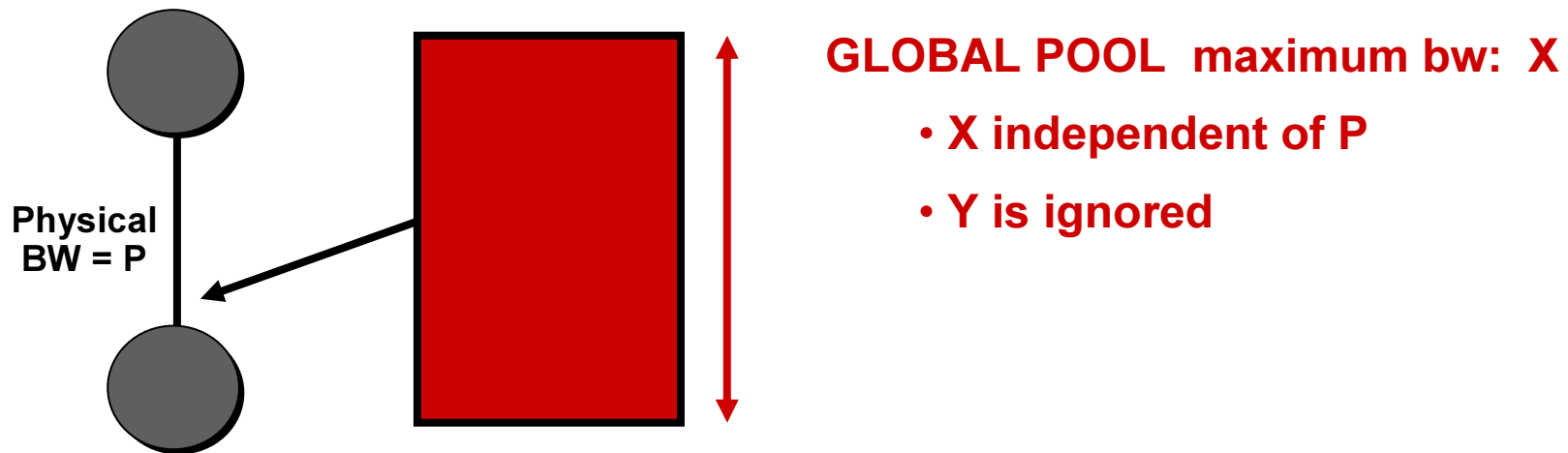
Link Attributes

- **Resource / policy attributes are configured on every link and define the capabilities of the network**
 - Bandwidth
 - Resource Class Affinity string (policy)
 - TE-specific link metric
- **When performing the constraint based path computation, the head-end compares the link attributes received via the IGP to those configured on the tunnel**

Tunnel Attributes

- **Configured at the head-end of the tunnel**
- **Define the requirements for the tunnel**
 - **Bandwidth**
 - **Priorities**
 - **Setup priority: priority for taking a resource**
 - **Hold priority: priority for holding a resource**
 - **Resource Class Affinity string (Policy)**

Available Bandwidth Attribute



- Bandwidth pools definition:
 - X is the bandwidth constraint on all tunnels at all pre-emption levels
(Sum of all tunnels $\leq X$)

Per-Priority Available Bandwidth Example

ip rsvp ban 1000 1

B/W	Priority
1000	0
1000	1
1000	2
1000	3
1000	4
1000	5
1000	6
1000	7

- **Successively, we examine:**
 - Tu1 requests 200k at P=3/3
 - Tu2 requests 750k at P=4/4
 - Tu3 requests 30k at P=5/5
 - Tu4 requests 30k at P=6/6

Initial view of bandwidth utilisation

Per-Priority Available Bandwidth Example

ip rsvp ban 1000 1

B/W	Priority
1000	0
1000	1
1000	2
800	3
800	4
800	5
800	6
800	7

- Successively, we examine:
 - Tu1 requests 200k at P=3/3
 - Tu2 requests 750k at P=4/4
 - Tu3 requests 30k at P=5/5
 - Tu4 requests 30k at P=6/6

tu1 accepted!

Per-Priority Available Bandwidth Example

ip rsvp ban 1000 1

B/W	Priority
1000	0
1000	1
1000	2
800	3
50	4
50	5
50	6
50	7

- **Successively, we examine:**
 - Tu1 requests 200k at P=3/3
 - Tu2 requests 750k at P=4/4**
 - Tu3 requests 30k at P=5/5
 - Tu4 requests 30k at P=6/6

tu2 accepted!

Per-Priority Available Bandwidth Example

ip rsvp ban 1000 1

B/W	Priority
1000	0
1000	1
1000	2
800	3
50	4
20	5
20	6
20	7

- **Successively, we examine:**
 - Tu1 requests 200k at P=3/3
 - Tu2 requests 750k at P=4/4
 - Tu3 requests 30k at P=5/5**
 - Tu4 requests 30k at P=6/6

tu3 accepted!

Per-Priority Available Bandwidth Example

ip rsvp ban 1000 1

B/W	Priority
1000	0
1000	1
1000	2
800	3
50	4
20	5
20	6
20	7

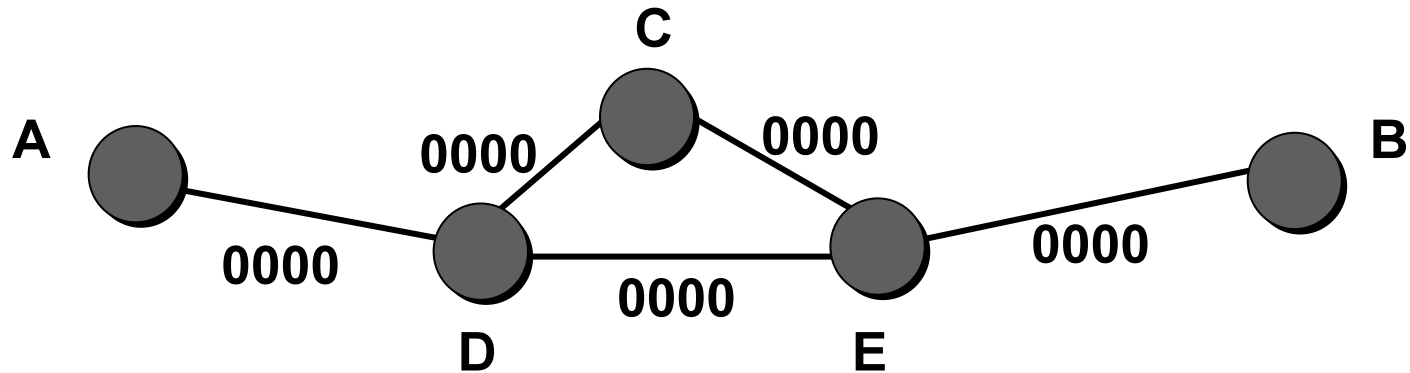
- **Successively, we examine:**
 - Tu1 requests 200k at P=3/3
 - Tu2 requests 750k at P=4/4
 - Tu3 requests 30k at P=5/5
 - Tu4 requests 30k at P=6/6**

tu4 REJECTED!

Resource class affinity attribute

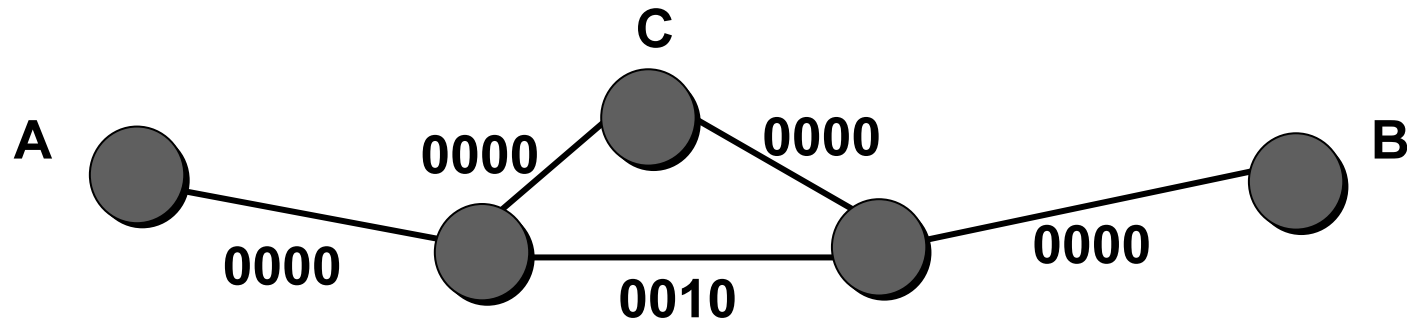
- **Supports the ability to include / exclude certain links for certain traffic trunks based on a user-defined Policy**
- **Tunnel is characterized by:**
 - **32-bit resource-class affinity string**
 - **32-bit resource-class mask (0 = don't care, 1 = care)**
- **Link is characterized by a 32-bit resource-class affinity string**
- **Default-value of tunnel / link bits is 0**
- **Default value of the tunnel mask = 0x0000FFFF**

Resource Class Affinity Attribute: Example 1: 4-bit string, default



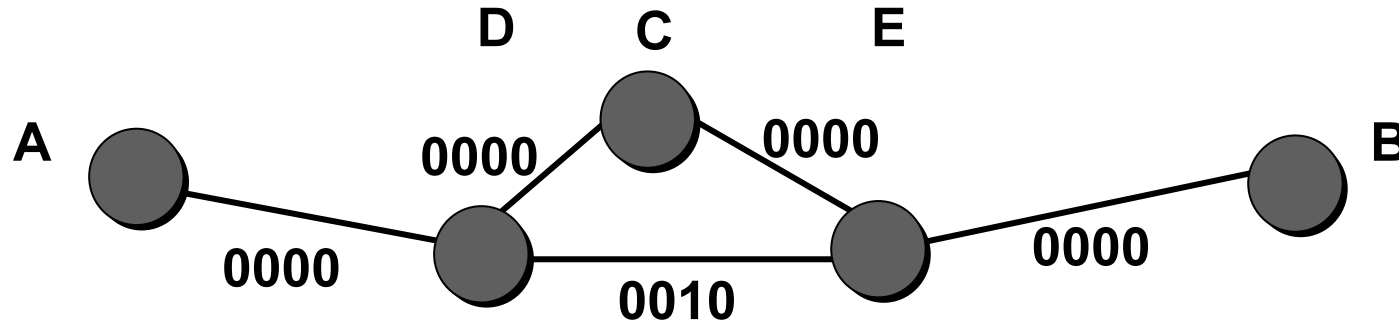
- **Tunnel A to B:**
 - tunnel affinity = 0000, tunnel mask = 0011
- **ADEB and ADCEB are possible**

Resource Class Affinity Attribute: Example 2: 4-bit string



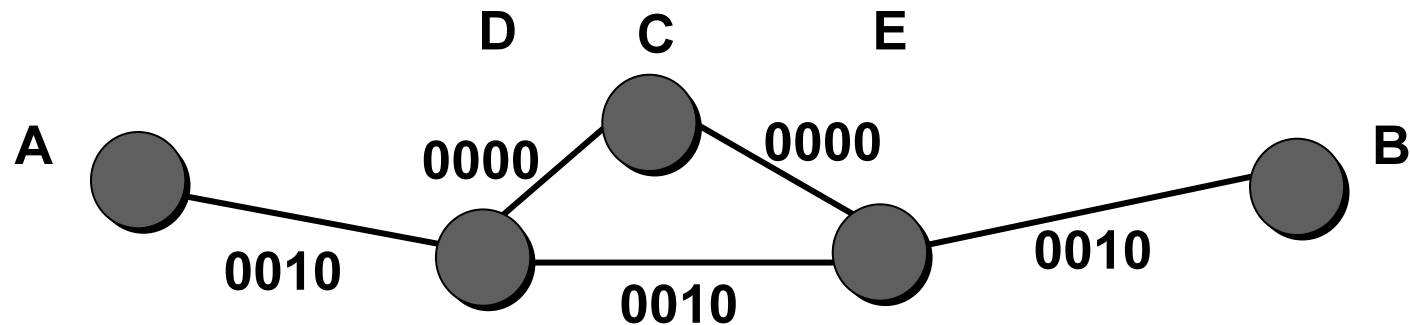
- Setting a bit on link D-E drives all tunnels off the link, except those specially configured
- Tunnel from A to B: D E
 - tunnel affinity = 0000, tunnel mask = 0011
- Only ADCEB is possible

Resource Class Affinity Attribute: Example 3: 4-bit string



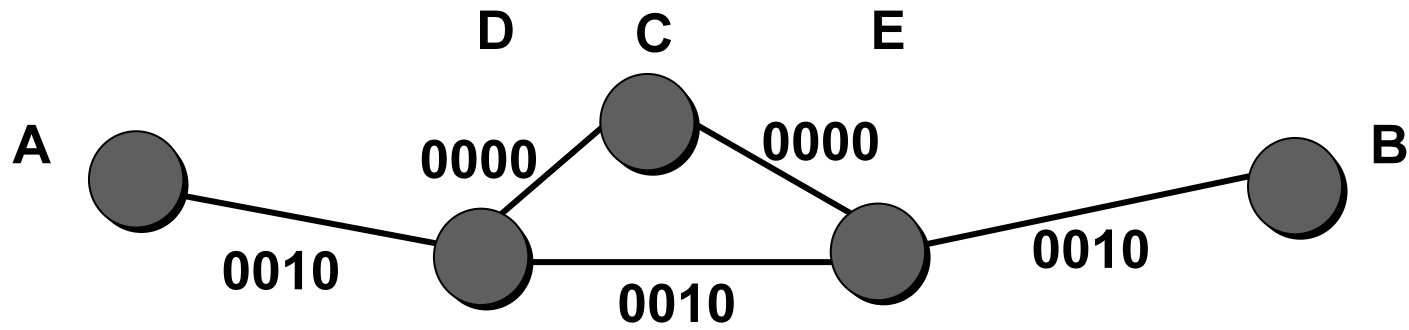
- A specific tunnel can be configured to allow such links by clearing that bit in its affinity attribute mask
- Tunnel from A to B:
 - tunnel affinity = 0000, tunnel mask = 0001
- Again, ADEB and ADCEB are possible

Resource Class Affinity Attribute: Example 4: 4-bit string



- **Alternatively, a specific tunnel can be restricted to only such links by instead setting the bit in its affinity attribute**
- **Tunnel from A to B:**
 - tunnel affinity = 00**1**0, tunnel mask = 00**1**1

Resource Class Affinity Attribute: Example 5: 4-bit string

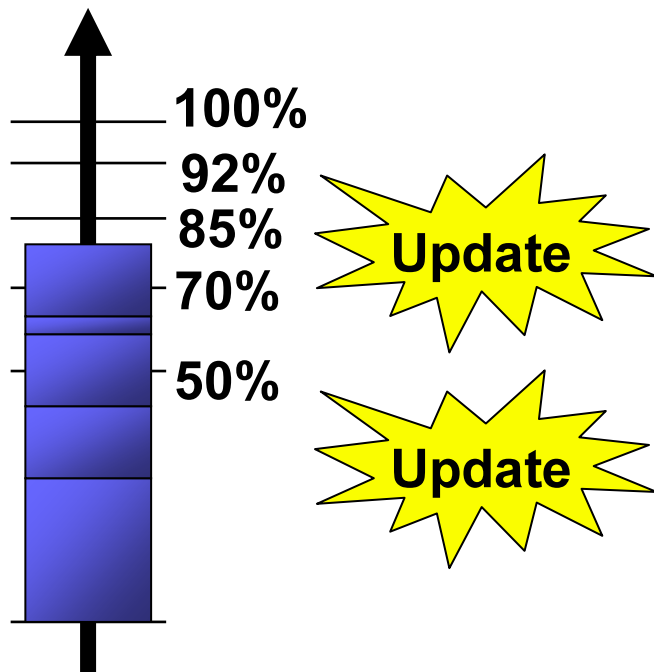


- By changing an additional bit in the affinity attribute, no tunnel paths from A to B are possible:
 - tunnel affinity = 00**11**, tunnel mask = 00**11**

Link Attribute Flooding

- **Flooding can be triggered by different events**
 - **Periodic (timer-based)**
 - **On significant changes of available bandwidth**
 - **Up/Down thresholds (in %)**
 - **On link configuration changes**
 - **On tunnel setup failure**

Link Attribute Flooding: Significant Change



- Each time a threshold is crossed, an update is sent
- Closer thresholds as utilization increases
- Different thresholds for UP and DOWN (more stable)
- Thresholds configurable to fine-tune flooding overhead vs. CSPF accuracy

Link Attribute Flooding: Tunnel Setup Failure

- **Due to the threshold scheme, it is possible that a router thinks that an LSP tunnel can be signalled via router Z while in fact, Z does not have the required resources**
- **When Z receives the Resv message and refuses the LSP tunnel, it broadcasts an update of its status**

Other Tunnel Characteristics

- **Ordered list of Path Options**
 - Possible administratively specified paths (via an off-line central server)
 - Constrained-based dynamically computed paths based on combination of bandwidth and policies
- **Re-optimization**
 - Each path option is enabled or not for re-optimization

MPLS TE Components

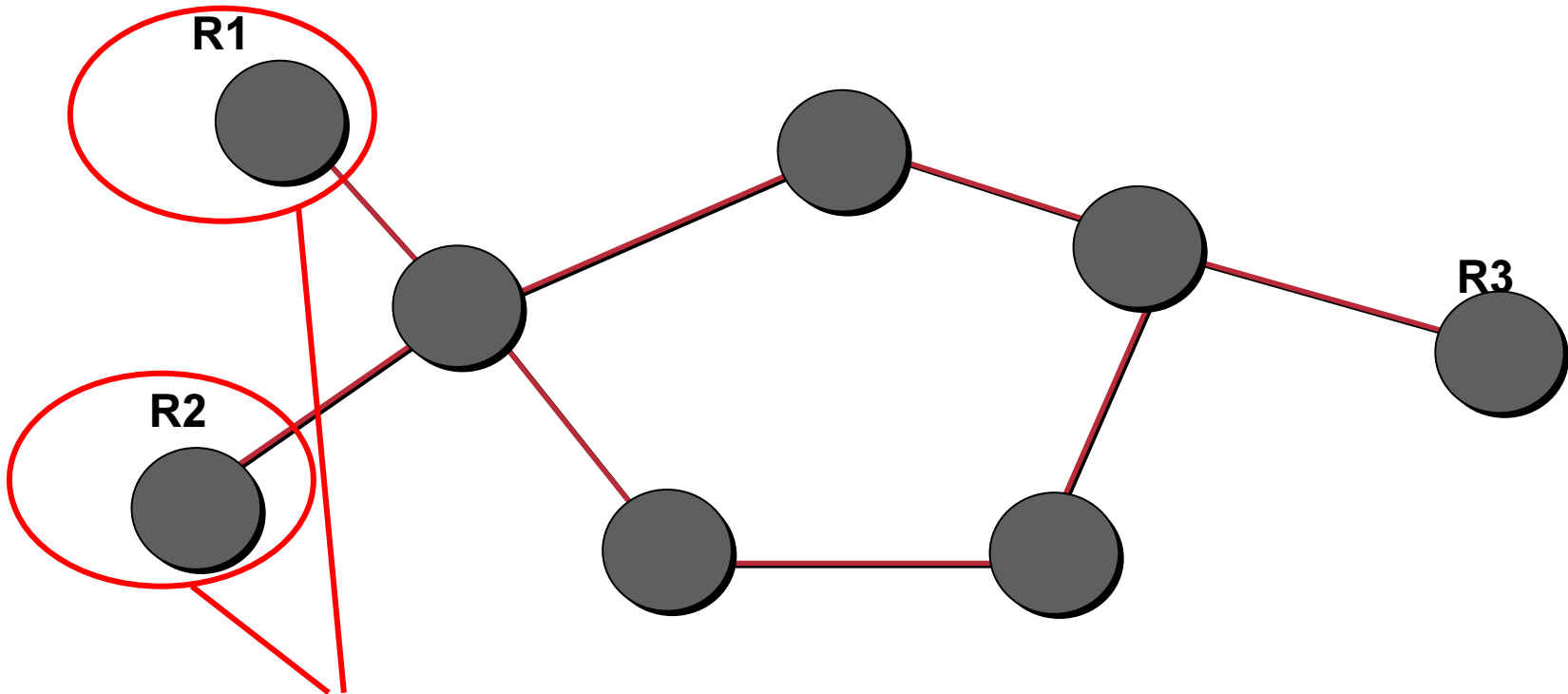
(1) Resource / policy information distribution

(2) Constraint based path computation

–Selects paths that obey the constraints

MPLS TE Components:

(2) Constraint based path computation



- PCALC on head-end routers calculates best path that satisfies constraints based upon the received topology and policy information
- Output is an explicit route used as an input to the tunnel signalling component

Constrained-Based Routing

“In general, path computation for an LSP may seek to satisfy a set of requirements associated with the LSP, taking into account a set of constraints imposed by administrative policies and the prevailing state of the network - which usually relates to topology data and resource availability. Computation of an engineered path that satisfies an arbitrary set of constraints is referred to as ‘constraint based routing’ ”.

Draft-li-mpls-igp-te-00.txt

Path Computation

- **For dynamic tunnels the head-end router determines the path**
 - can alternatively be statically configured on head-end
- **Path computation is “on demand”:**
 - for a new trunk
 - for an existing trunk whose (current) LSP failed
 - for an existing trunk when doing re-optimization

Path Computation

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

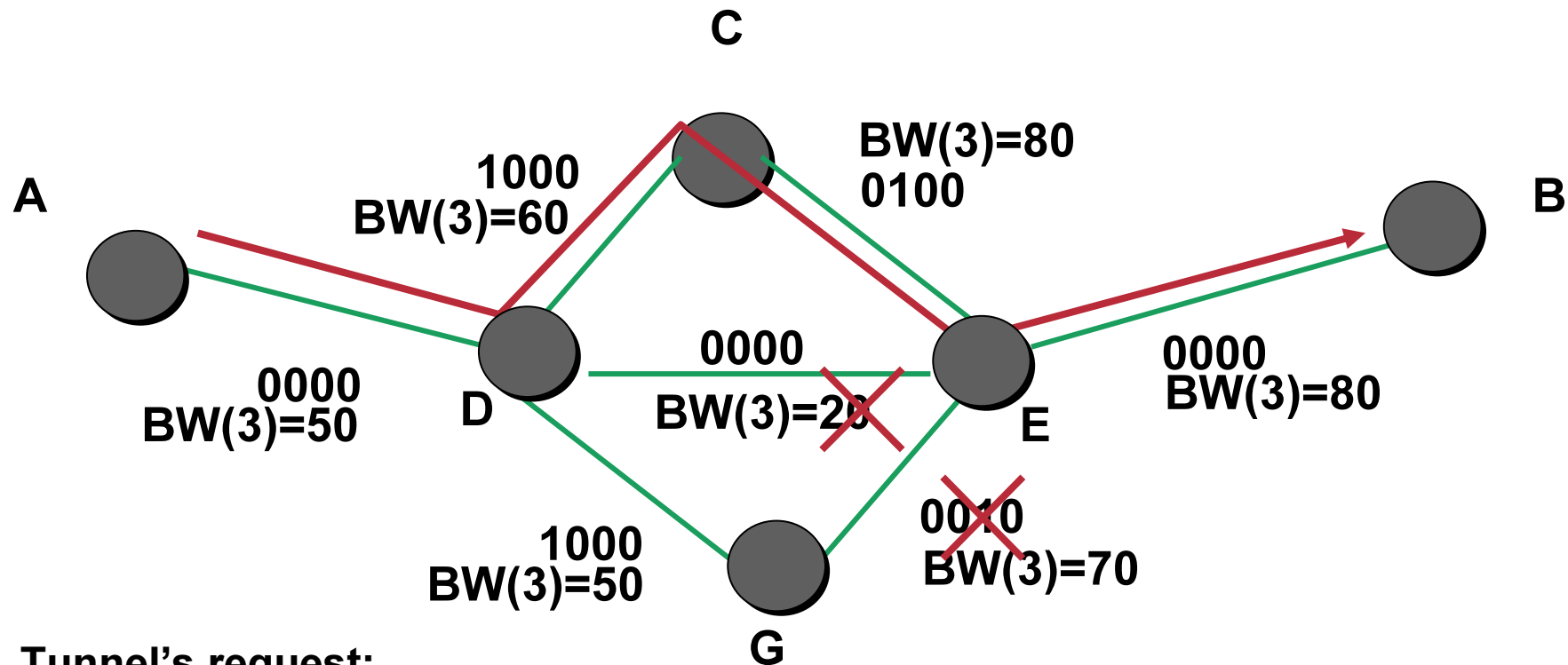
Path Computation

- **Prune links if:**
 - insufficient resources (e.g. bandwidth)
 - violates policy constraints
- **Compute shortest distance path**
 - Uses its own metric
 - In case of a tie-break: selects the path with the biggest left-over bandwidth, then with the smallest hop-count

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 path setup**

Example

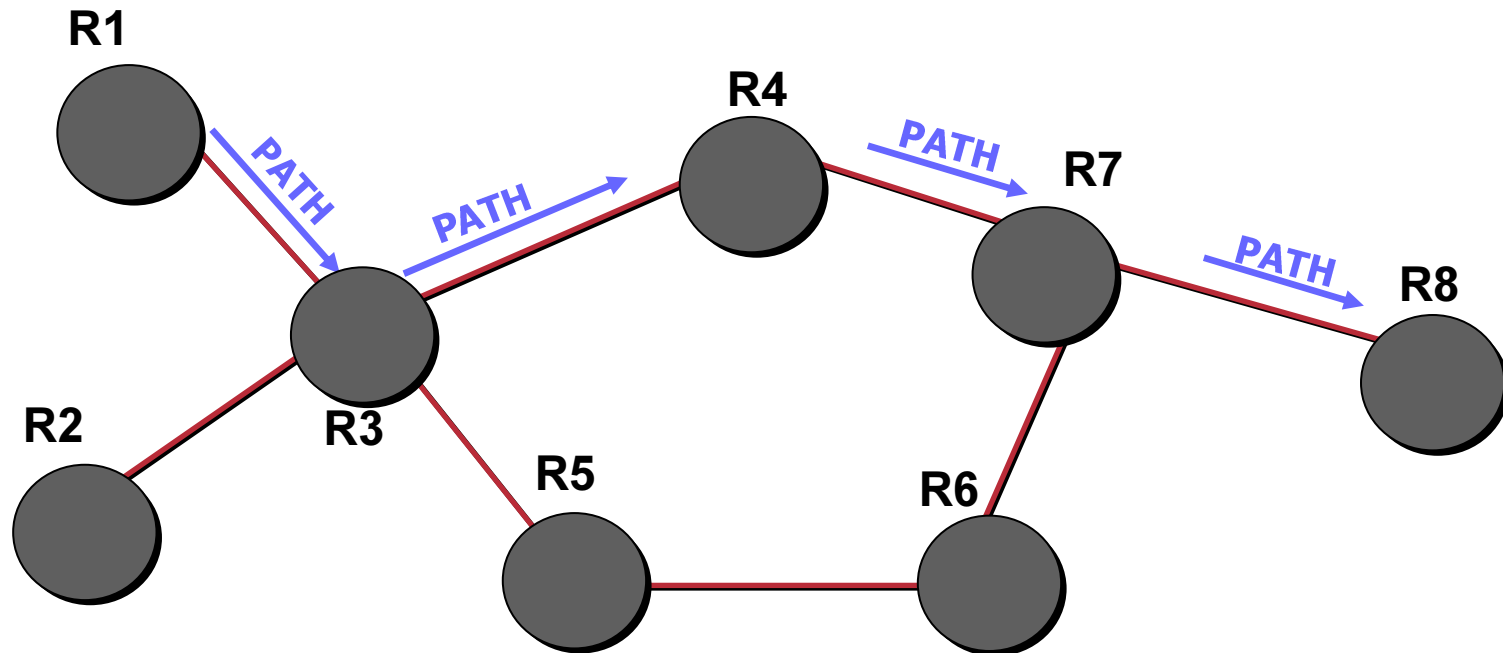


- Tunnel's request:
 - Priority 3, BW = 30 units,
 - Policy string: 0000, mask: 0011

MPLS TE Components

- (1) Resource / policy information distribution
- (2) Constraint based path computation
- (3) RSVP for tunnel signaling**
 - RSVP (with extensions) is used for signaling LSPs**

MPLS TE Components: (3) Tunnel Signaling



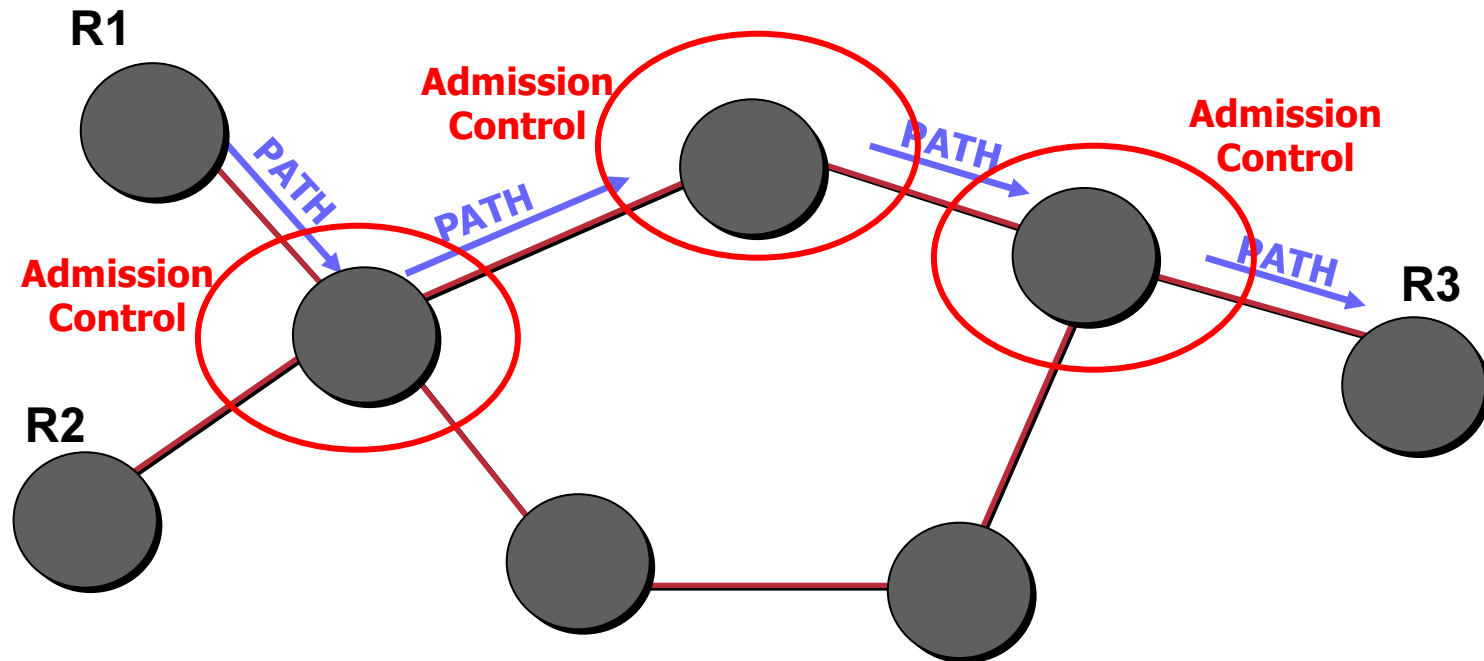
- **RSVP (with extensions) used for tunnel signaling**
 - **Uses explicit route object output from PCALC**
 - **ERO = R1->R3->R4->R7->R8**

MPLS TE Components

- (1) Resource / policy information distribution
- (2) Constraint based path computation
- (3) RSVP for tunnel signaling
- (4) Link admission control**
 - Decides which tunnels may use which resources (i.e. links)**

MPLS TE Components:

(4) Link admission control



- **At each hop – determines if resources are available**
 - If Admission Control fails, send PathError
 - May tear down (existing) TE LSPs with a lower priority
 - Triggers IGP information distribution when resource thresholds are crossed

Link Admission Control

- **Invoked by Path message**

- if BW is available, this BW is put aside in a waiting pool (waiting for the RESV msg)

- if this process required the pre-emption of resources, LCAC notified RSVP of the pre-emption which then sent PathErr and/or ResvErr for the pre-empted tunnel

- if BW is not available, LCAC says “No” to RSVP and a Path error is sent. A flooding of the node’s resource info is triggered, if needed

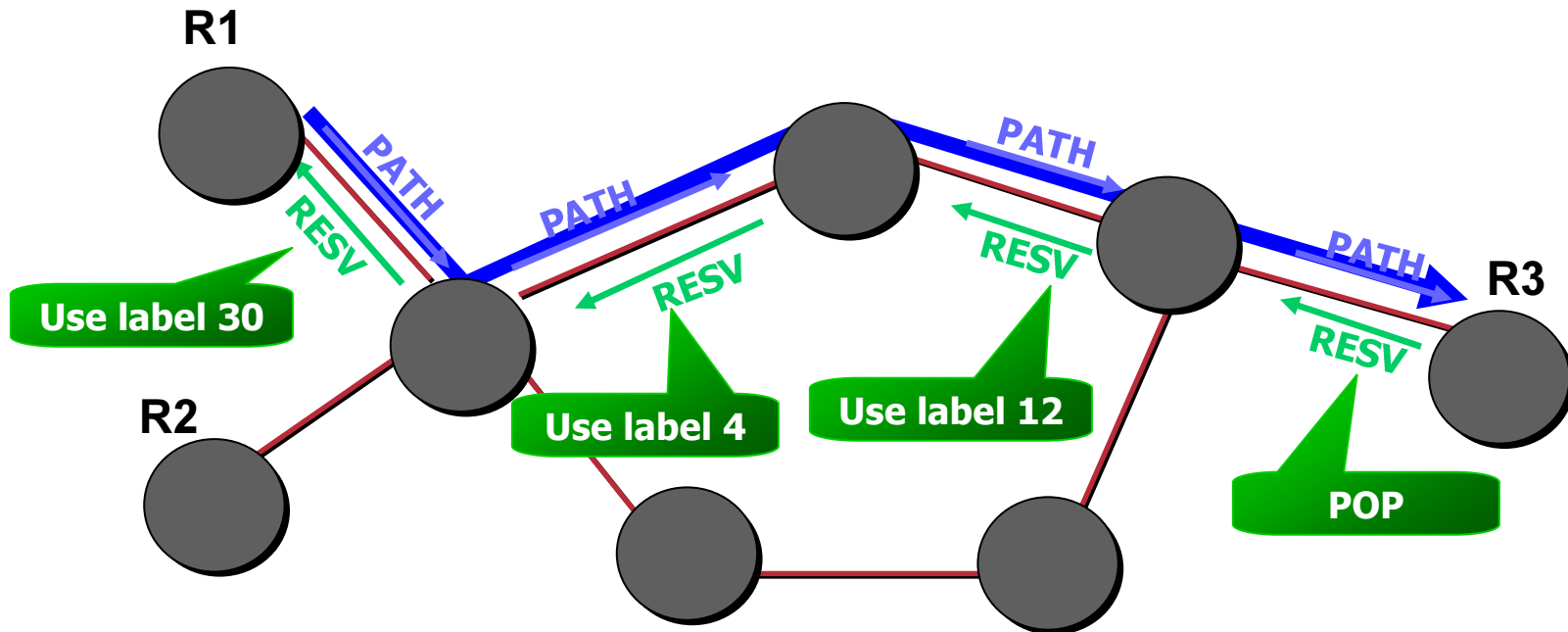
- “draft-ietf-mpls-rsvp-lsp-tunnel-02.txt”

MPLS TE Components

- (1) Resource / policy information distribution
- (2) Constraint based path computation
- (3) RSVP for tunnel signaling
- (4) Link admission control
- (5) LSP establishment**

MPLS TE Components:

(5) LSP Establishment



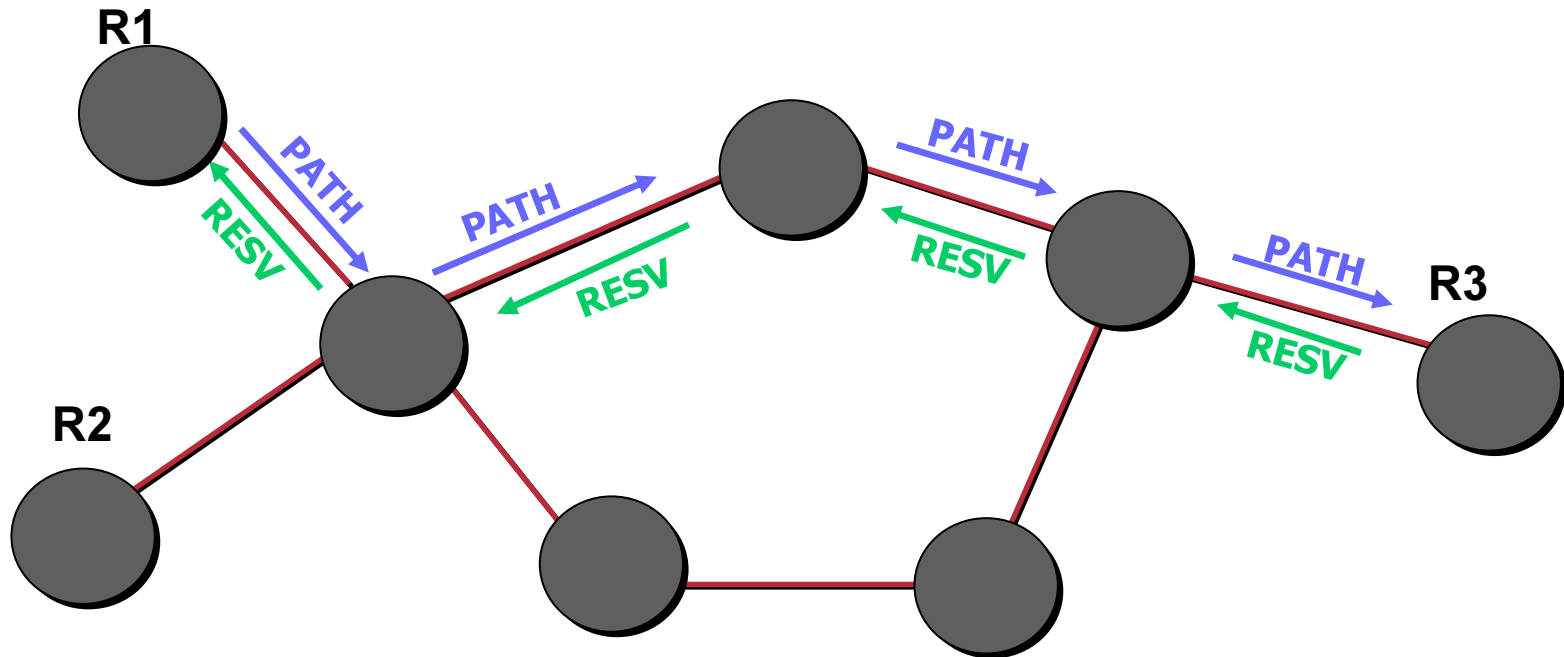
- **RESV confirms bandwidth reservation and distributes labels**
 - Downstream on demand label allocation
- **MPLS used for forwarding – overcomes issues of IP destination based forwarding**

MPLS TE Components

- (1) Resource / policy information distribution
- (2) Constraint based path computation
- (3) RSVP for tunnel signaling
- (4) Link admission control
- (5) LSP establishment
- (6) TE tunnel control and maintenance**
 - Establishes and maintains tunnels**

MPLS TE Components:

(6) TE tunnel control



- Periodic PATH and RESV refreshes establish and maintain tunnels

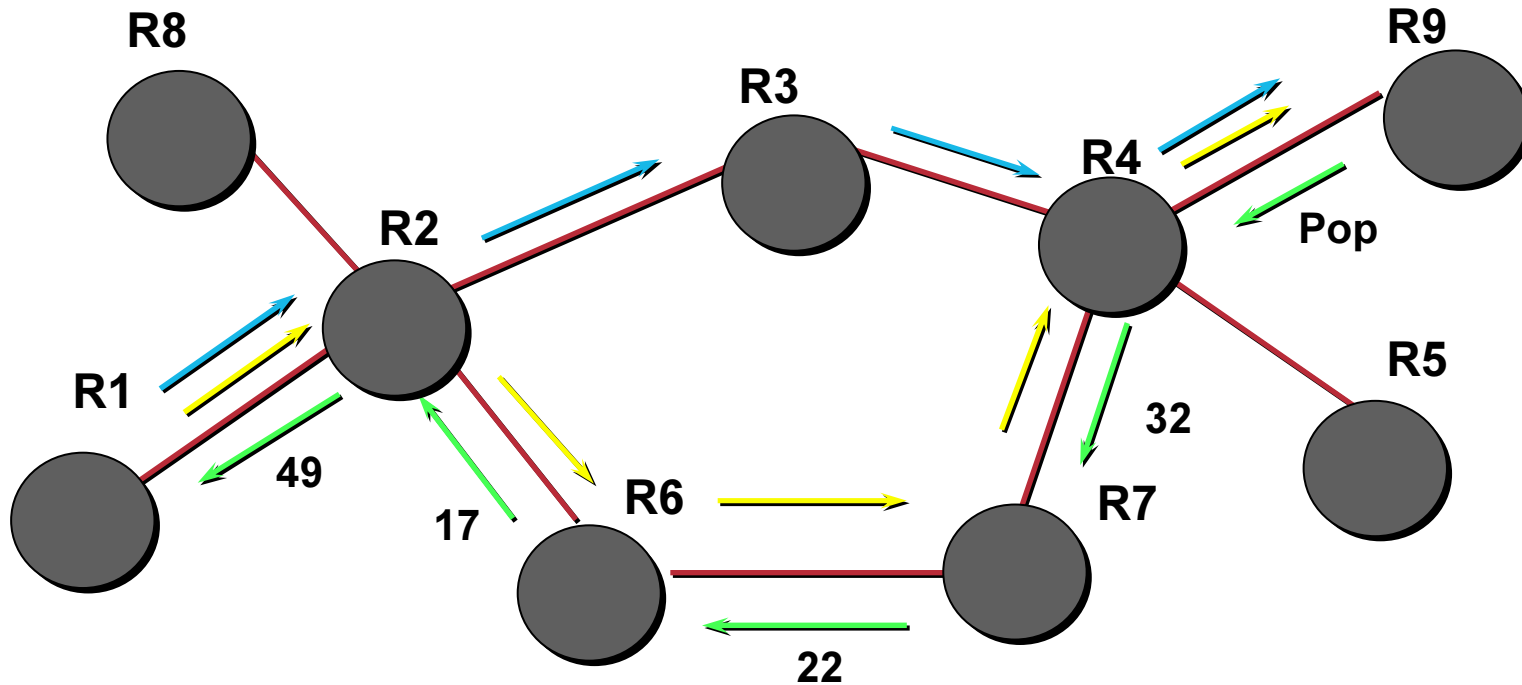
Path Monitoring

- **Use of new Record Route Object**
 - keep track of the exact tunnel path
 - detects loops
 - copy of RRO to ERO allows for route pinning

Path Re-Optimization

- **Paths can be re-optimized periodically or on demand**
- **Re-optimization characteristics:**
 - make before break
 - no double counting of reservations
 - via RSVP “shared explicit” style!

Non-disruptive rerouting: new path setup



Current Path (ERO = R1->R2->R6->R7->R4->R9)

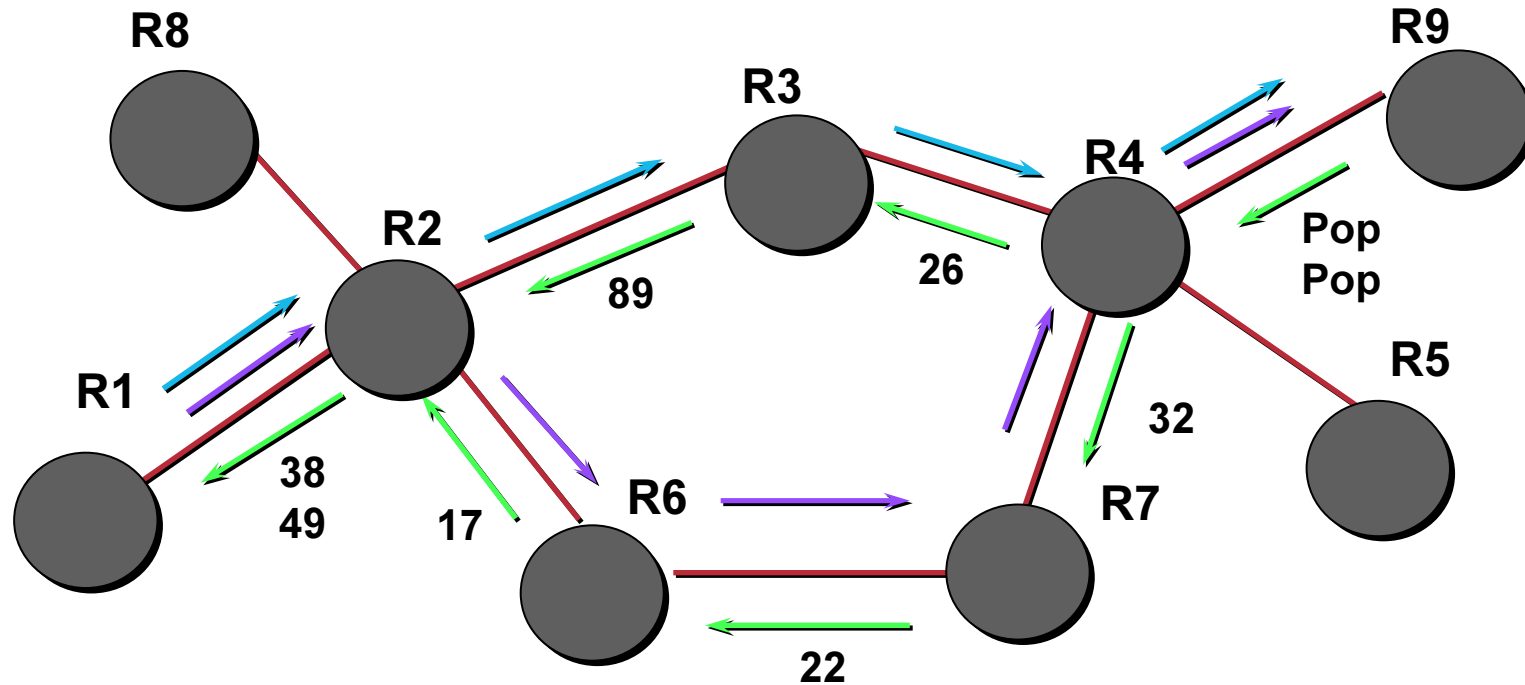


New Path (ERO = R1->R2->R3->R4->R9) - shared with Current Path



Until R9 gets new Path Message, current Resv is refreshed

Non-disruptive rerouting: switching paths



Resv: allocates labels for both paths
Reserves bandwidth once per link



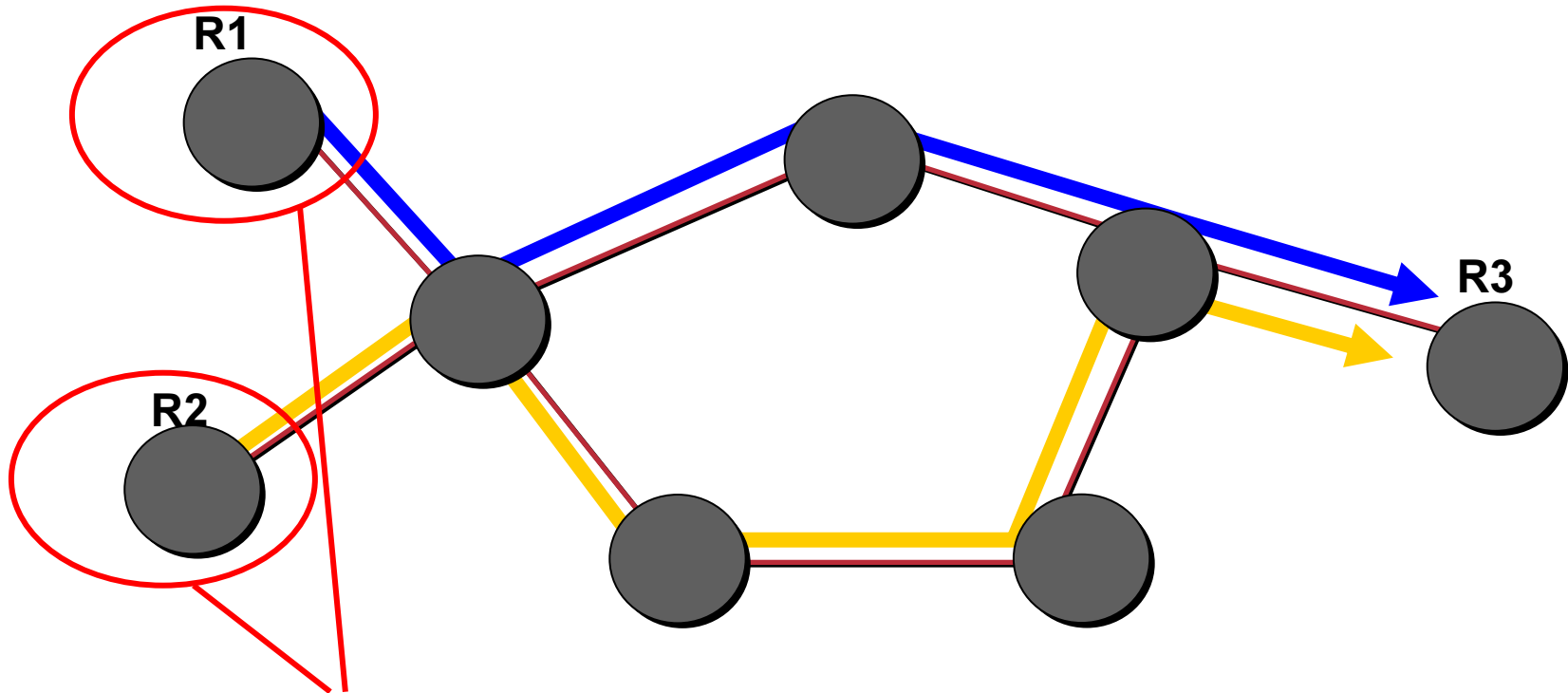
PathTear can then be sent to remove old path (and release resources)

MPLS TE Components

- (1) Resource / policy information distribution
- (2) Constraint based path computation
- (3) RSVP for tunnel signaling
- (4) Link admission control
- (5) LSP establishment
- (6) TE tunnel control and maintenance
- (7) Assign traffic to tunnels**

MPLS TE Components:

(7) Assign traffic to tunnels



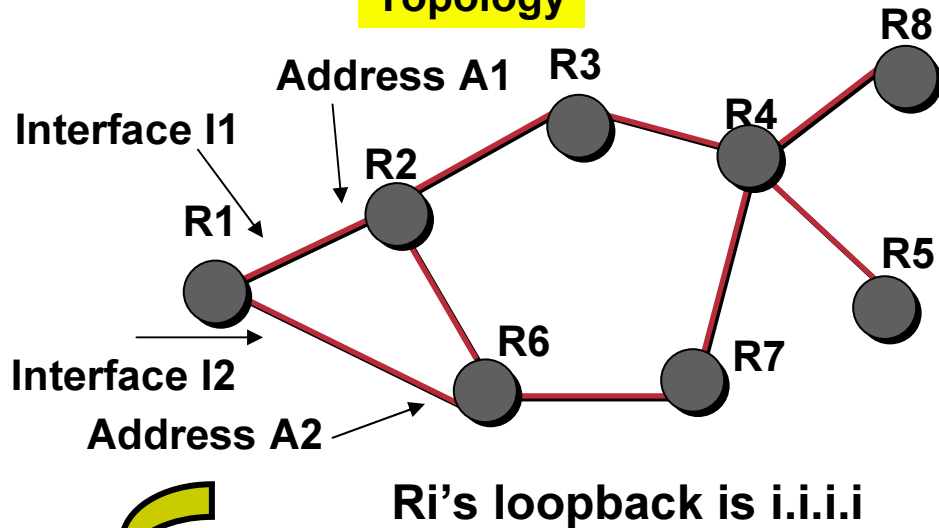
- **Head-end routers assign traffic to tunnels:**
 - Can use static routing
 - Or be integration with IGP by using Autoroute
 - PBR

Modified SPF calculation

- **Automatic assignment based on IGP**
- **at the head-end LSP looks like an interface**
- **when SPF reaches the tail-end of an LSP, the next hop to the tail-end is set to the interface associated with the LSP**
- **destinations whose shortest paths flow via the tail-end will also have the interface associated with the LSP as the next hop**
- **the possible use of a tunnel as (Ointf, NH) does not change the path metric in the routing table!**

Example

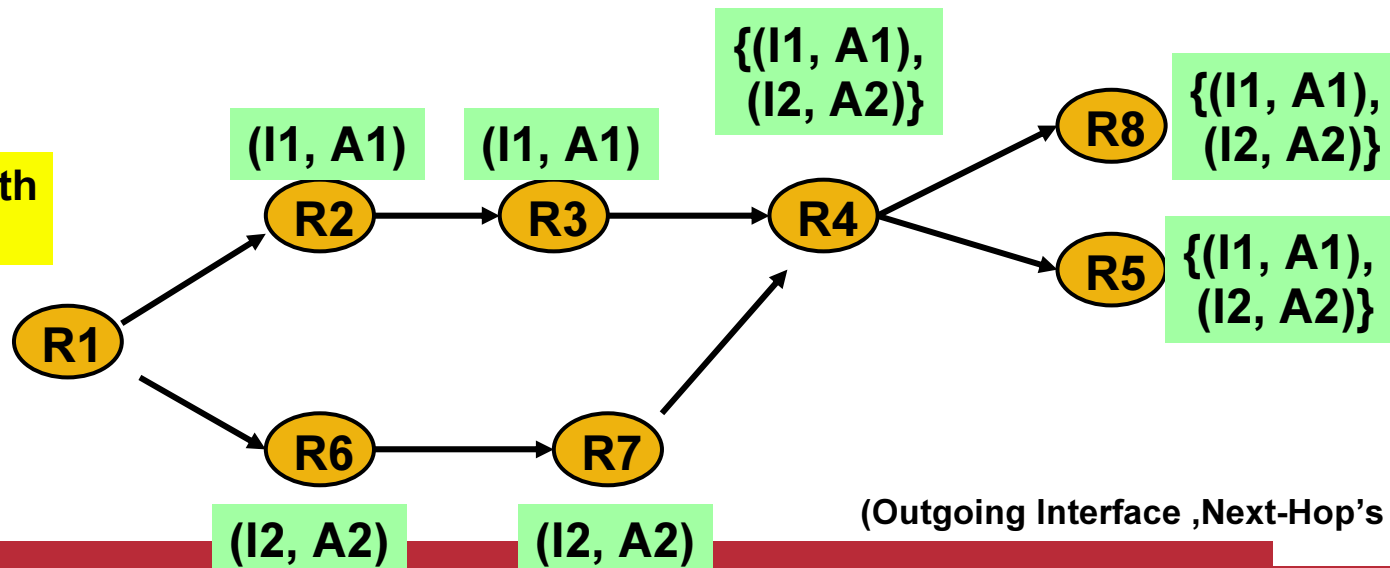
Topology



Routing Table

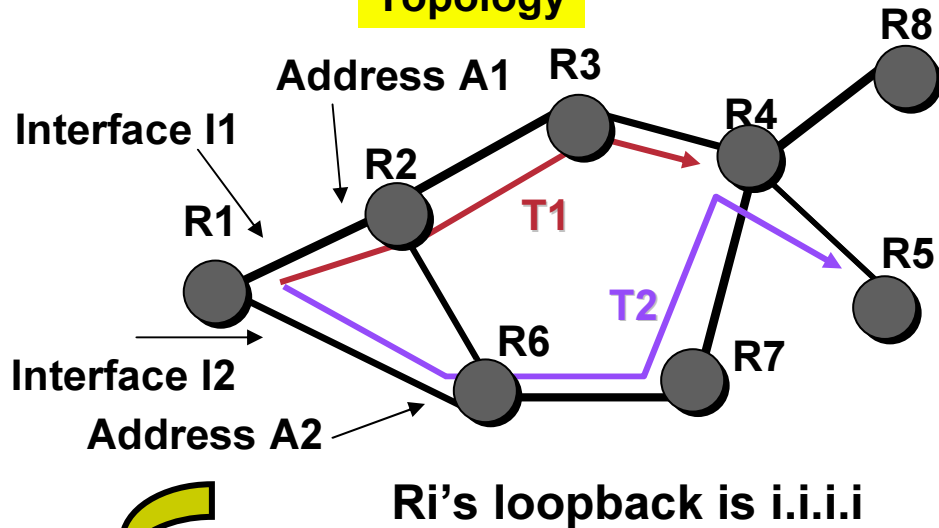
Dest	O Intf	Next Hop	Metric
2.2.2.2	I1	A1	1
3.3.3.3	I1	A1	2
4.4.4.4	I1 I2	A1 A2	3 3
5.5.5.5	I1 I2	A1 A2	4 4
6.6.6.6	I2	A2	1
7.7.7.7	I2	A2	2
8.8.8.8	I1 I2	A1 A2	4 4

Shortest-Path Tree



Example

Topology



Routing Table

Dest	O Intf	Next Hop	Metric
2.2.2.2	I1	A1	1
3.3.3.3	I1	A1	2
4.4.4.4	T1	R4	3
5.5.5.5	T2	R5	4
6.6.6.6	I2	A2	1
7.7.7.7	I2	A2	2
8.8.8.8	T1	R4	4

Shortest-Path Tree

