Ativando MPLS Traffic Engineering

Alexandre Longo – alongo@cisco.com

Cisco Systems

Some Assumptions

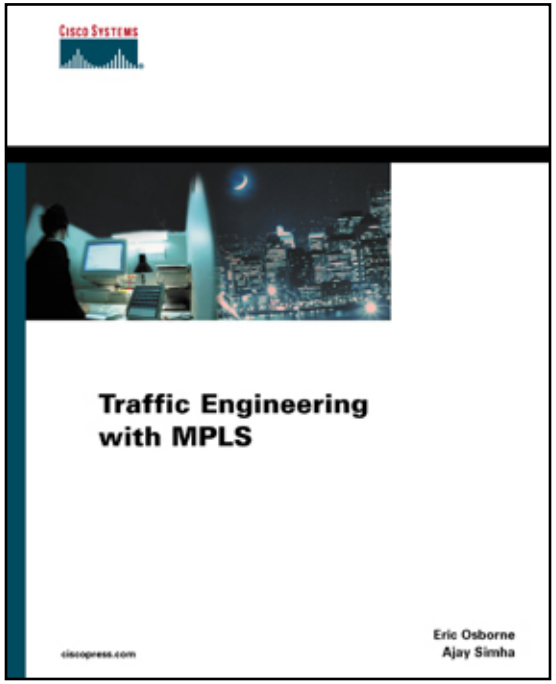
- You understand basic IP routing
- You understand MPLS concepts and operation
- You understand how a link-state protocol works
- Some knowledge of QoS is useful
- You will still be awake at the end of this

A Blatant Plug

- **Traffic Engineering with MPLS**

  ISBN: 1-58705-031-5

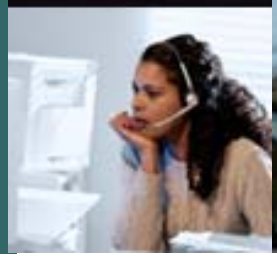
- **Now available in Portuguese and Chinese!**



Agenda

- Traffic Engineering Overview
- Traffic Engineering Theory
- Configuration (*)
- Protection
- Design and Scalability (*)
- Summary

TRAFFIC ENGINEERING OVERVIEW



Network vs. Traffic Engineering

- **Network engineering**
  - **Build your network to carry your predicted traffic**
- **Traffic engineering**
  - **Manipulate your traffic to fit your network**
- **Traffic patterns are impossible to accurately predict**
- **Symmetric bandwidths/topologies, asymmetric load**
- **TE can be done with IGP costs, ATM/FR, or MPLS**

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**

  Override the shortest path selected by the IGP

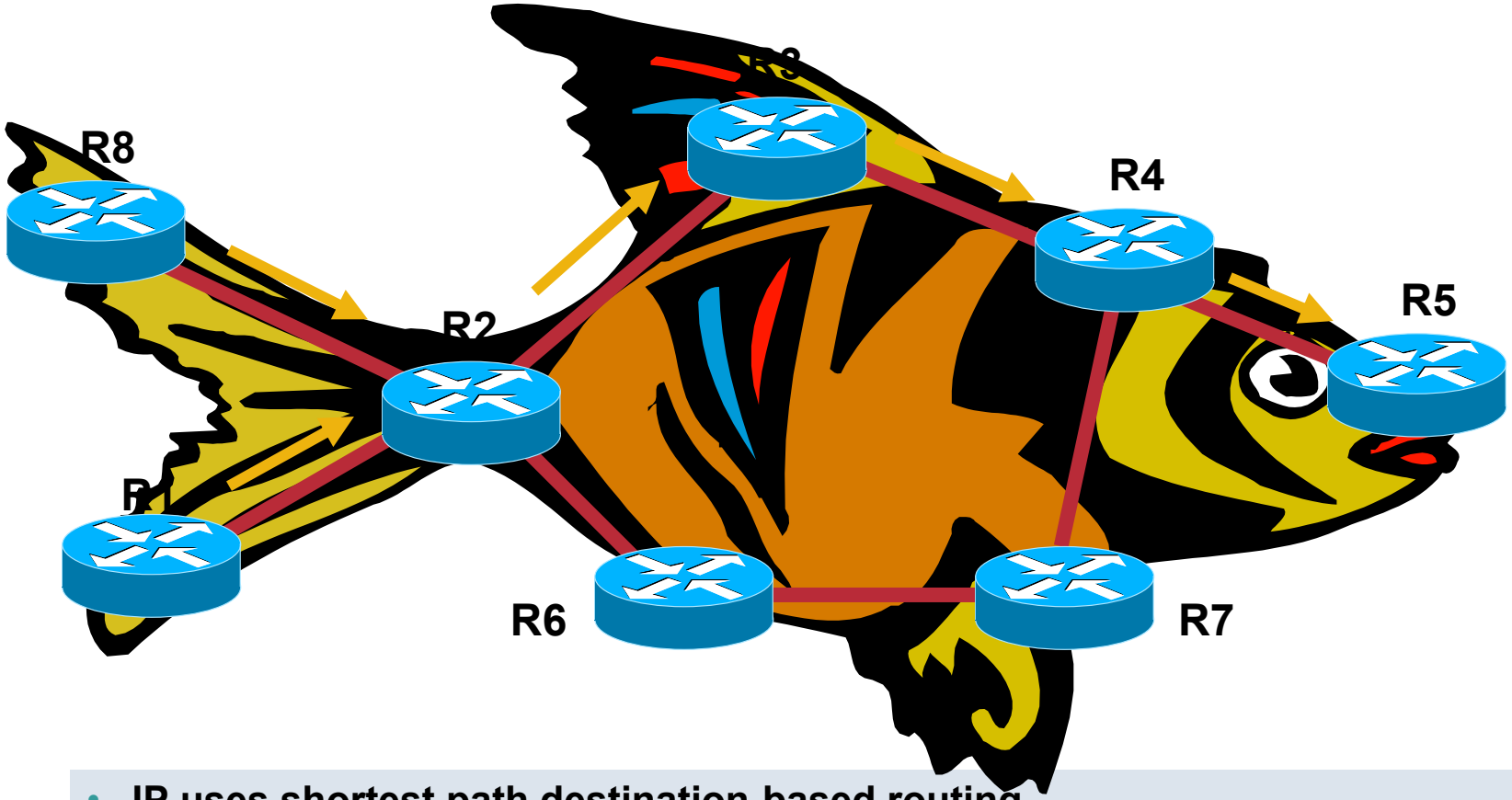
- **Replace ATM/FR cores**

  PVC-like traffic placement without IGP full mesh and associated O(N^2) flooding

- **The ultimate goal is COST SAVING**

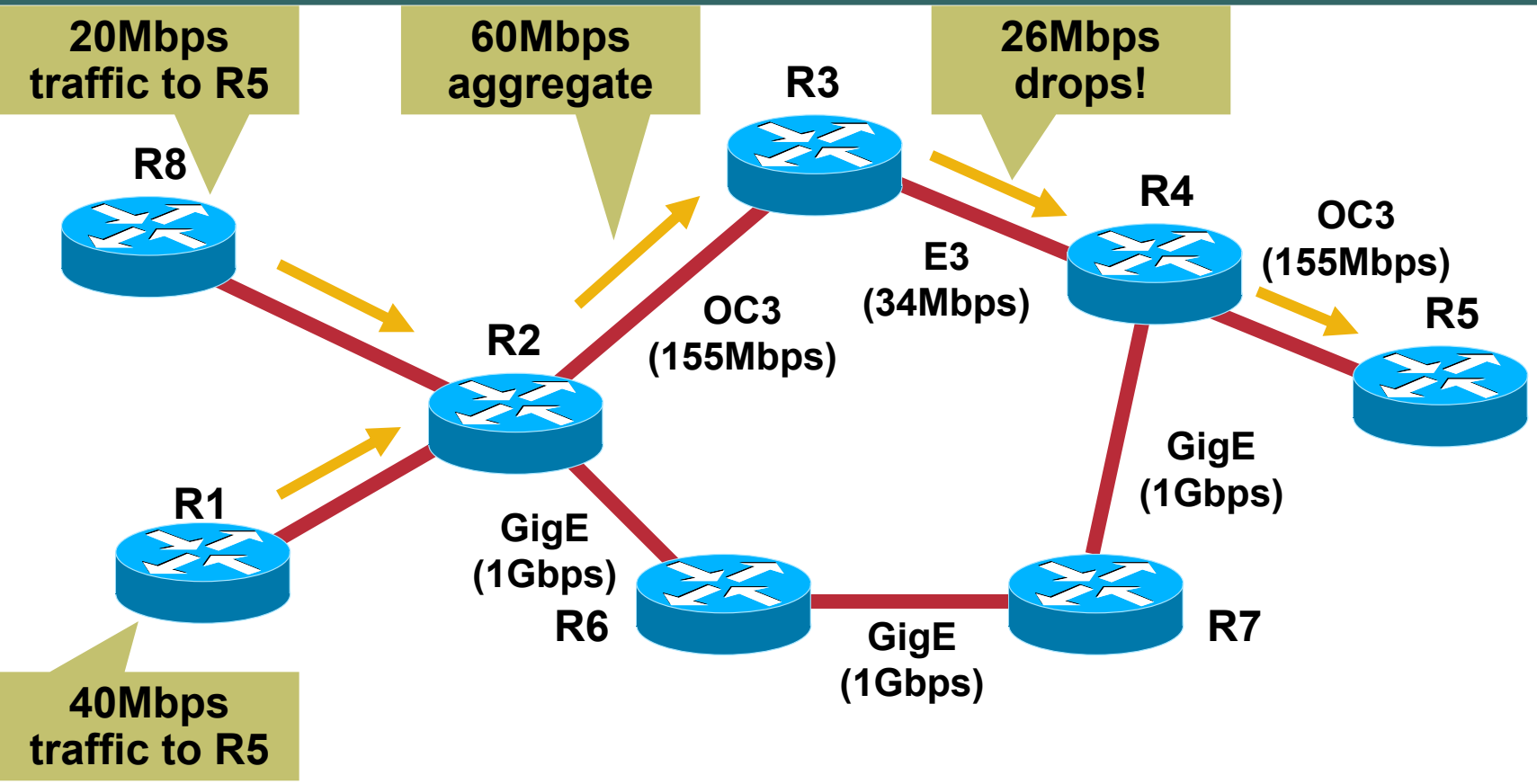
  Service development also progressing

The “Fish” Problem (Shortest Path)

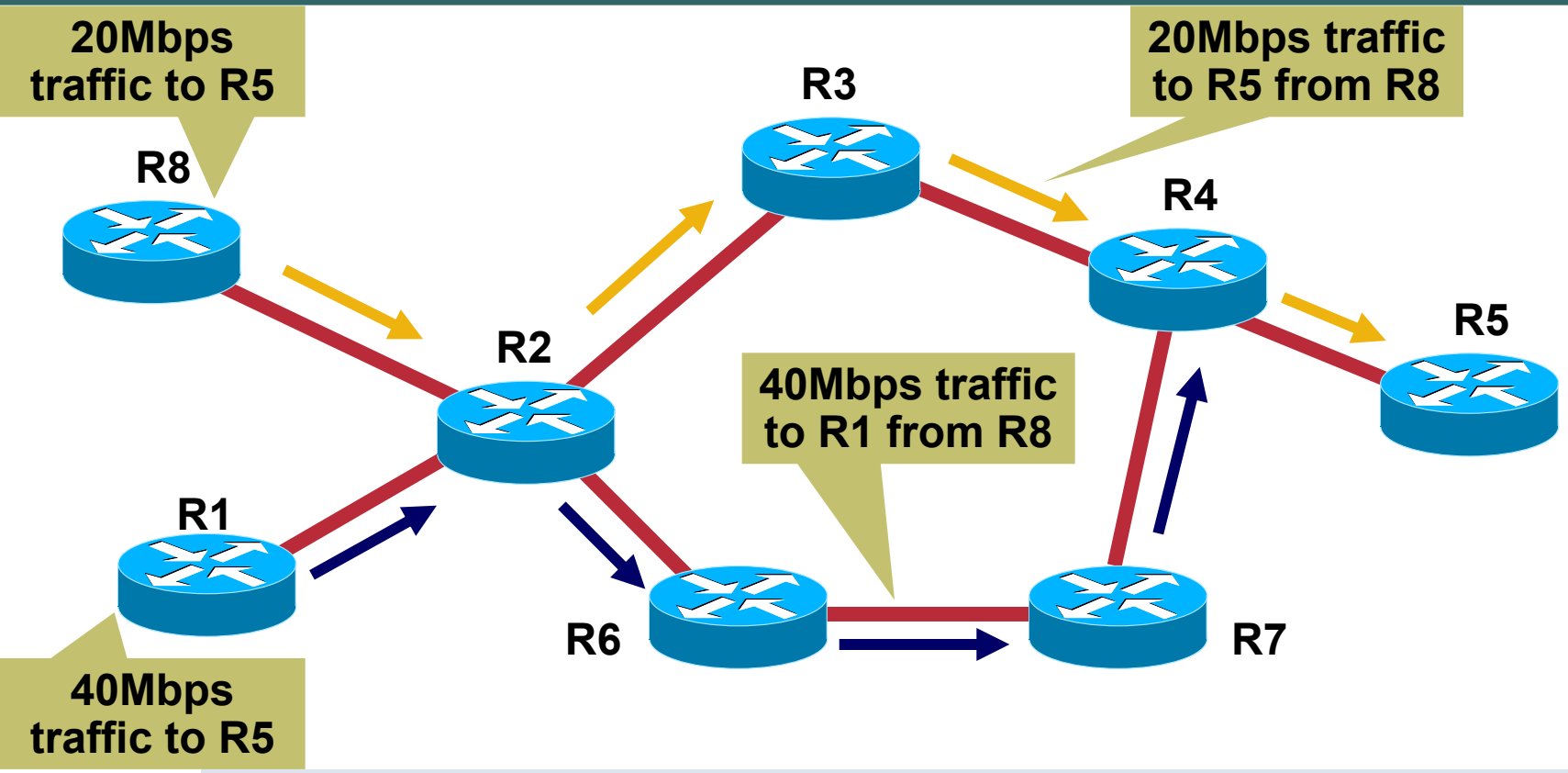


- IP uses shortest path destination-based routing
- Shortest path may not be the only path
- Alternate paths may be under-utilized
- Whilst the shortest path is over-utilized

Shortest Path and Congestion



The TE Solution



- MPLS Labels can be used to engineer explicit paths
- Tunnels are UNI-DIRECTIONAL

Normal path: R8 → R2 → R3 → R4 → R5

Tunnel path: R1 → R2 → R6 → R7 → R4

Terminology

- **Constrained-Based Shortest Path First (CSPF)**

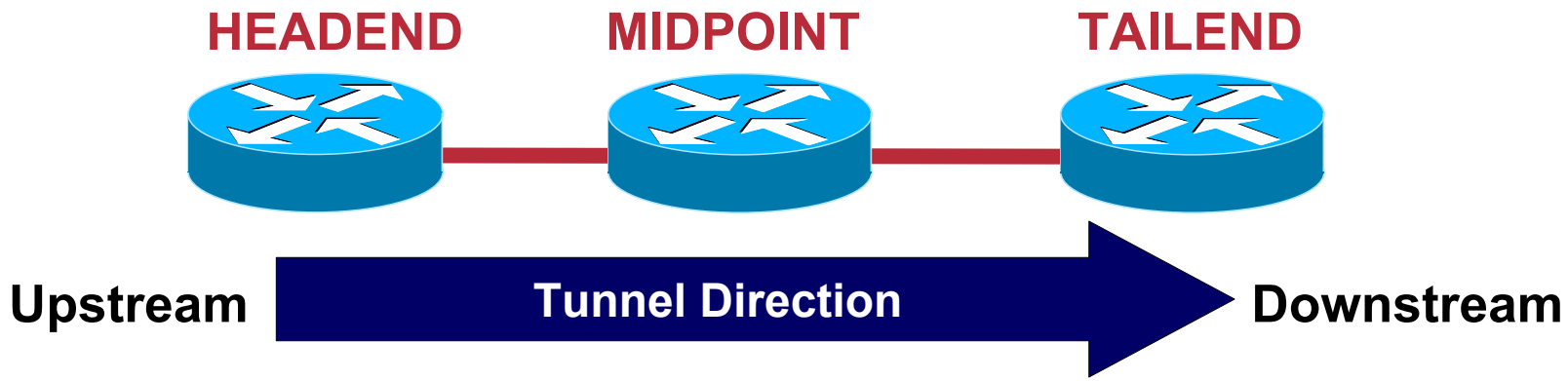
  MPLS-TE uses CSPF to create a shortest path based on a series of constraints:

  Bandwidth

  Affinity/link attributes

  …or an explicitly configured path

- **Tunnels are UNI-DIRECTIONAL!**



TRAFFIC ENGINEERING THEORY

Traffic Engineering Components

- Information distribution
- Path selection/calculation
- Path setup
- Trunk admission control
- Forwarding traffic on to tunnel
- Path maintenance

Information Distribution

- Need to flood TE information (Resource Attributes) across the network
  - Available bandwidth per priority level, a few other things
- IGP extensions flood this information
  - OSPF uses Type 10 (area-local) Opaque LSAs
  - ISIS uses new TLVs
- Basic IGP: {self, neighbors, cost to neighbors}
- TE extensions: {self, neighbors, cost to neighbors, available bandwidth to neighbors}
- TE bandwidth is a control-plane number only

Path Calculation

- Once available bandwidth information and attributes are flooded, router may calculate a path from head to tail
  - Path may be explicitly configured by operator
- TE Headend does a “Constrained SPF” (CSPF) calculation to find the best path
- CSPF is just like regular IGP SPF, except
  - Takes required bandwidth and attributes into account
- Looks for best path from a head to a single tail (unlike OSPF)
- Minimal impact on CPU utilization using CSPF
- Path can also be explicitly configured

Path Setup

- Once the path is calculated, it must be signaled across the network
  - Reserve any bandwidth to avoid “double booking” from other TE reservations
  - Priority can be used to pre-empt low priority existing tunnels
- RSVP used to set up TE LSP
  - PATH messages (from head to tail) carries LABEL_REQUEST
  - RESV messages (from tail to head) carries LABEL
- When RESV reaches headend, tunnel interface = UP
- RSVP messages exist for LSP teardown and error sig

Trunk Admission Control

- **On receipt of PATH message**

  Router will check there is bandwidth available to honour the reservation

  If bandwidth available then RSVP accepted

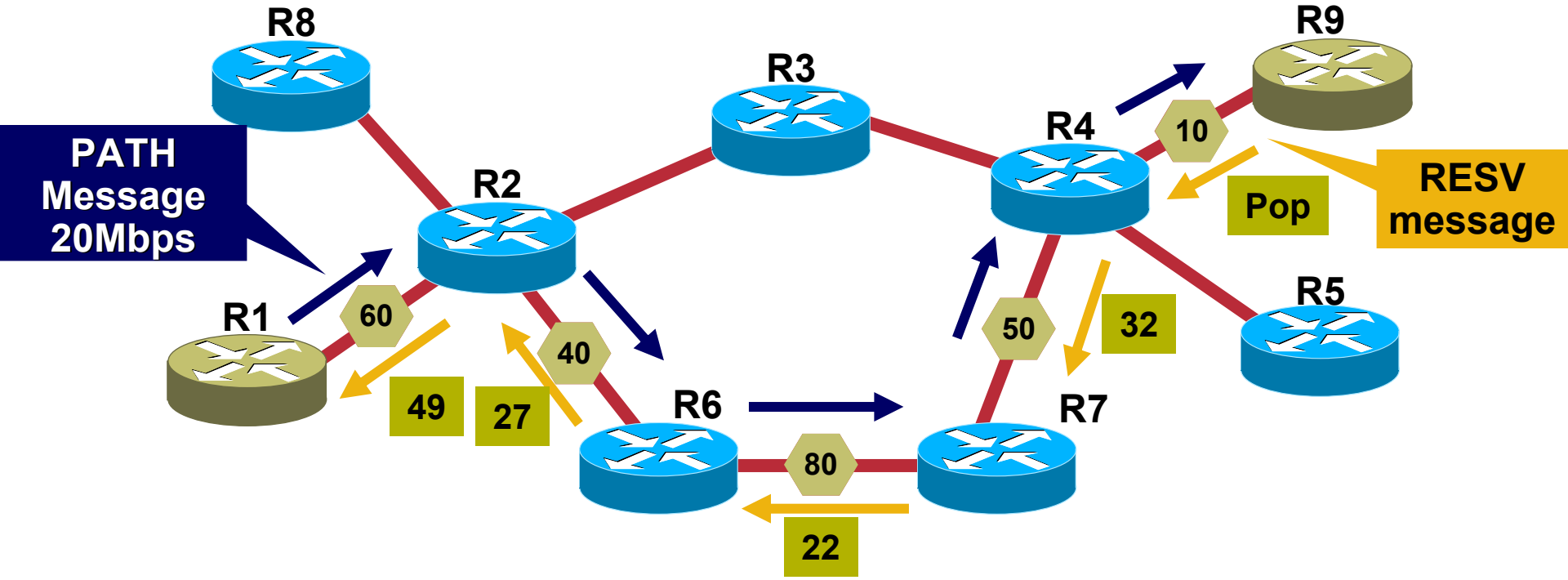
- **On receipt of a RESV message**

  Router actually reserves the bandwidth for the TE LSP

  If pre-emption is required lower priority LSP are torn down

- **OSPF/ISIS updates are triggered**

Path Setup Example



RSVP PATH: R1 → R2 → R6 → R7 → R4 → R9

RSVP RESV: Returns labels and reserves bandwidth on each link

80 Bandwidth available

49 Returned label via RESV message

Forwarding Traffic to a Tunnel

- Static routing
- Policy routing
  - Global table only—not from VRF at present
- Autoroute
- Forwarding Adjacency

Static, autoroute, and forwarding adjacency get you unequal-cost load-balancing

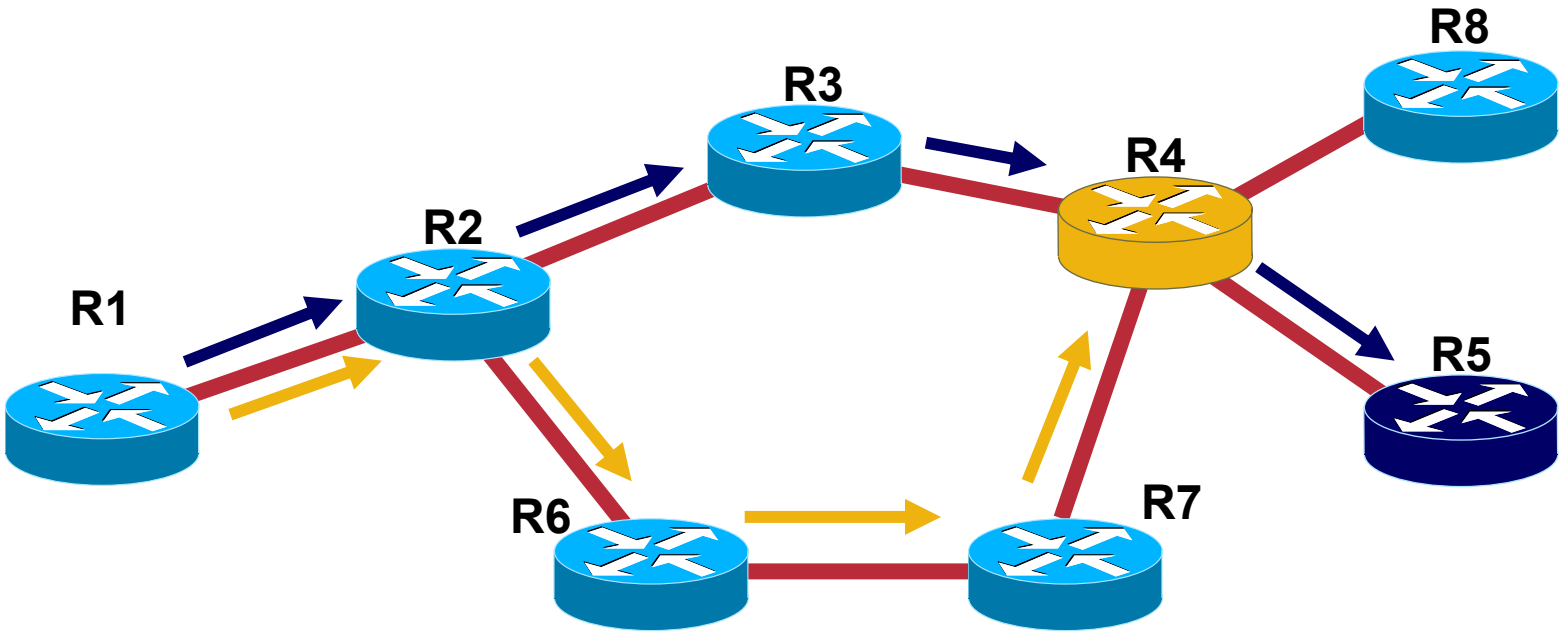
Autoroute

- Used to include TE LSP in SPF calculations
- IGP adjacency is NOT run over the tunnel!
- Tunnel is treated as a directly connected link to the tail

  When tunnel tail is seen in PATH list during IGP SPF, replace outgoing physical interface with tunnel interface

  Inherit tunnel to all downstream neighbors of said tail

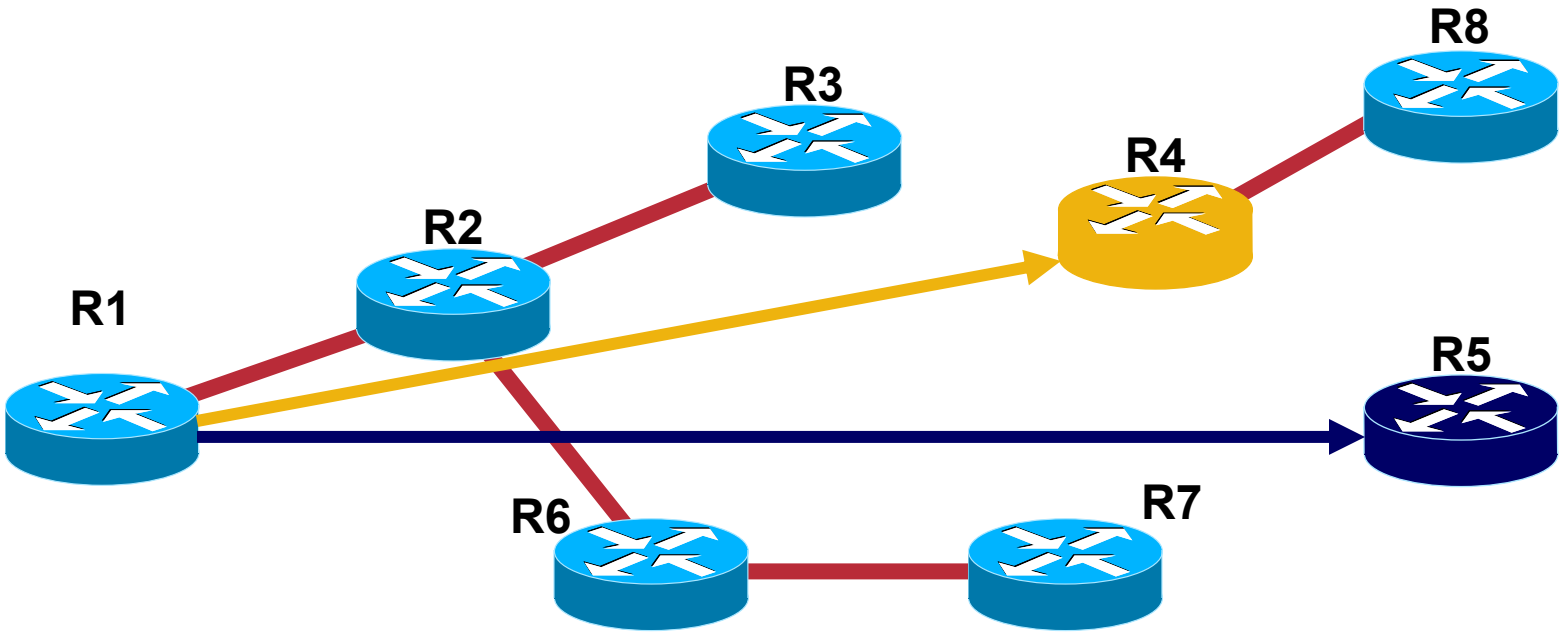
Autoroute Topology (OSPF and ISIS)



Tunnel1: R1 → R2 → R3 → R4 → R5

Tunnel2: R1 → R6 → R7 → R4

Autoroute Topology (OSPF and ISIS)



From R1 router perspective:

Next hop to R4 and R8 is Tunnel1

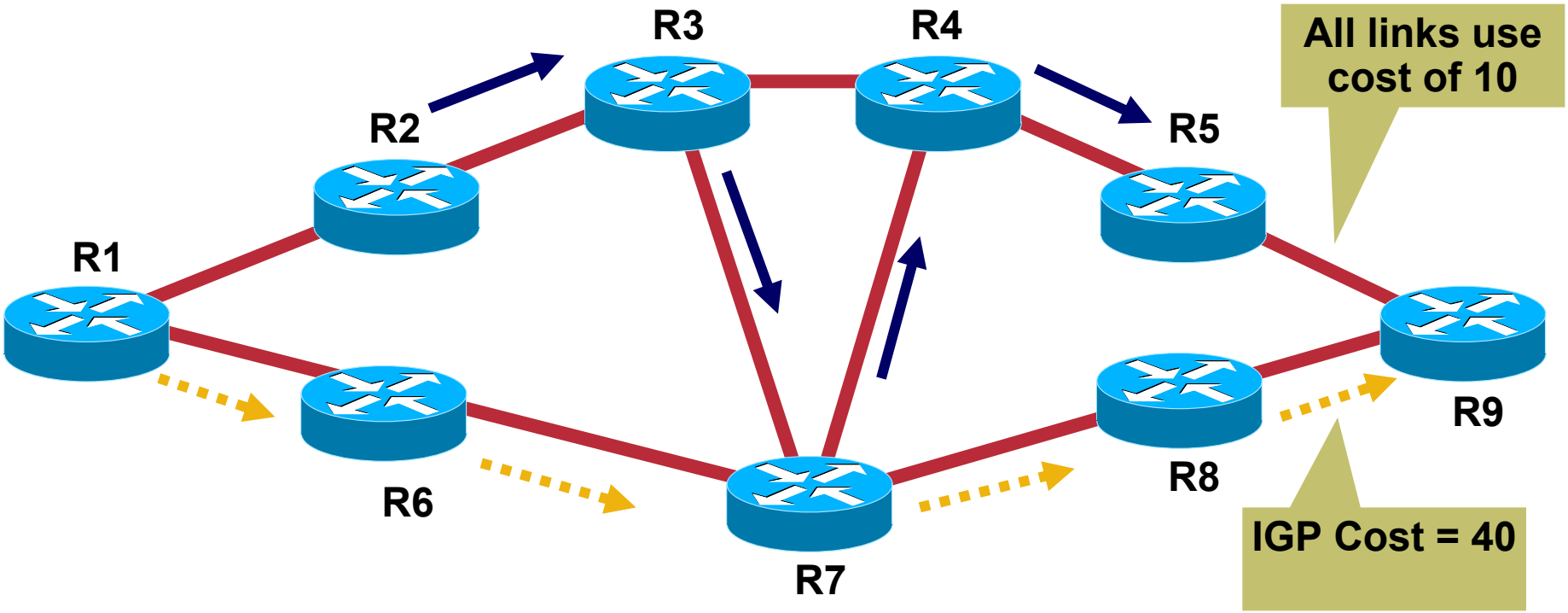
Next hop to R5 is Tunnel2

All nodes behind tunnel routed via tunnel

Forwarding Adjacency

- Autoroute does not advertise the LSP into the IGP
- There may be a requirement to advertise the existence of TE tunnels to upstream routers
  - Like an ATM/FR PVC—attract traffic to a router regardless of the cost of the underlying physical network cost
- Useful as a drop-in replacement for ATM/FR (and during migration)
- Can get suboptimal forwarding (NOT loops) if you're not careful

Forwarding Adjacency

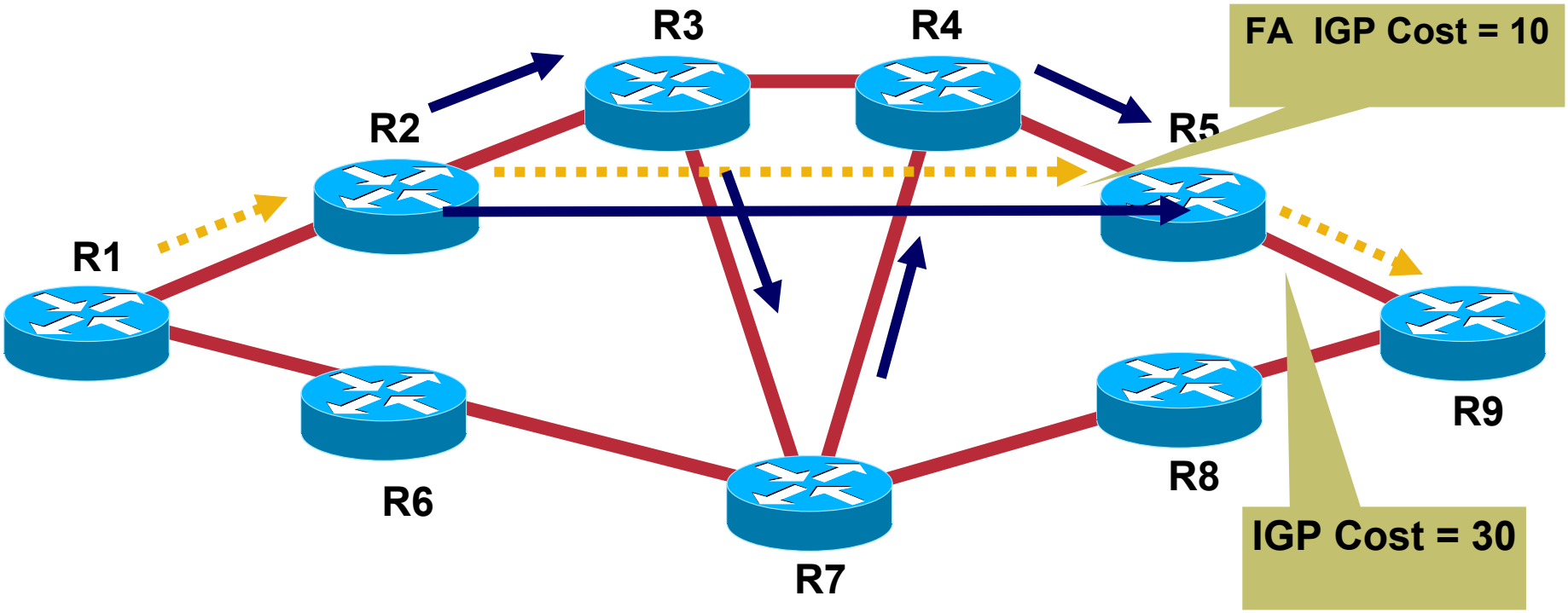


Tunnel: R2 → R3 → R7 → R4 → R5

R1 shortest path to R9 via IGP

Tunnel at R2 is never used as R1 can't see it

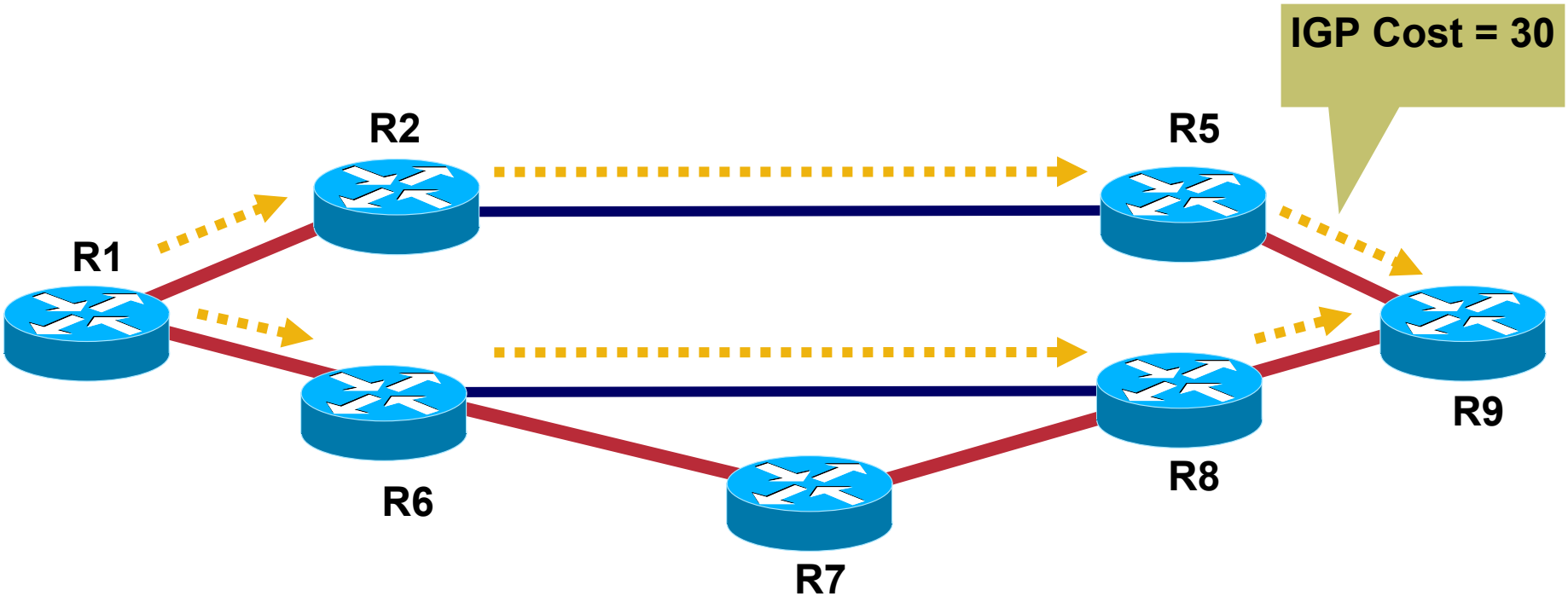
Advertise TE Links into IGP



Tunnel: R2 → R3 → R4 → R5

R1 shortest path to R9

Load Balancing Across FA



Tunnel: R2 ➔R3 ➔ R4 ➔ R5

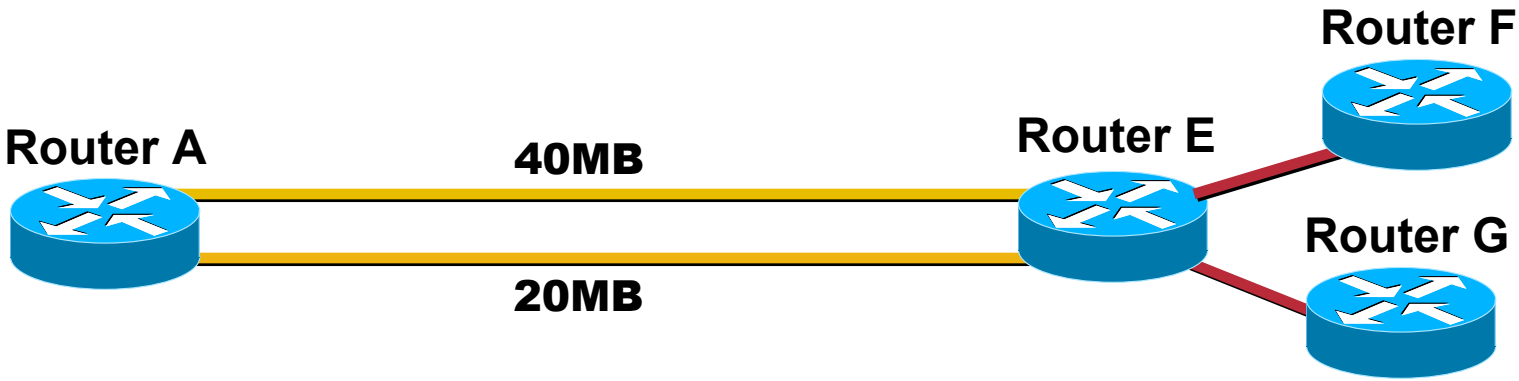
R1 shortest path to R9

Unequal Cost Load Balancing

- IP routing has equal-cost load balancing, but not unequal cost*
- Unequal cost load balancing difficult to do while guaranteeing a loop-free topology
- Since MPLS doesn't forward based on IP header, permanent routing loops don't happen
- 16 hash buckets for next-hop, shared in rough proportion to configured tunnel bandwidth or load-share value

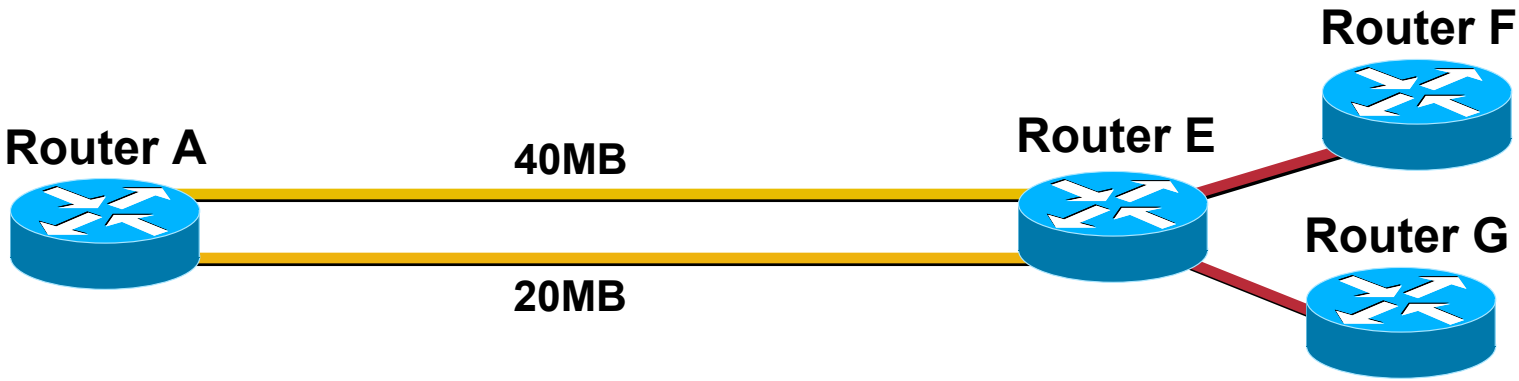
*EIGRP Has 'Variance', but That's Not as Flexible

Unequal Cost: Example 1




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

Unequal Cost: Example 1




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

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

Path Maintenance

- **Steady-state information load is low**
  - **Especially with refresh reduction (RFC2961)**
- **Path re-optimization**
  - **Process where some traffic trunks are rerouted to new paths so as to improve the overall efficiency in bandwidth utilization**
  - **For example, traffic may be moved to secondary path during failure; when primary path is restored traffic moved back**
- **Path restoration**
  - **Comprised of two techniques; local protection (link and node) and path protection**
  - **Discussed later in protection section**

CONFIGURATION

Prerequisite Configuration (Global)


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

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
```

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)
```

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
```

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
```

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]
```

Multiple Path Calculations

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

```
tunnel mpls traffic-eng path-option 10 explicit name foo
tunnel mpls traffic-eng path-option 20 explicit name bar
tunnel mpls traffic-eng path-option 30 dynamic
```

- Path-options can each have their own bandwidth

```
tunnel mpls traffic-eng path-option 10 explicit name foo
     bandwidth 100
tunnel mpls traffic-eng path-option 20 explicit name bar
     bandwidth 50
tunnel mpls traffic-eng path-option 30 dynamic
     bandwidth 0
```

LSP Attributes

Configure on Tunnel:

```
tunnel mpls traffic-eng path-
  option 10 dynamic attributes
  foo
```

Attribute list 'foo' is defined at:

```
mpls traffic-eng lsp
  attributes foo
 bandwidth 25
 priority 2 2
```

- **Attribute list options**

```
affinity
auto-bw
bandwidth
lockdown
priority
protection
record-route
```

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
```

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)
```

Summary Configuration (1/2)


```
ip cef (distributed}
mpls traffic-eng tunnels
interface Tunnel0
  tunnel mode mpls traffic-eng
  ip unnumbered Loopback0
  tunnel destination RID-of-tail
  tunnel mpls traffic-eng autoroute announce
  tunnel mpls traffic-eng path-option 10 dynamic
```

Summary Configuration (2/2)


```
! Configure in IGP
mpls traffic-eng tunnels
mpls traffic-eng router-id Loopback0
mpls traffic-eng area ospf-area (OSPF)
mpls traffic-eng level-x (ISIS)
metric-style wide
!
! On Physical interface
interface POS0/0
  mpls traffic-eng tunnels
  ip rsvp bandwidth Kbps
```

PROTECTION

Protection

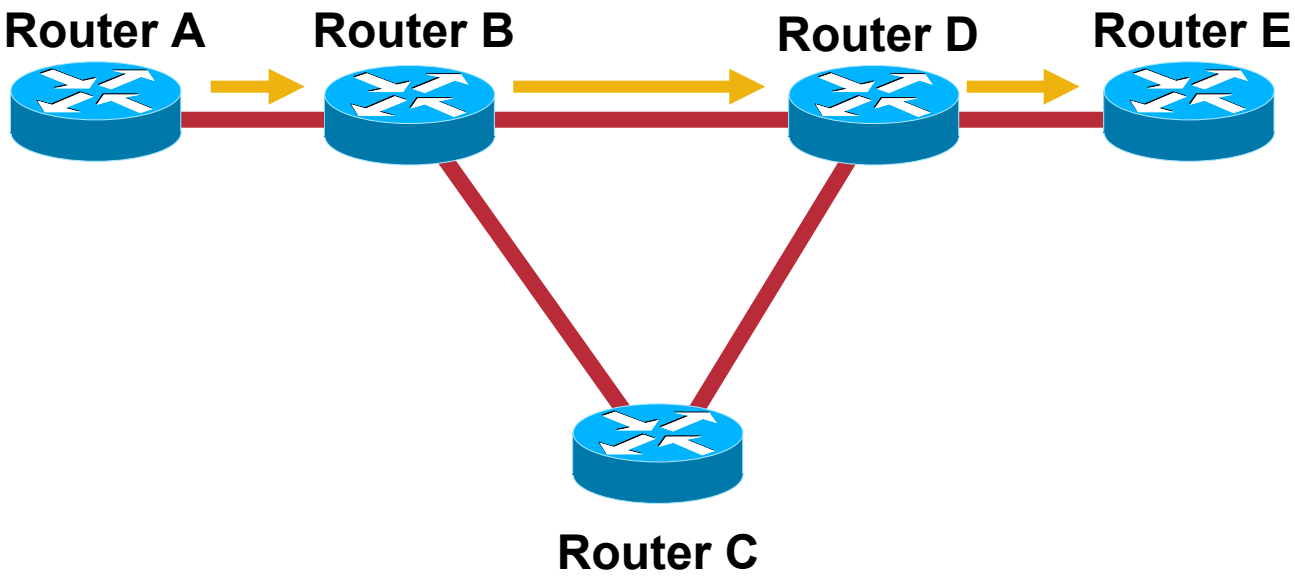
- Mechanism to minimize packet loss during a failure
- Pre-provisioned protection tunnels that carry traffic when a protected link or node goes down
- MPLS TE protection also known as FAST REROUTE (FRR)
- FRR protects against LINK FAILURE

  For example, Fibre cut, Carrier Loss, ADM failure
- FRR protects against NODE FAILURE

  For example, power failure, hardware crash, maintenance
- Fast failure recovery due to local repair

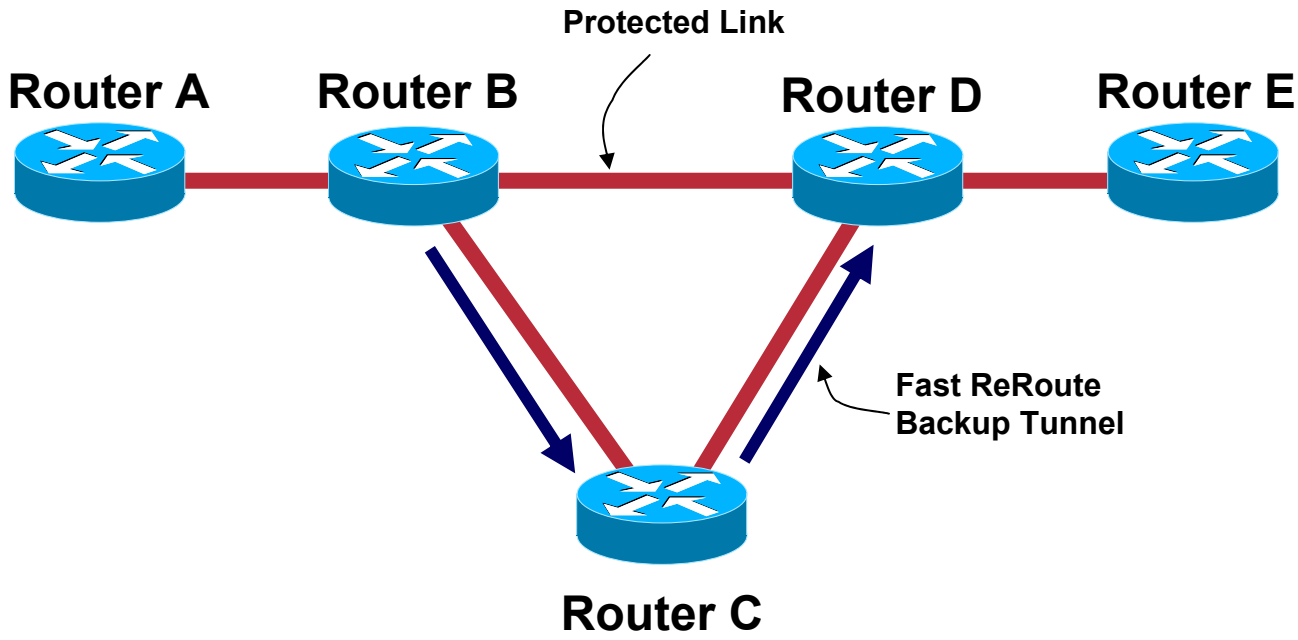
Link Protection

- TE Tunnel A ➔ B ➔ D ➔ E



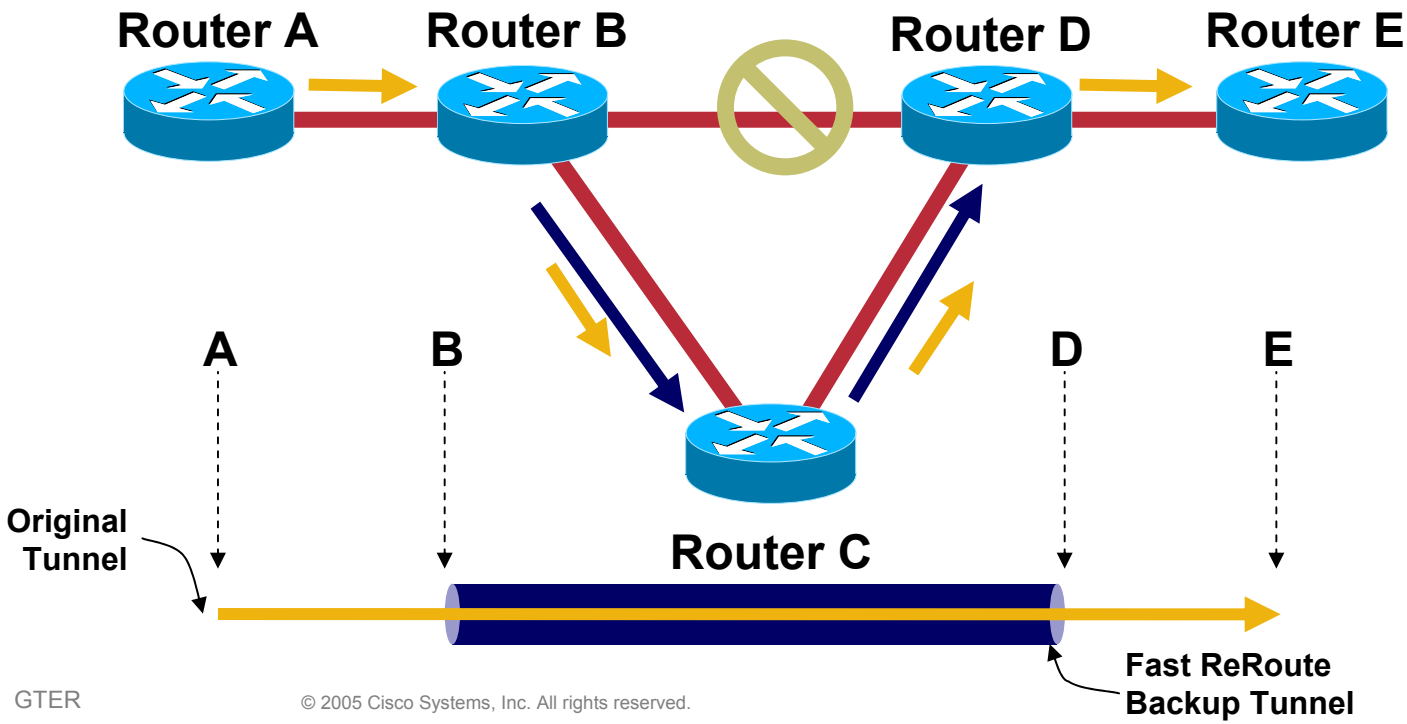
Link Protection

- B has a pre-provisioned backup tunnel to the other end of the protected link (Router D) B ➔ C ➔ D
- FRR relies on the fact that D is using global label space

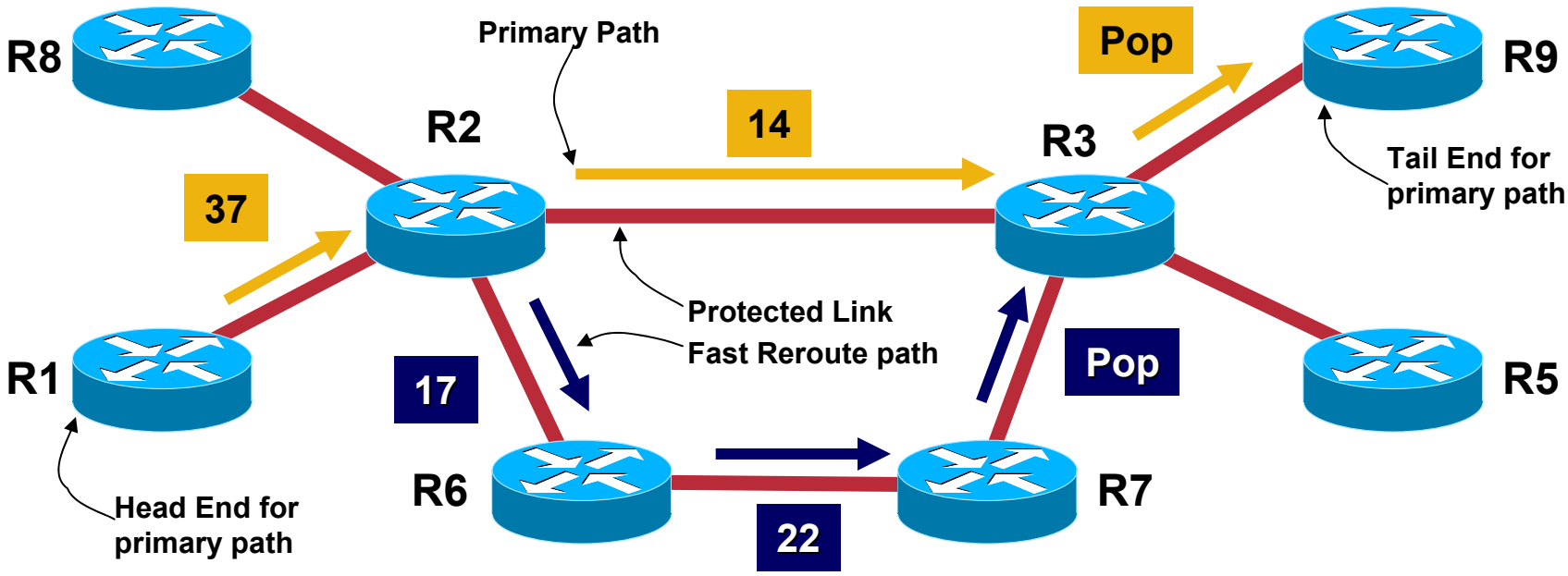


Link Protection

- When B ➔ D link fails, A ➔ E tunnel is encapsulated in B ➔ D tunnel
- Backup tunnel is used until A can re-compute tunnel path as A ➔ B ➔ C ➔ D ➔ E (~5-15 seconds or so)



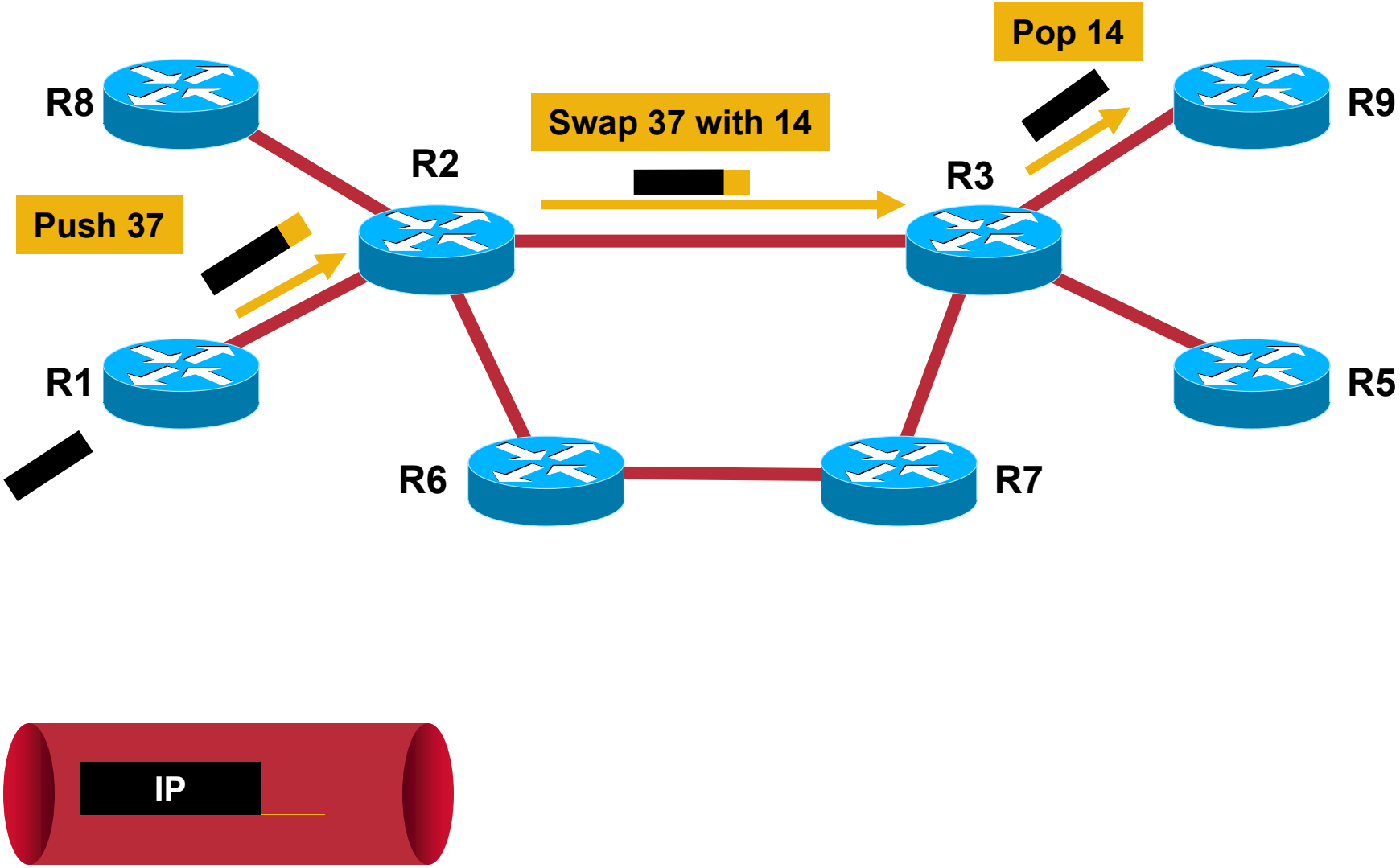
Link Protection Example



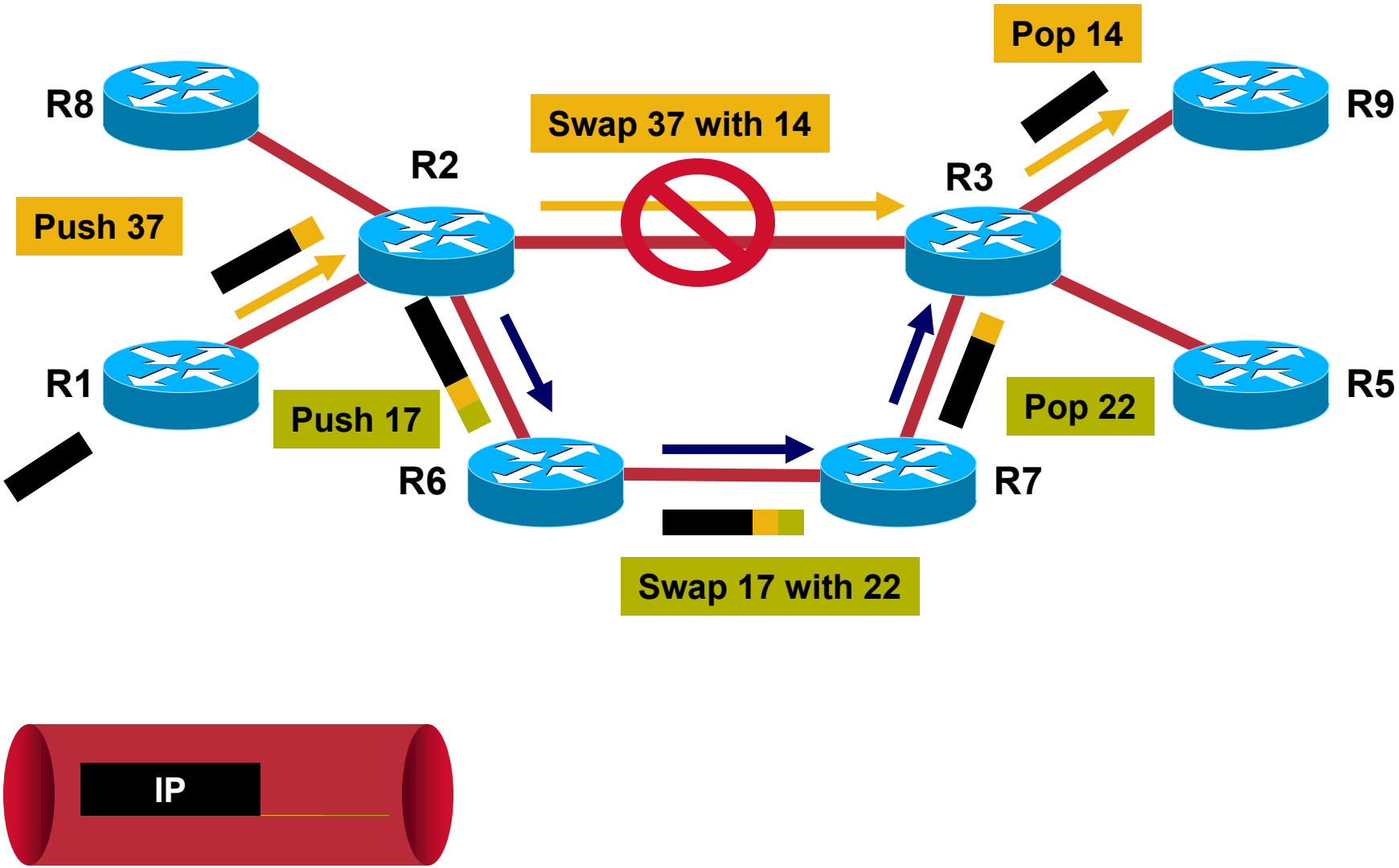
Primary path: R1 → R2 → R3 → R9

Fast Reroute path: R2 → R6 → R7 → R3

Normal TE Operation

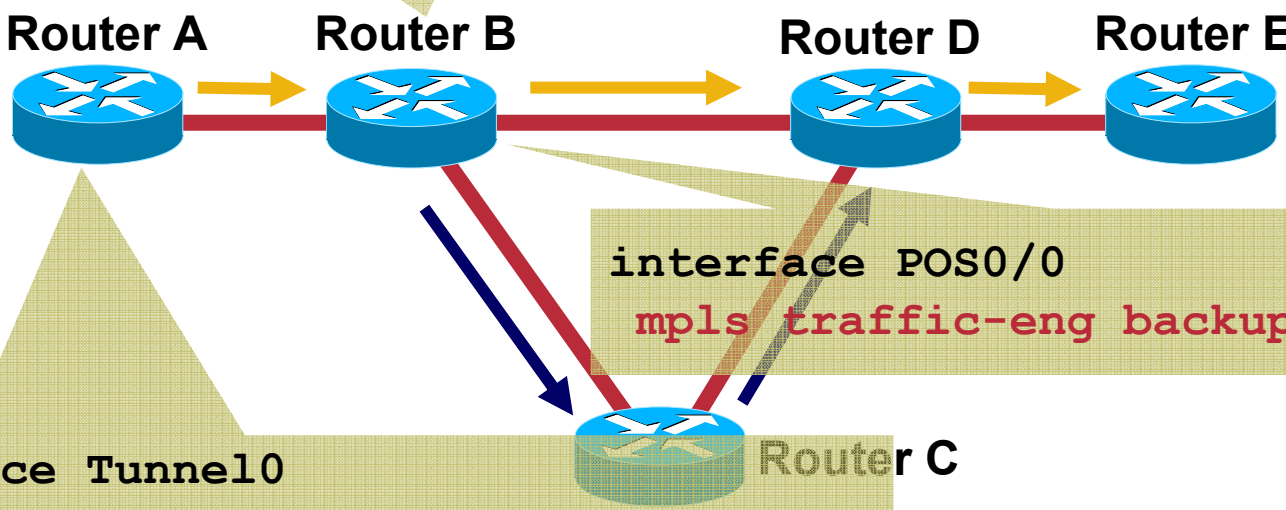


Fast Reroute Link Failure



Link Protection Configuration


```
interface Tunnel0
 tunnel destination Router D
 … explicit-path R2-R3-R4
 no tunnel mpls traffic-eng autoroute announce
```

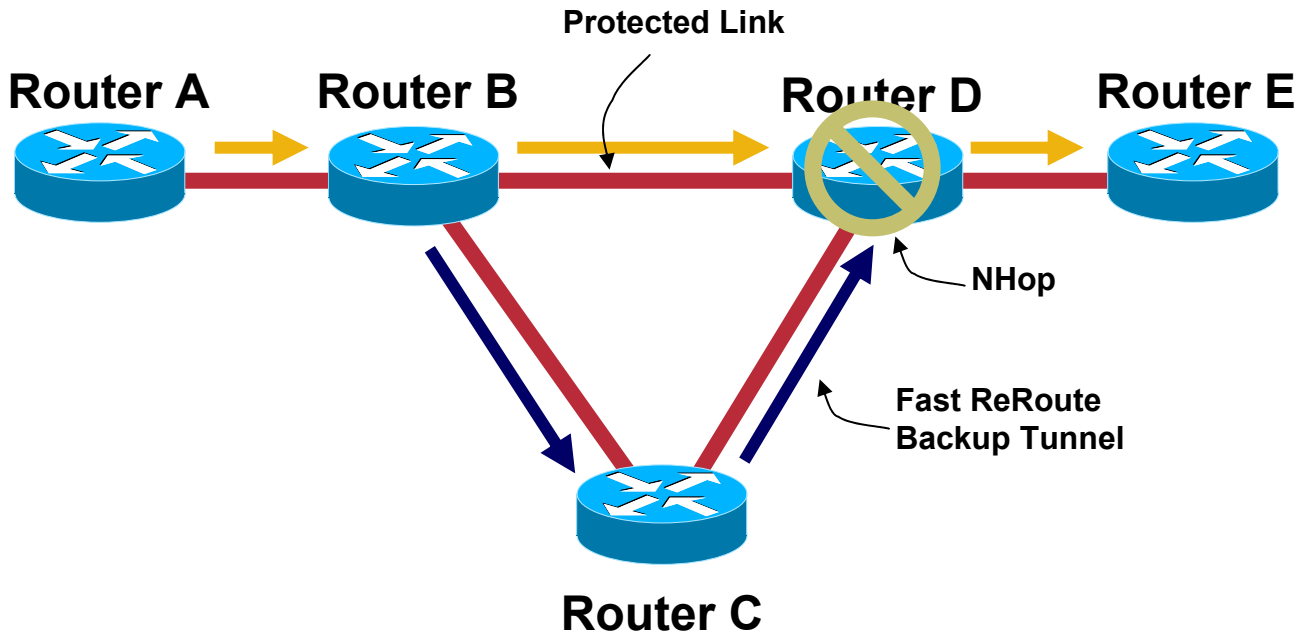



```
interface POS0/0
 mpls traffic-eng backup-path Tunnel0
```

```
interface Tunnel0
 tunnel destination Router E
 .. etc ...
 tunnel mpls traffic-eng fast-reroute
```

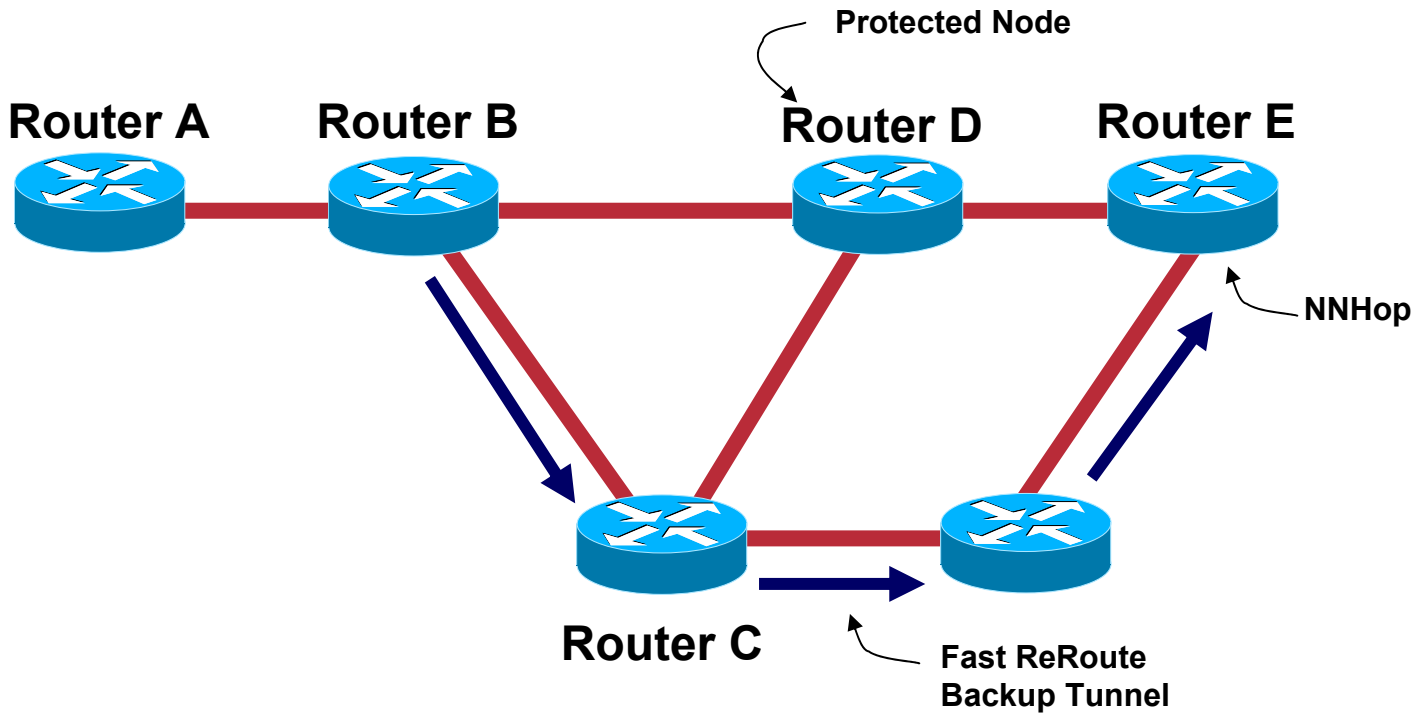
Node Protection

- 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)



Node Protection

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



Node Protection

- Node protection still has the same convergence properties 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)

Link and Node Protection Times

- Link and Node protection are very similar
- Protection times are commonly linear to number of protected items
- One nationwide provider gets ~35ms of loss
- New code on GSR E3 linecards gets a prefix-independent 2ms-4ms loss

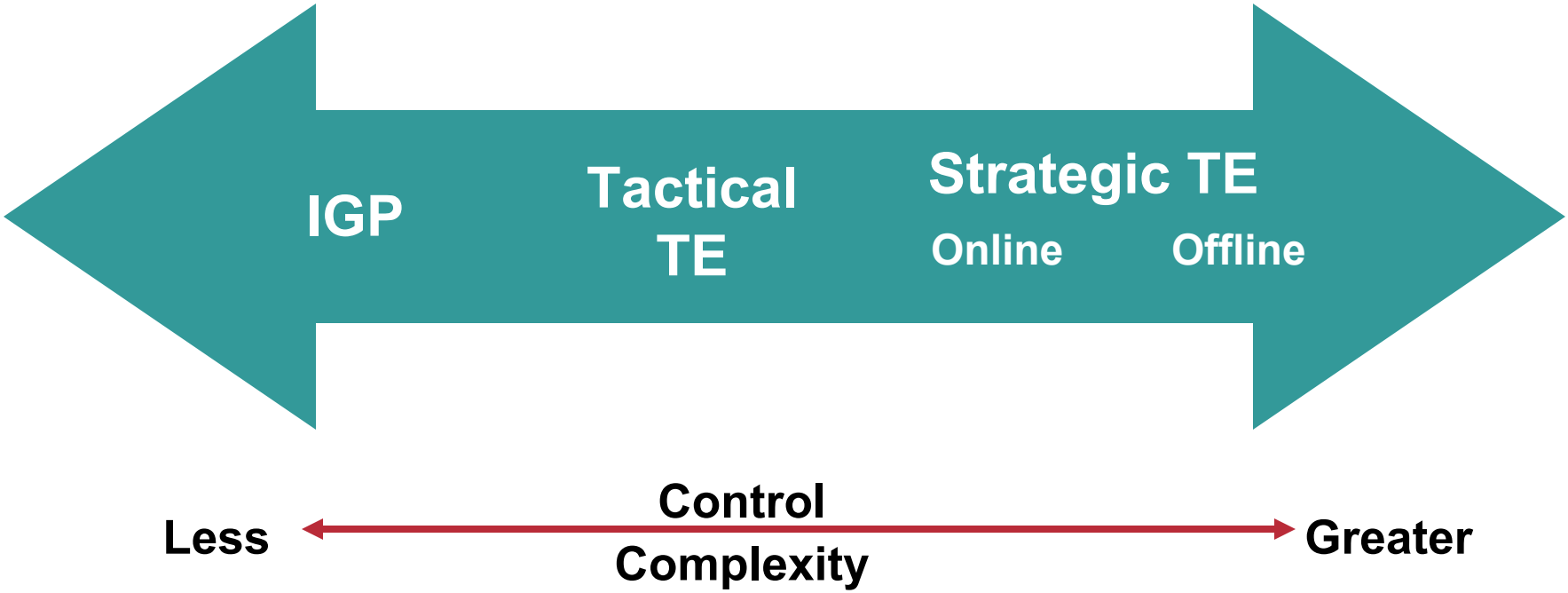
DESIGN AND SCALABILITY

Design Approach and Scalability

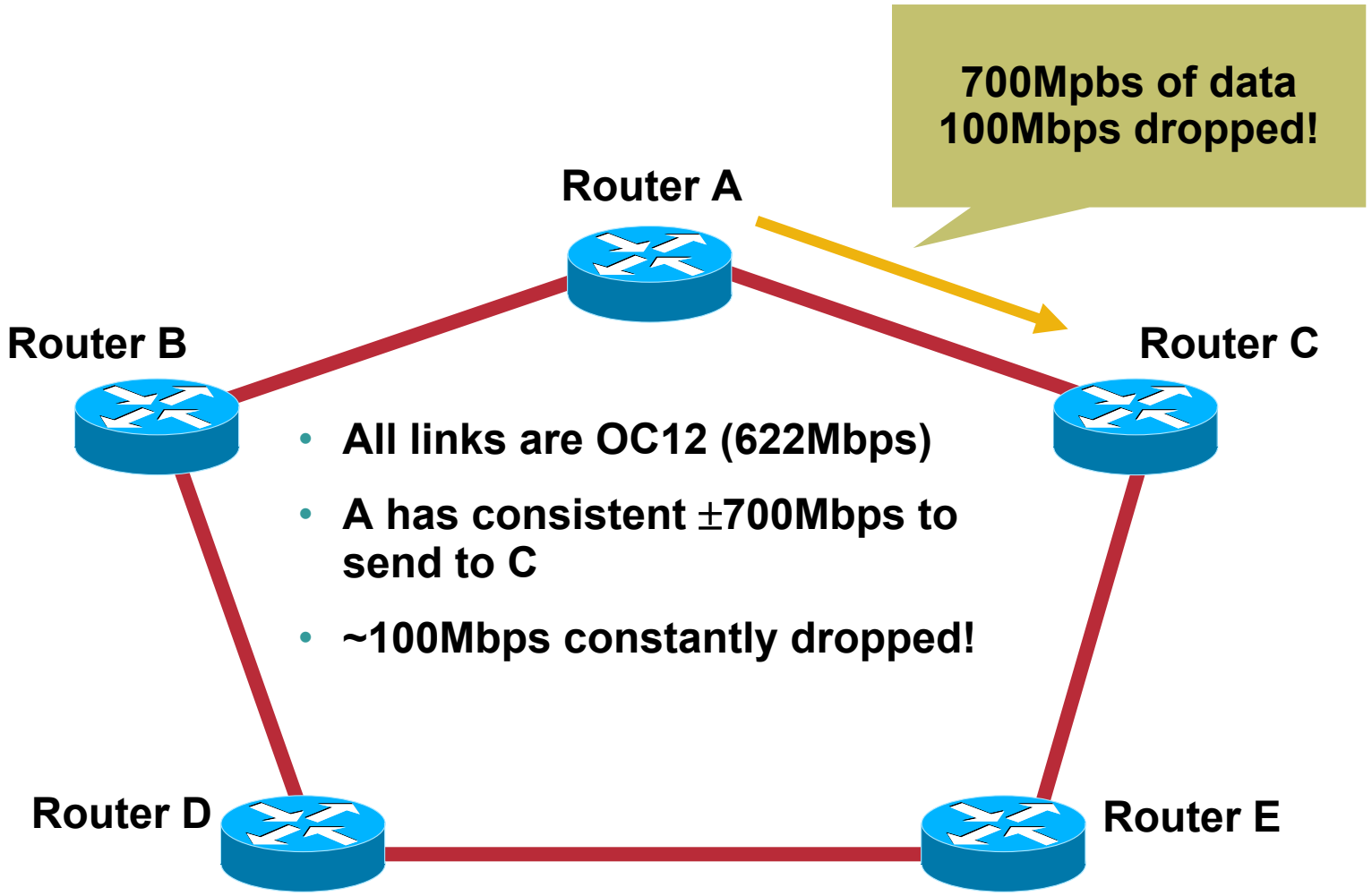
Two Methods to Deploy MPLS-TE

- **Tactical**
  - As needed to clear up congestion
  - You only have tunnels when there is a problem (and you must remember to remove them
- **Strategic**
  - Mesh of TE tunnels between a level of routers
  - Typically P to P but can be PE to PE in smaller networks
  - N(N-1) LSPs (one in each direction)

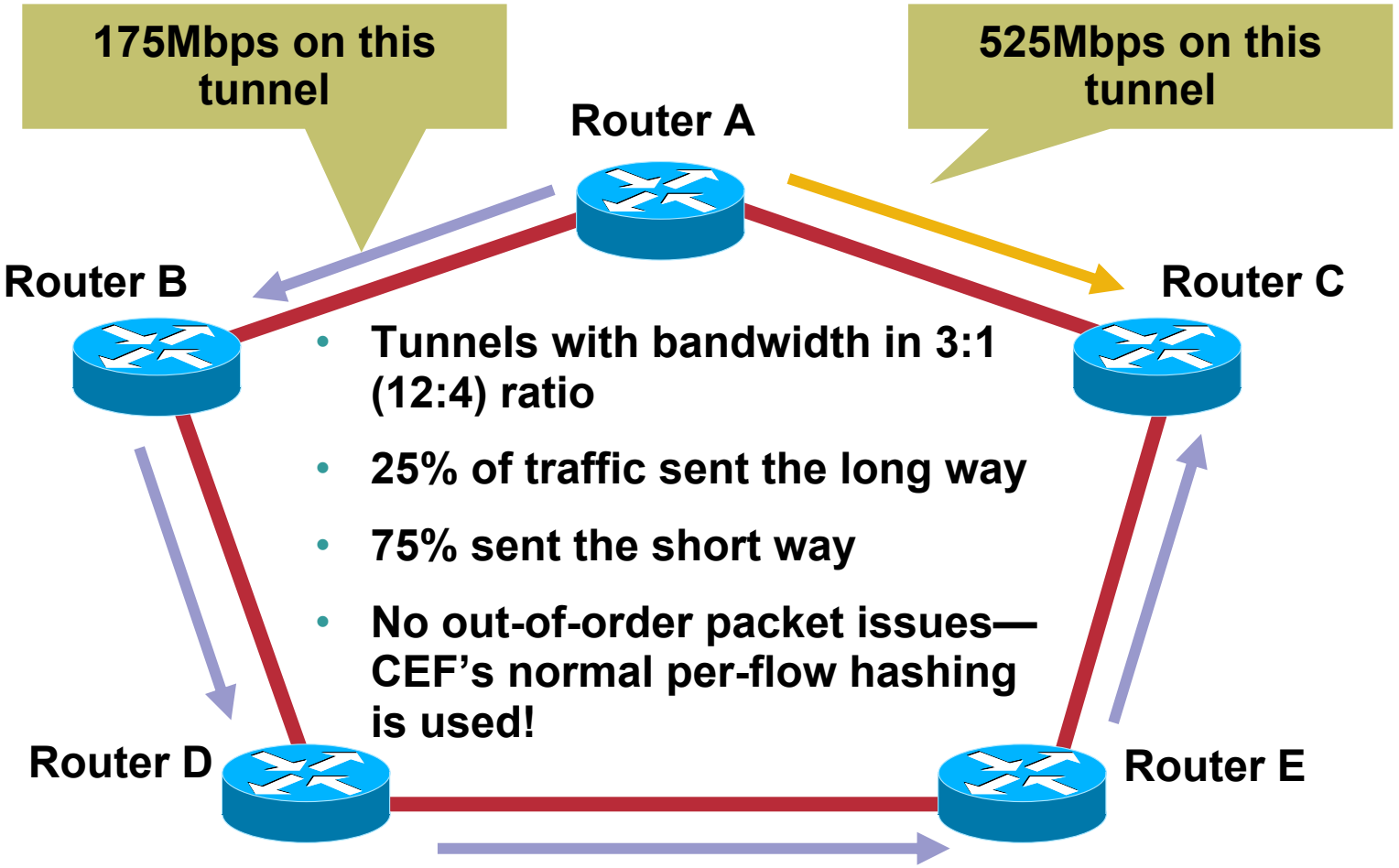
Design Spectrum



Tactical: Large ISP Case Study



Multiple TE and Unequal Cost Load Balancing



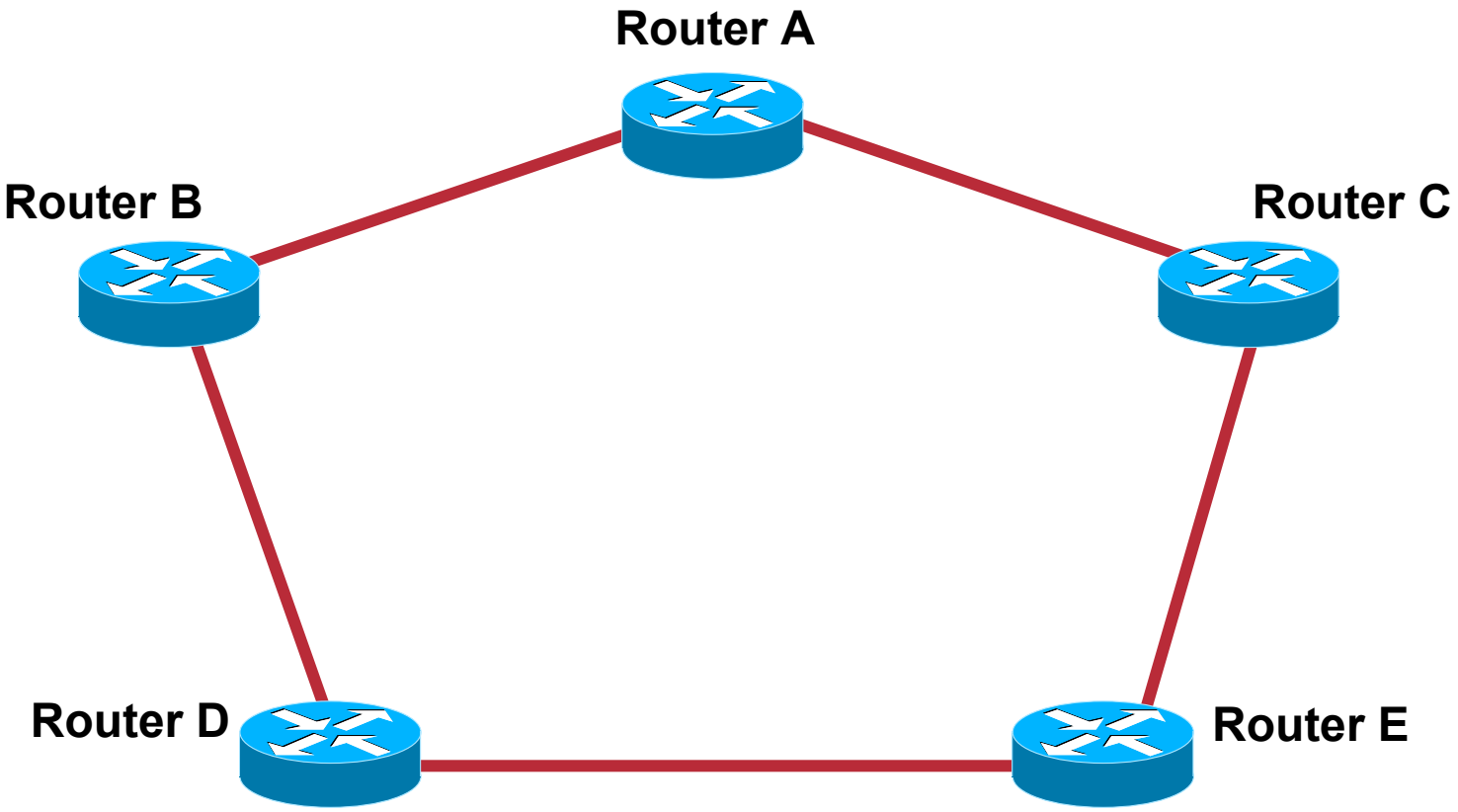
Tactical

- As needed—Easy, quick, but hard to track over time
- Easy to forget why a tunnel is in place
- Inter-node BW requirements may change, tunnels may be working around issues that no longer exist

Strategic

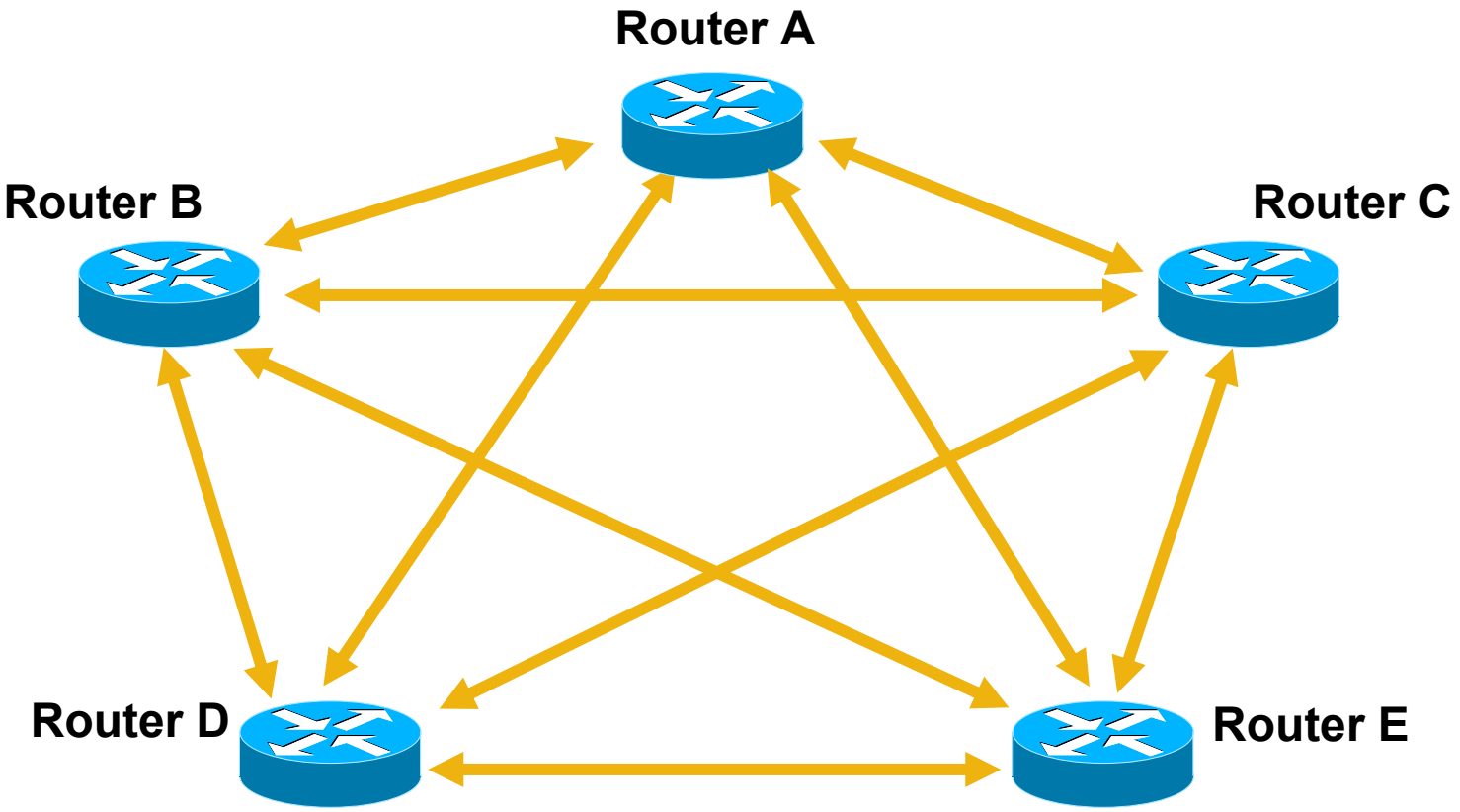
- Full mesh of TE tunnels between routers
- Initially deploy tunnels with 0 bandwidth
- Monitor tunnel interface statistics
  - ~Bandwidth used between router pairs
    - TE tunnels have interface MIBs
    - Make sure that Σtunnel <= Σnetwork BW
- As tunnel bandwidth is changed, tunnels will find the best path across the network

Strategic: Physical Topology



Strategic: Logical Topology

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



Strategic

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

SUMMARY

MPLS TE

- Helps optimize network utilization (strategic)
- Assists in handling unexpected congestion (tactical)
- Provides fast reroute for link and node failures
- TE is only part of a method of guaranteeing bandwidth
  - It is a control plane mechanism only
  - Must be used with traditional QoS mechanisms

OBRIGADO !!