MPLS-TE: Fundamentals and Advanced Fast Reroute

Andrew Gasson

agasson@cisco.com

Agenda

- How MPLS-TE Works
- Design Guidelines
- Fast ReRoute

TE Basics

- Information Distribution
- Path Calculation
- Path Setup
- Forwarding Traffic Down Tunnels

Information Distribution

- **OSPF**

  Uses type 10 (opaque area—local) ISAs

- **ISIS**

  Uses Type 22 TLVs

TE Basics

- Information Distribution
- Path Calculation
- Path Setup
- Forwarding Traffic Down Tunnels

Path Calculation

Cisco.com

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

  Constrained SPF

- **…or PCALC (path calculation)**

Path Calculation

Cisco.com

- **What if there's more than one path that meets the minimum requirements (bandwidth, etc.)?**
- **PCALC algorithm:**

  Find all paths with the lowest IGP cost

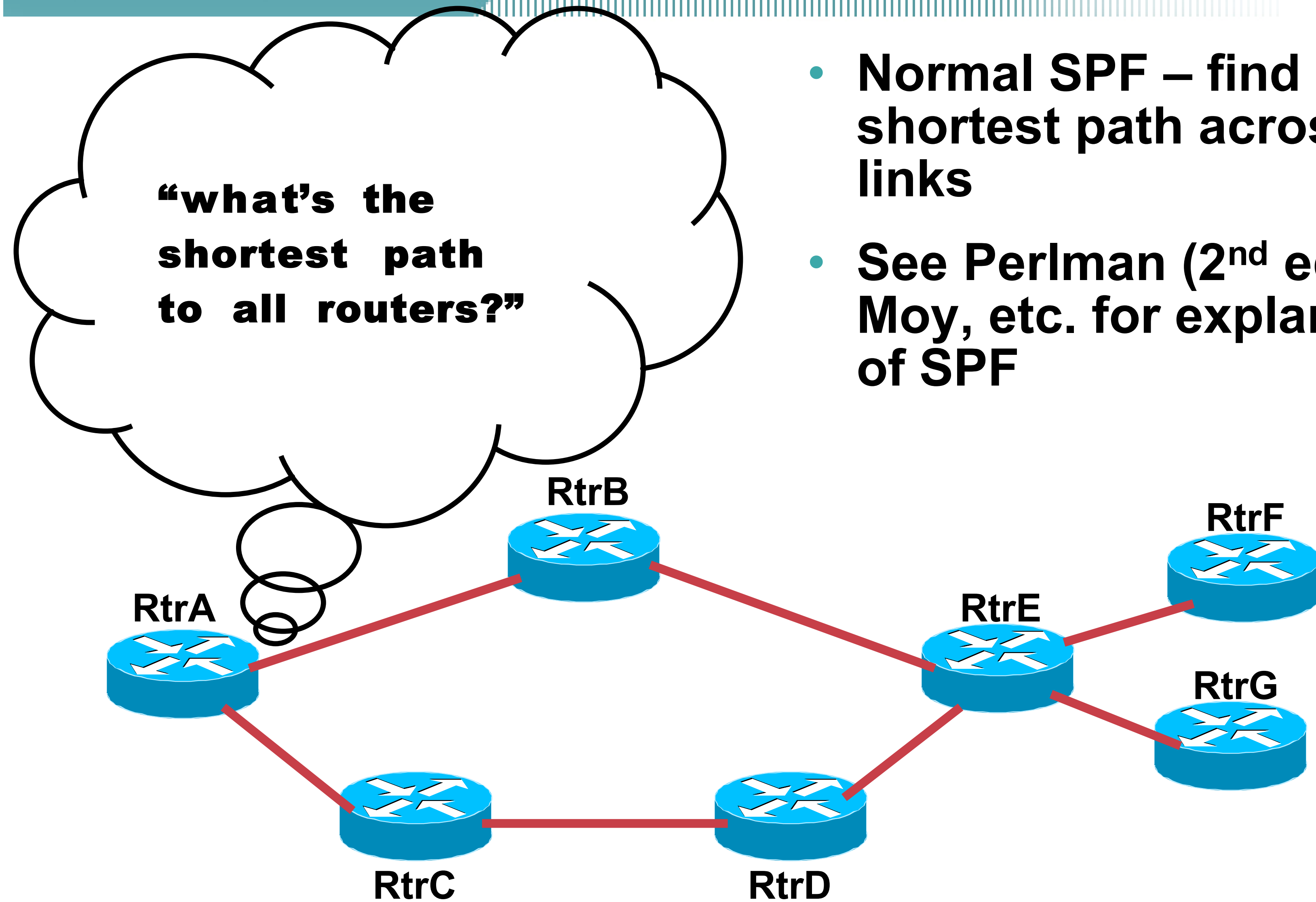
  Then pick the path with the highest minimum bandwidth along the path

  Then pick the path with the lowest hop count (not IGP cost, but hop count)

  Then just pick one path at random

Path Calculation

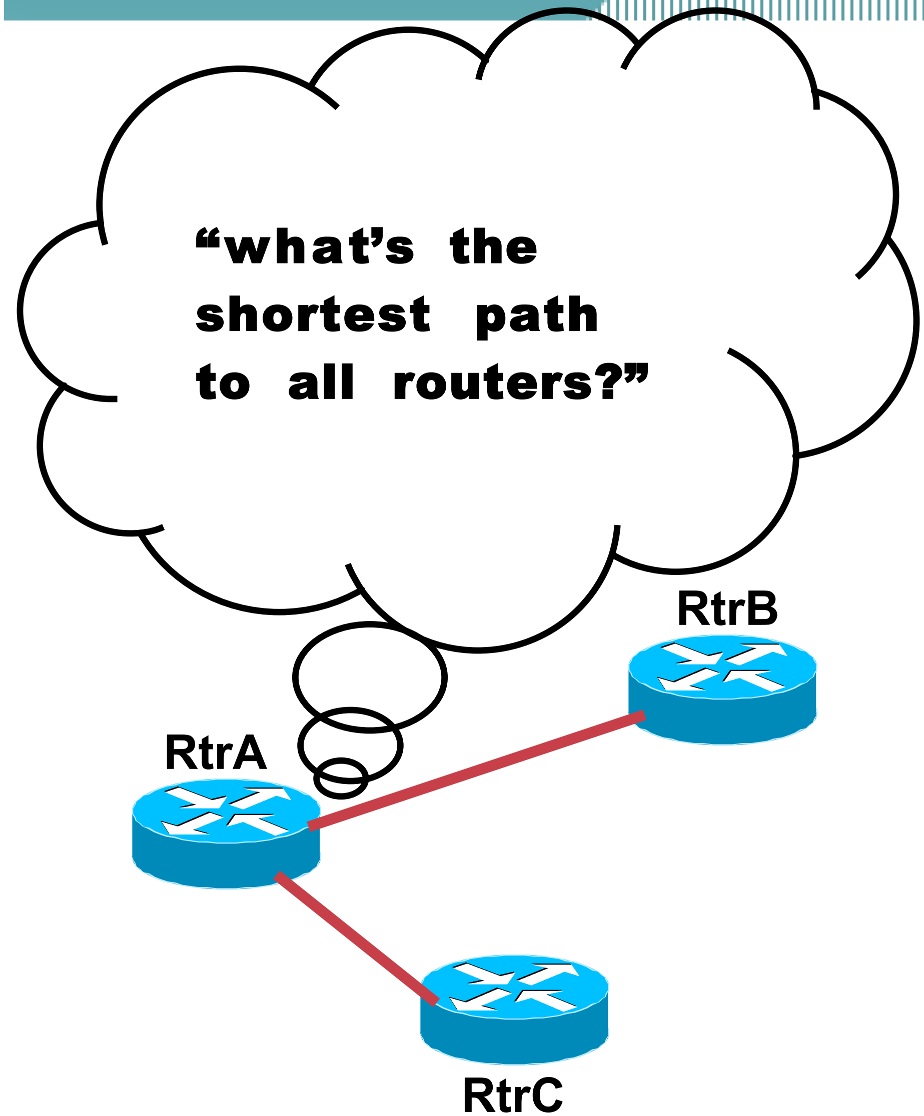
- **Normal SPF – find shortest path across all links**
- **See Perlman (2nd ed), Moy, etc. for explanation of SPF**



Path Calculation

- **Normal SPF – find shortest path across all links**
- **See Perlman (2nd ed), Moy, etc. for explanation of SPF**

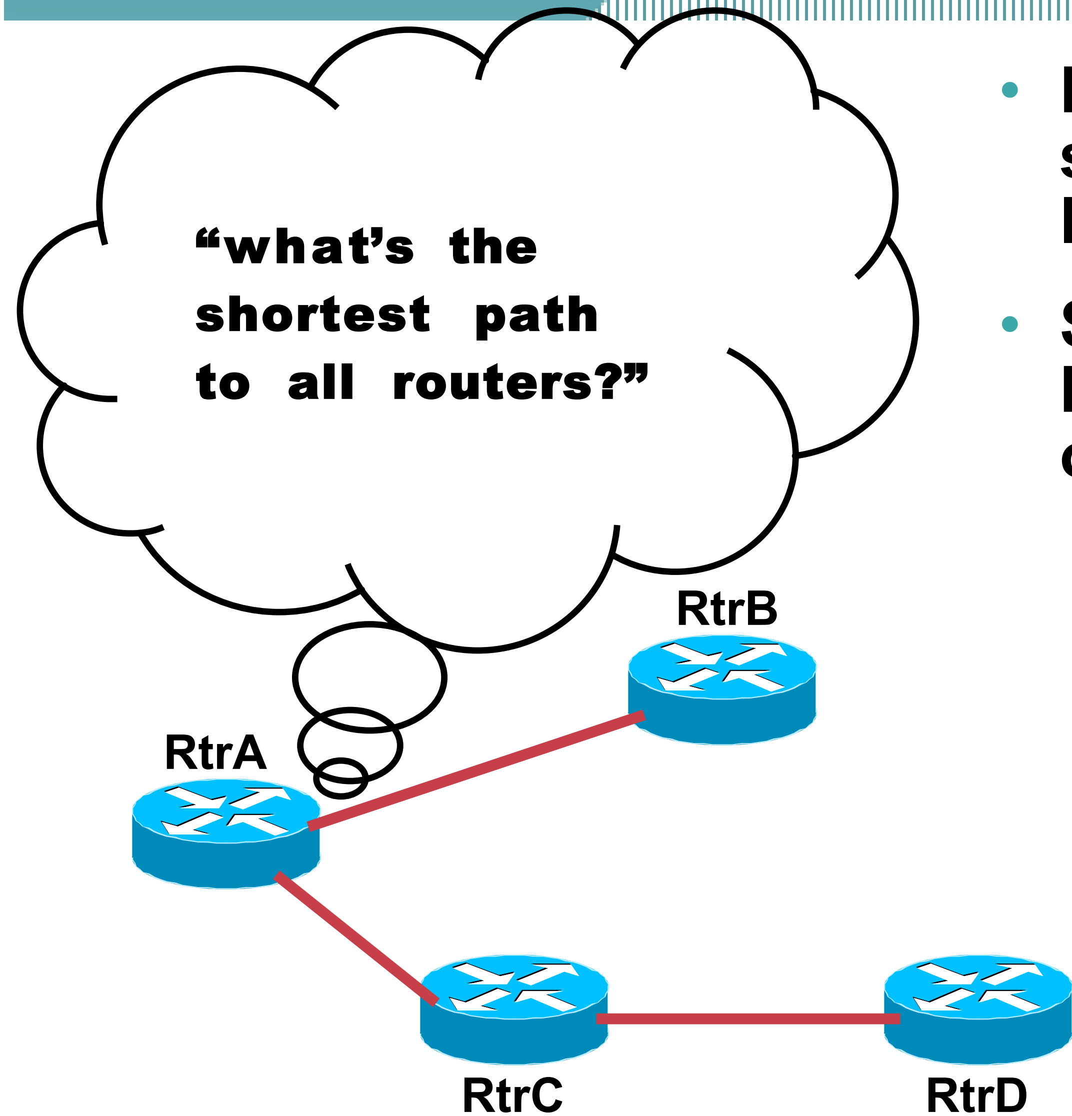
Path Calculation



- Normal SPF – find shortest path across all links
- See Perlman (2nd ed), Moy, etc. for explanation of SPF

Path Calculation

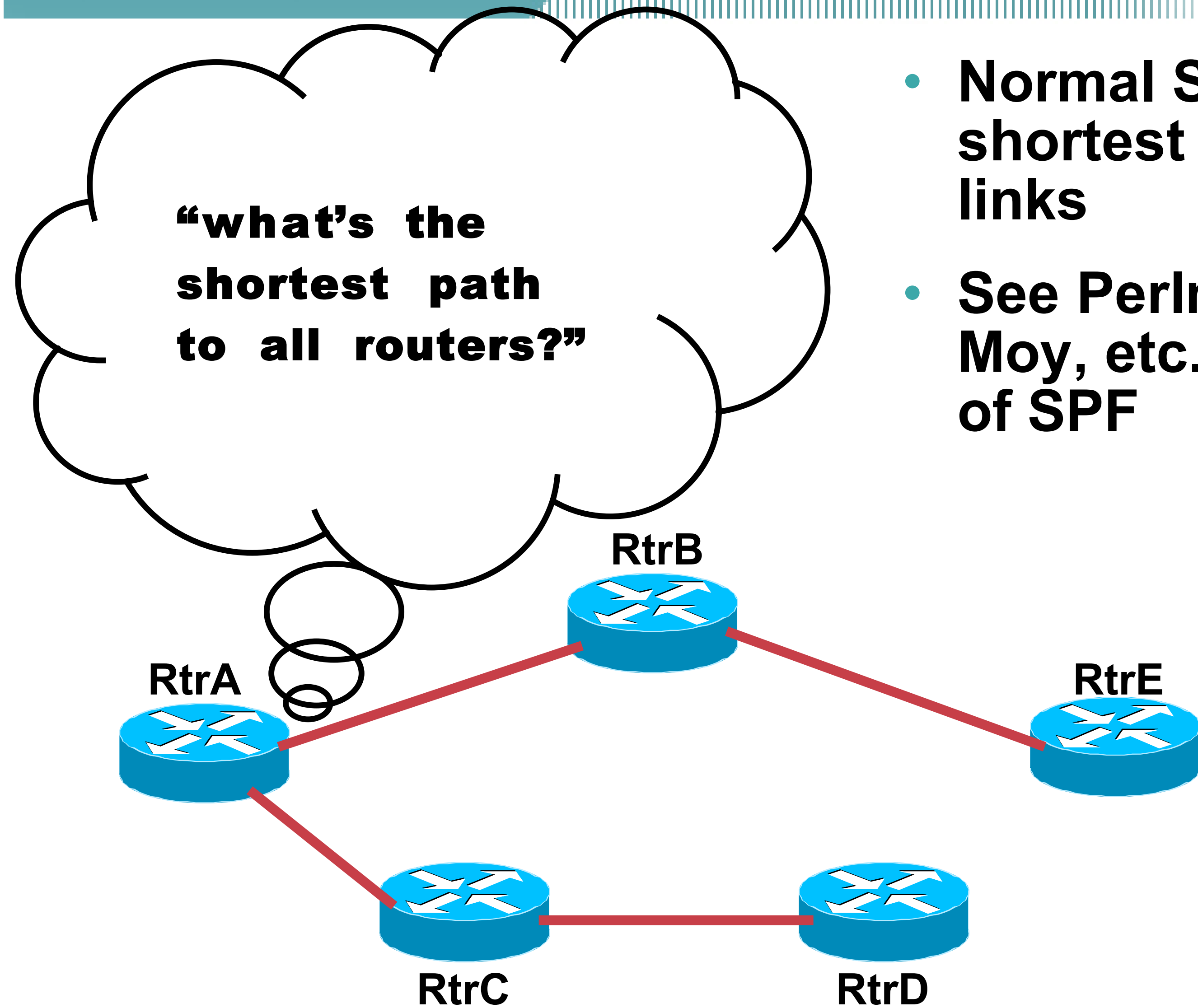
Cisco.com



- Normal SPF – find shortest path across all links
- See Perlman (2nd ed), Moy, etc. for explanation of SPF

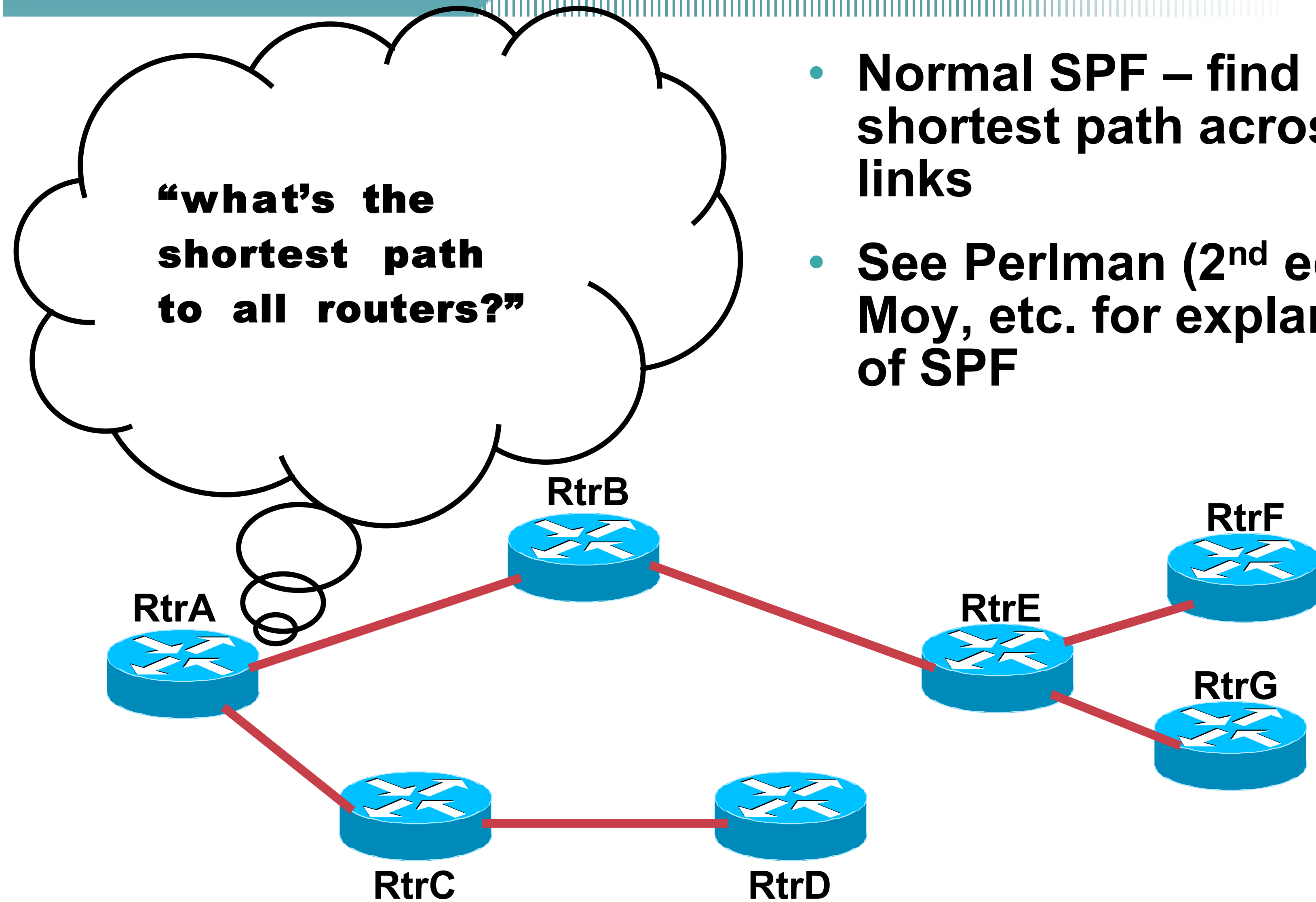
Path Calculation

- Normal SPF – find shortest path across all links
- See Perlman (2nd ed), Moy, etc. for explanation of SPF



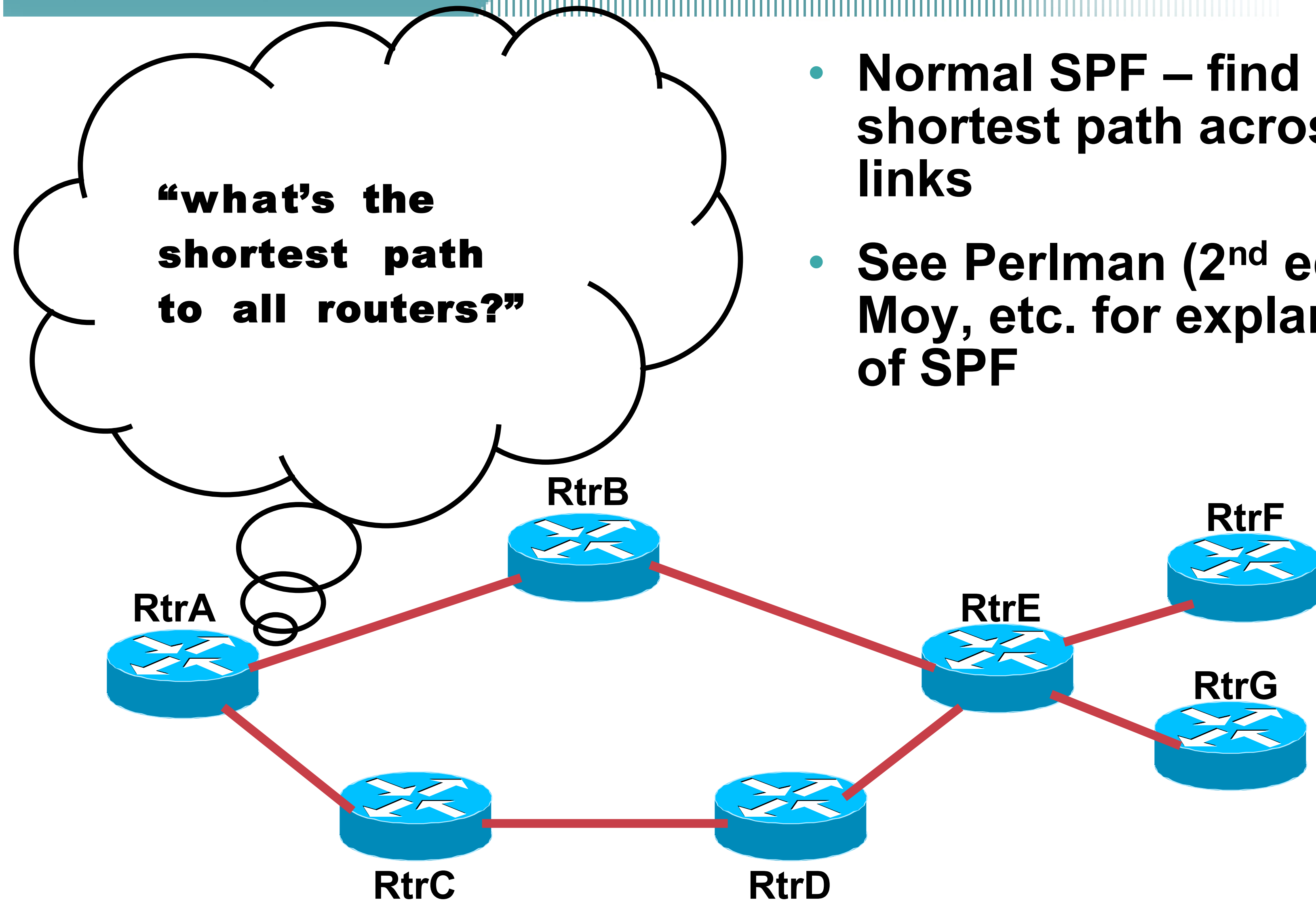
Path Calculation

- Normal SPF – find shortest path across all links
- See Perlman (2nd ed), Moy, etc. for explanation of SPF

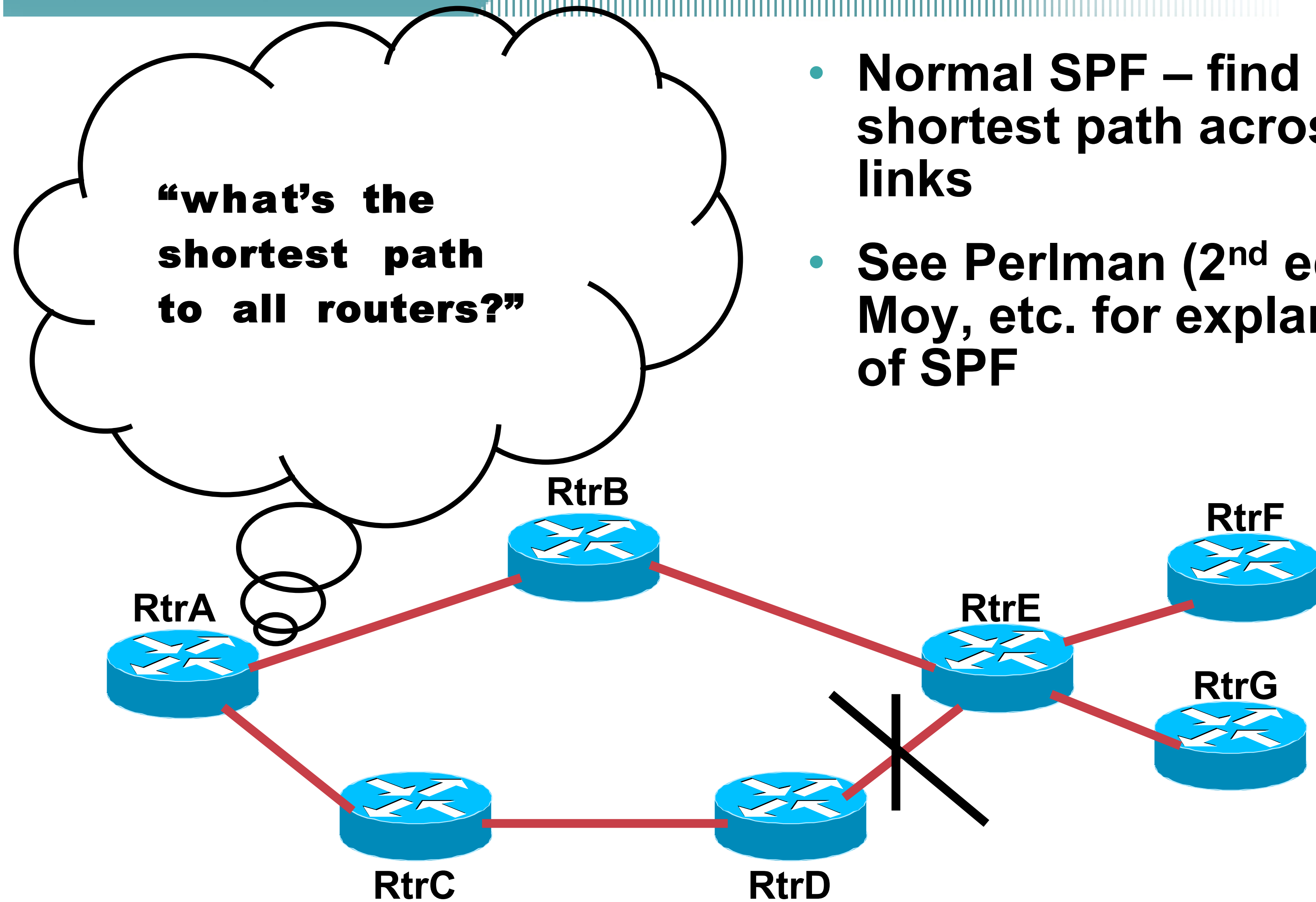


Path Calculation

- Normal SPF – find shortest path across all links
- See Perlman (2nd ed), Moy, etc. for explanation of SPF



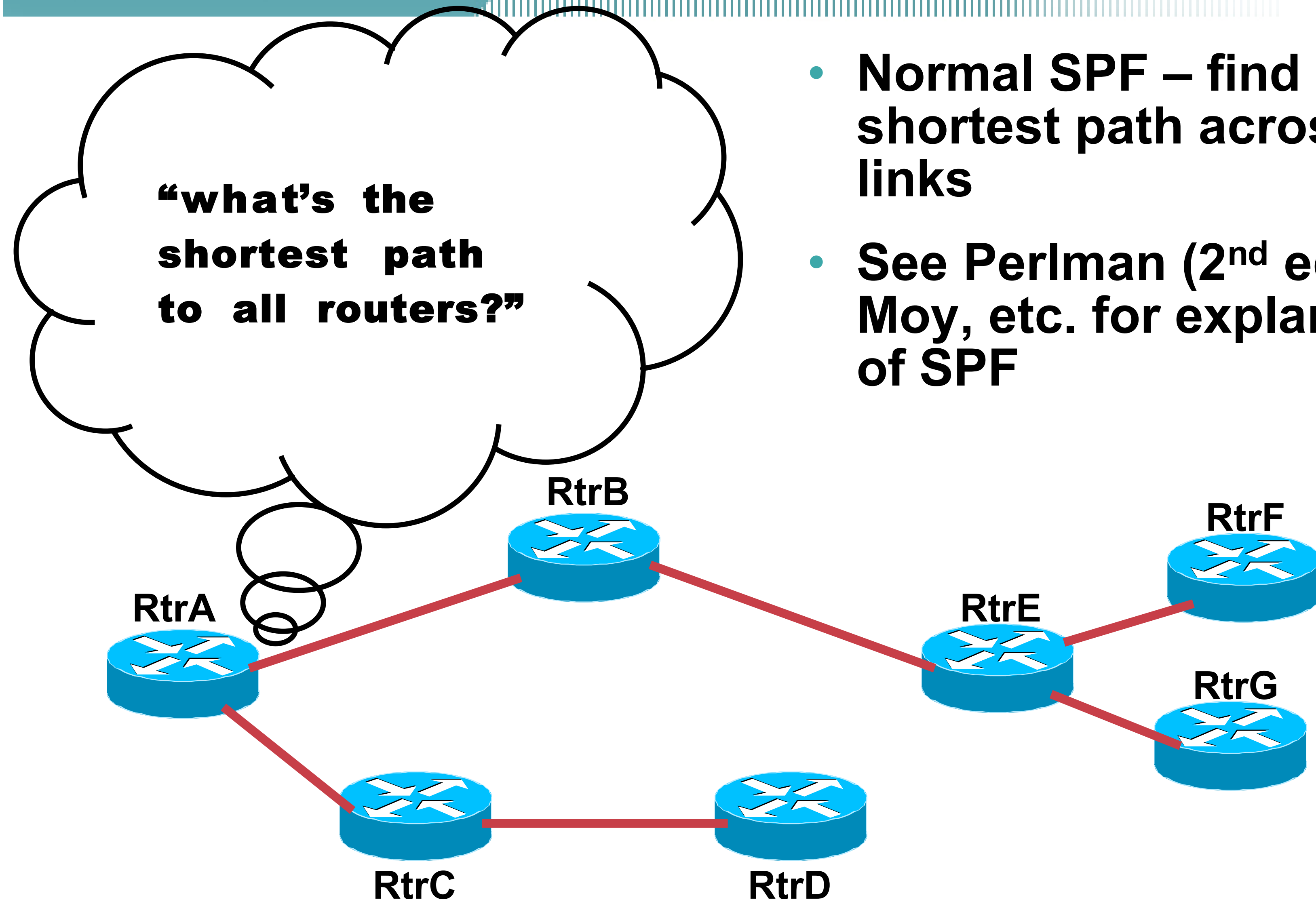
Path Calculation



- Normal SPF – find shortest path across all links
- See Perlman (2nd ed), Moy, etc. for explanation of SPF

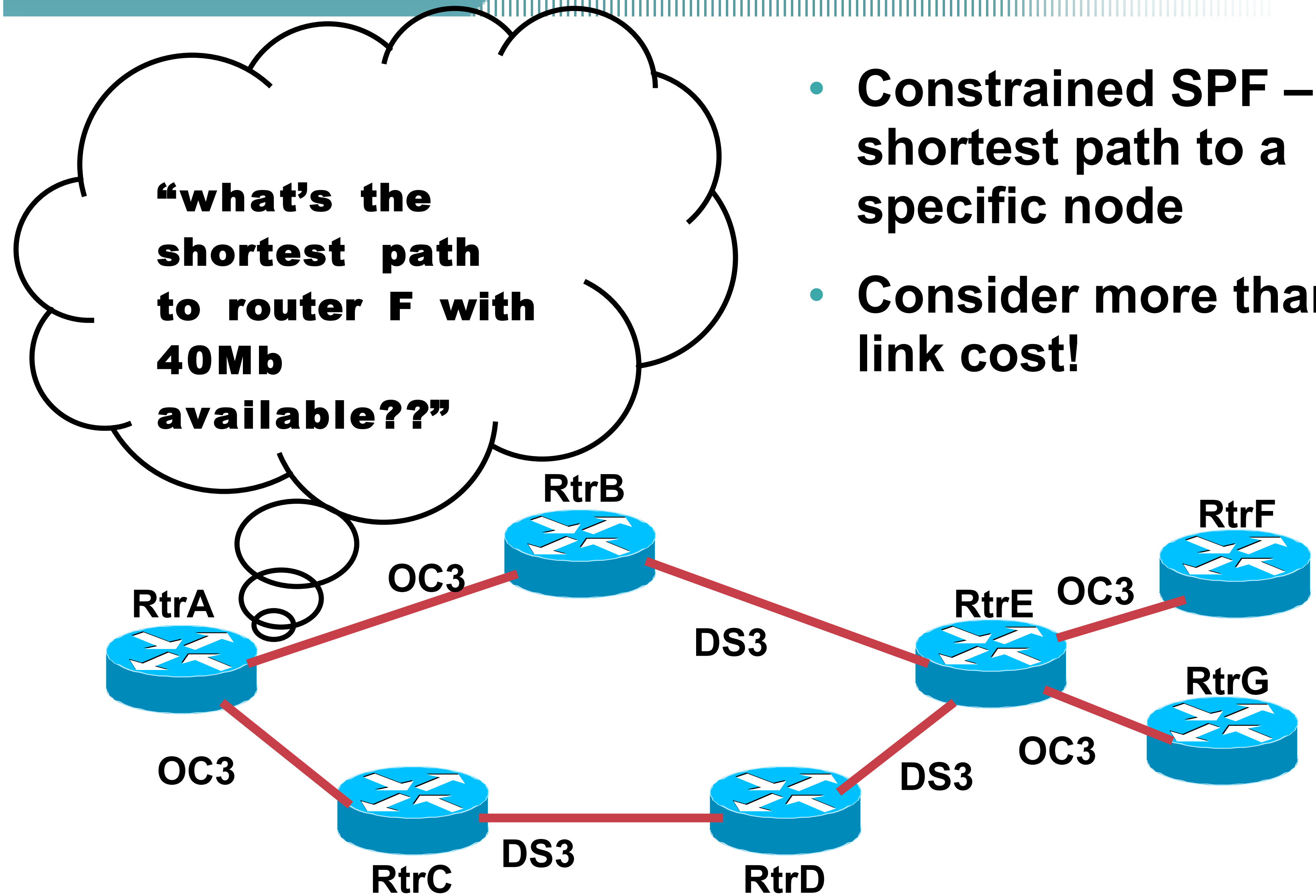
Path Calculation

- Normal SPF – find shortest path across all links
- See Perlman (2nd ed), Moy, etc. for explanation of SPF



Path Calculation

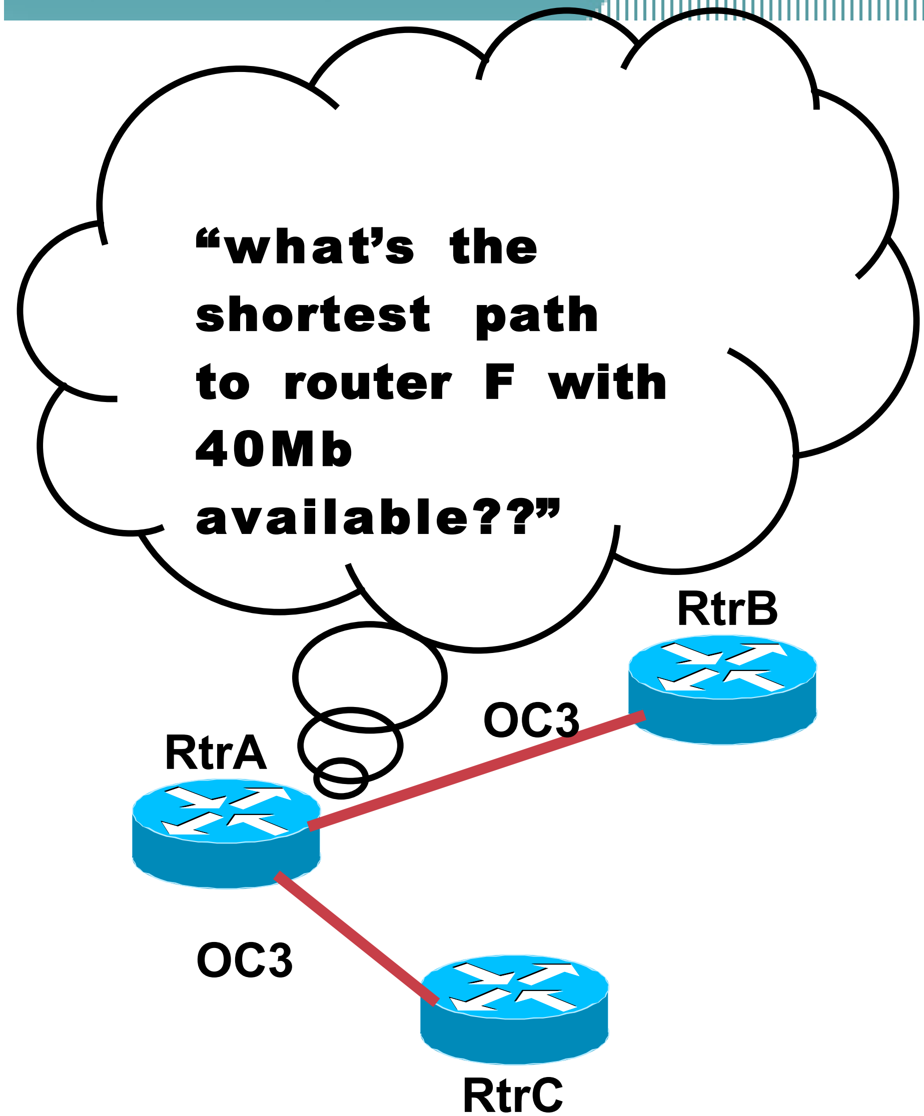
- Constrained SPF – find shortest path to a specific node
- Consider more than just link cost!



Path Calculation

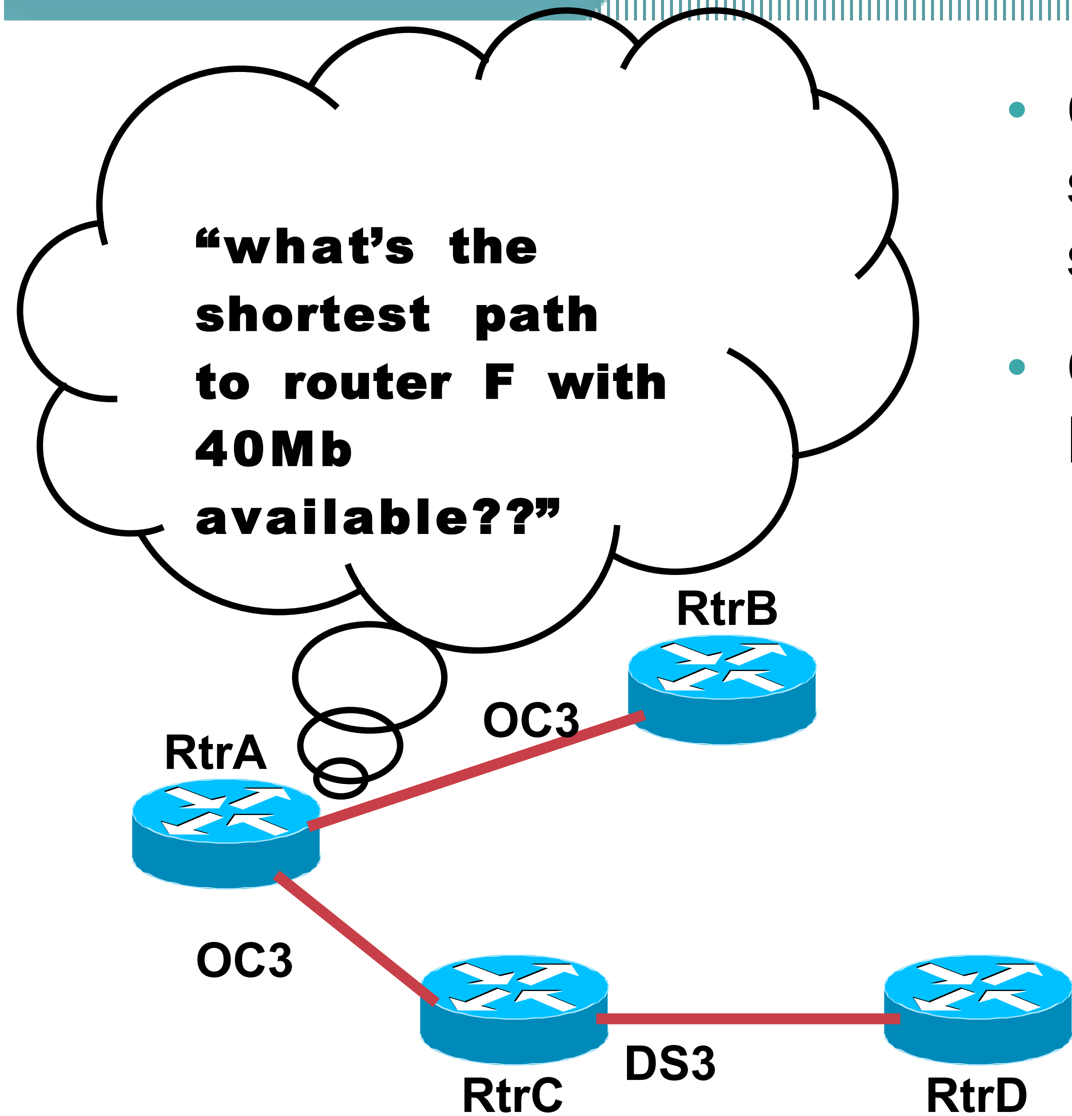
- Constrained SPF – find shortest path to a specific node
- Consider more than just link cost!

Path Calculation



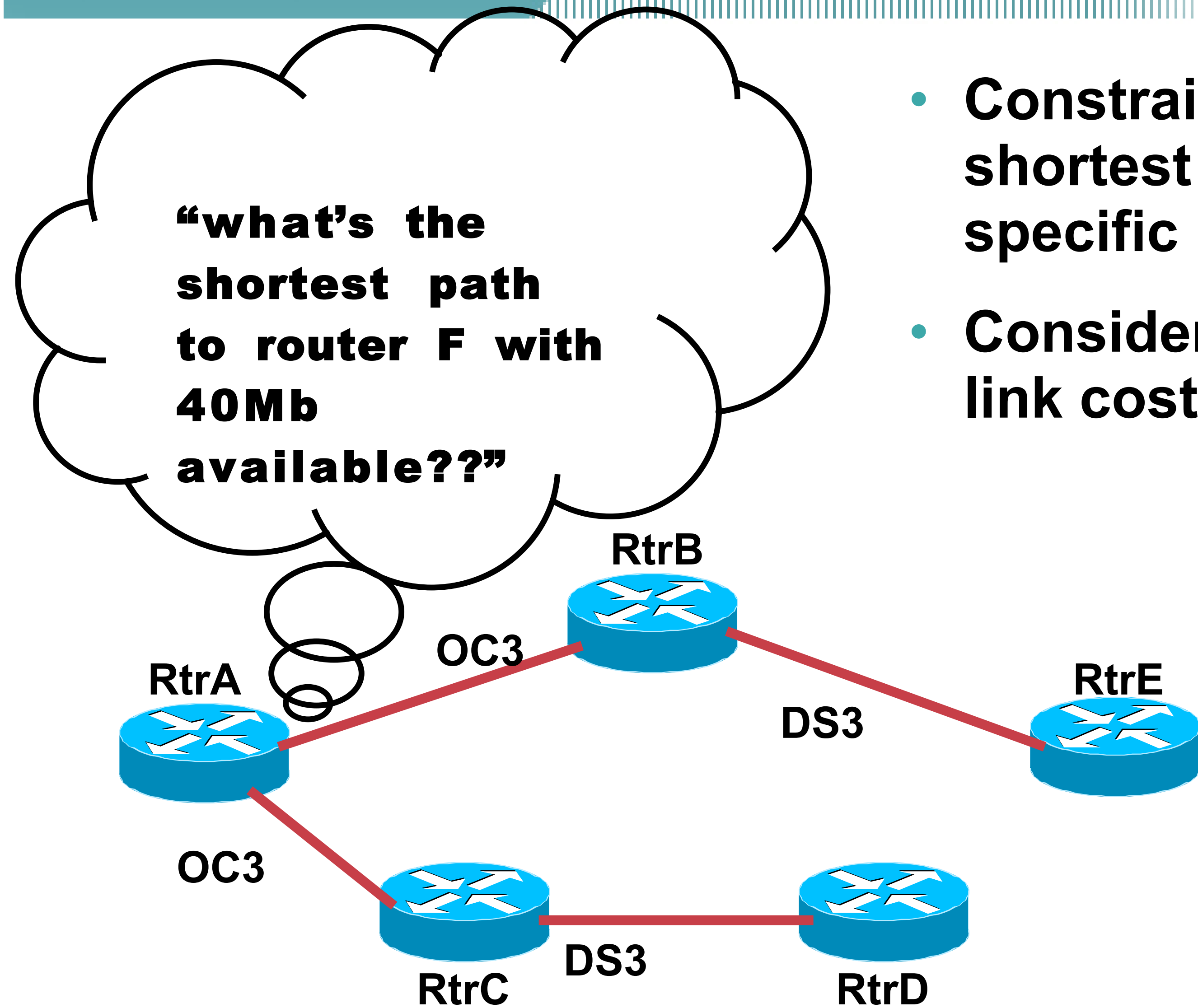
- **Constrained SPF – find shortest path to a specific node**
- **Consider more than just link cost!**

Path Calculation



- Constrained SPF – find shortest path to a specific node
- Consider more than just link cost!

Path Calculation

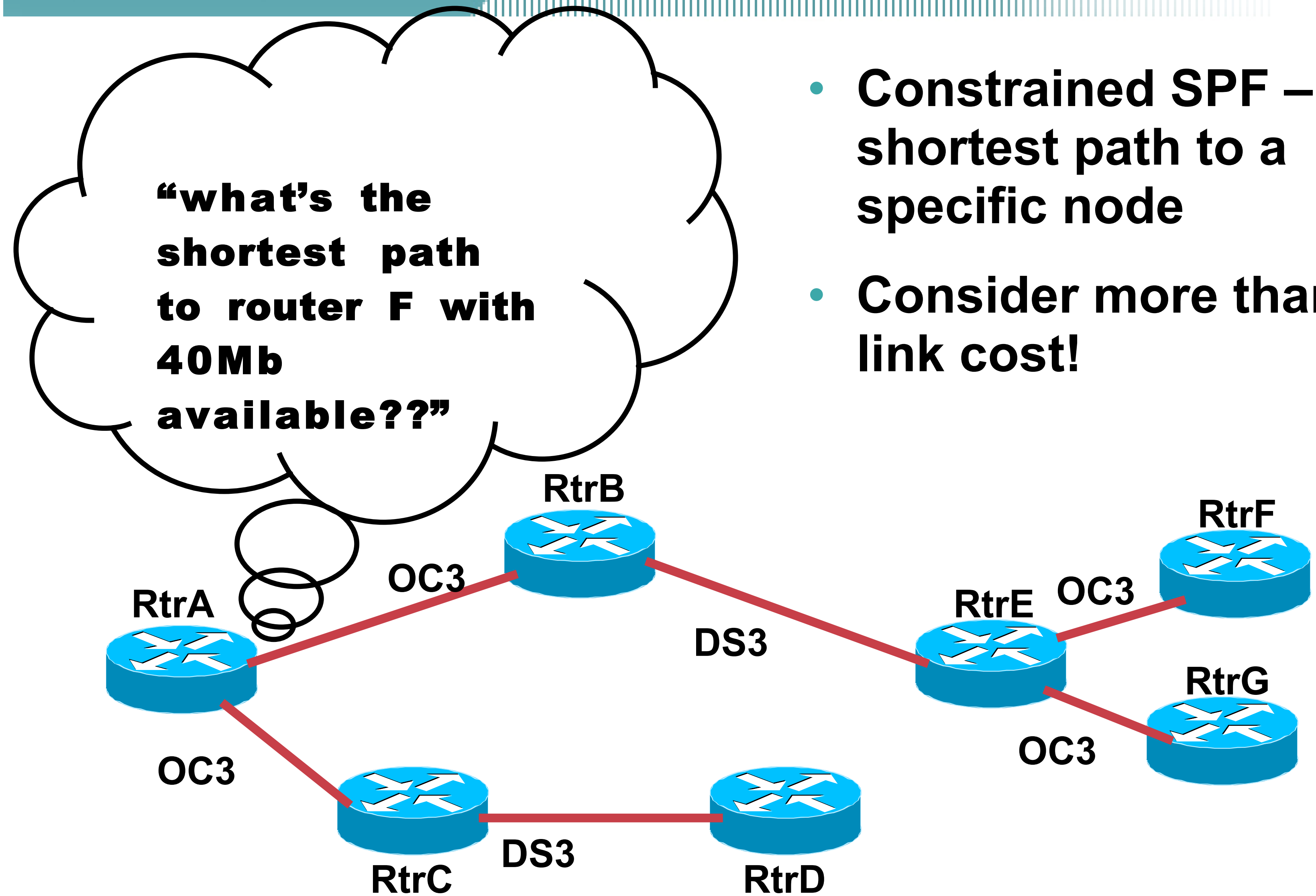


- Constrained SPF – find shortest path to a specific node
- Consider more than just link cost!

Path Calculation

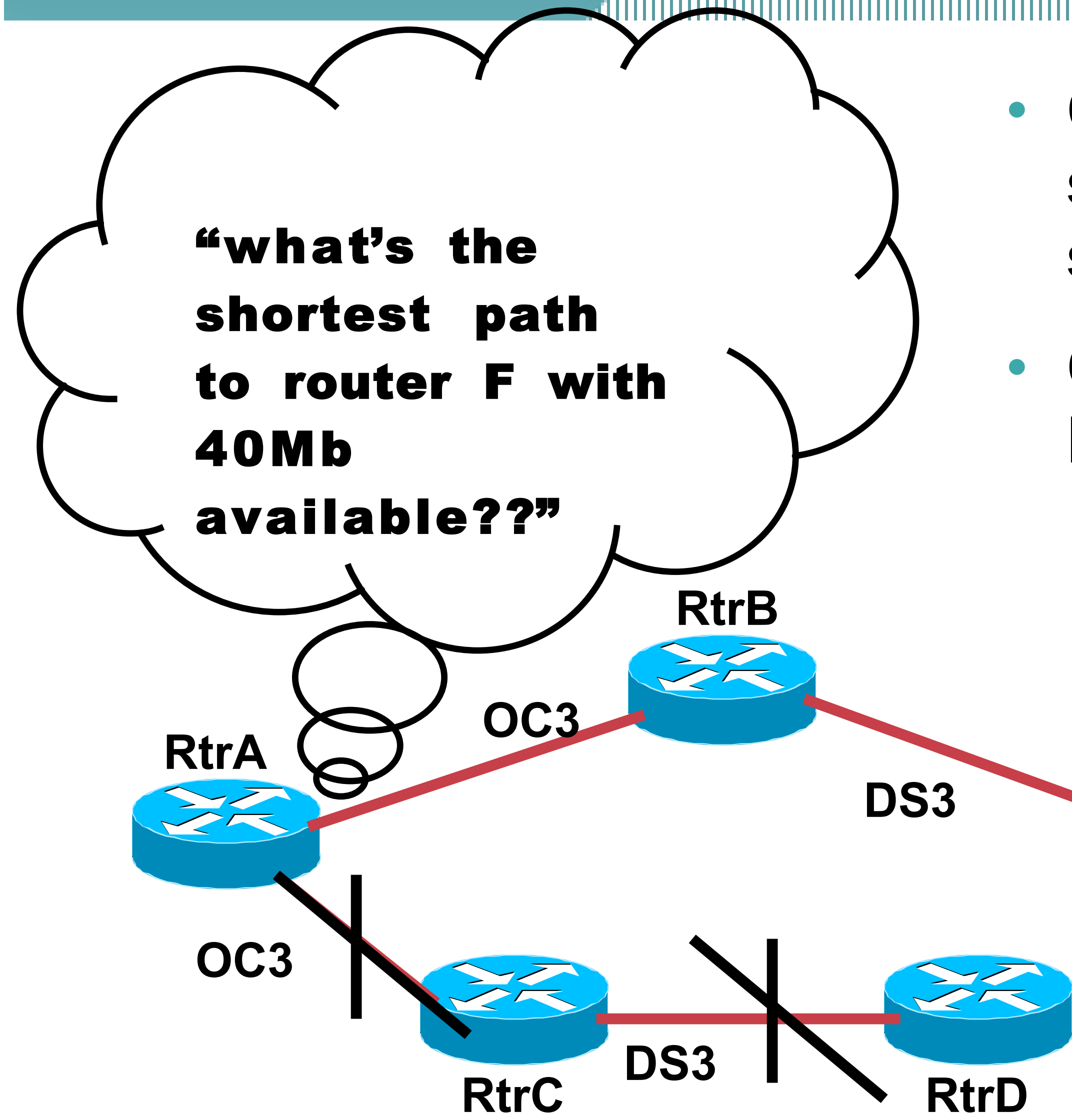
Cisco.com

- Constrained SPF – find shortest path to a specific node
- Consider more than just link cost!



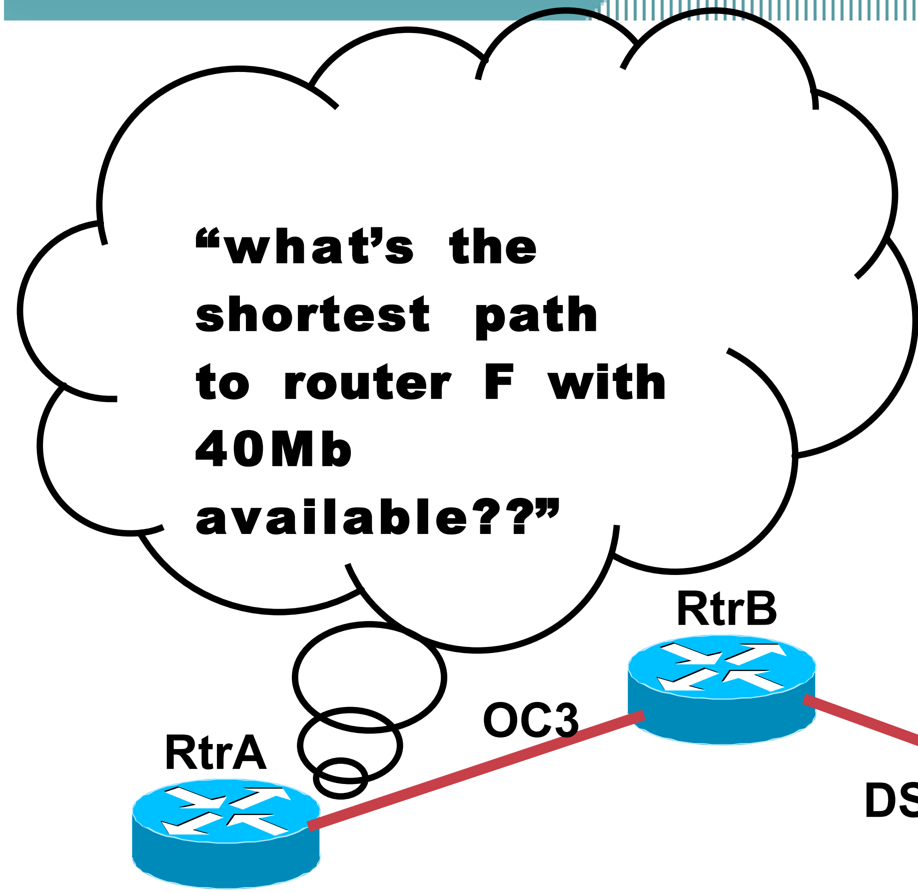
Path Calculation

Cisco.com



- Constrained SPF – find shortest path to a specific node
- Consider more than just link cost!

Path Calculation



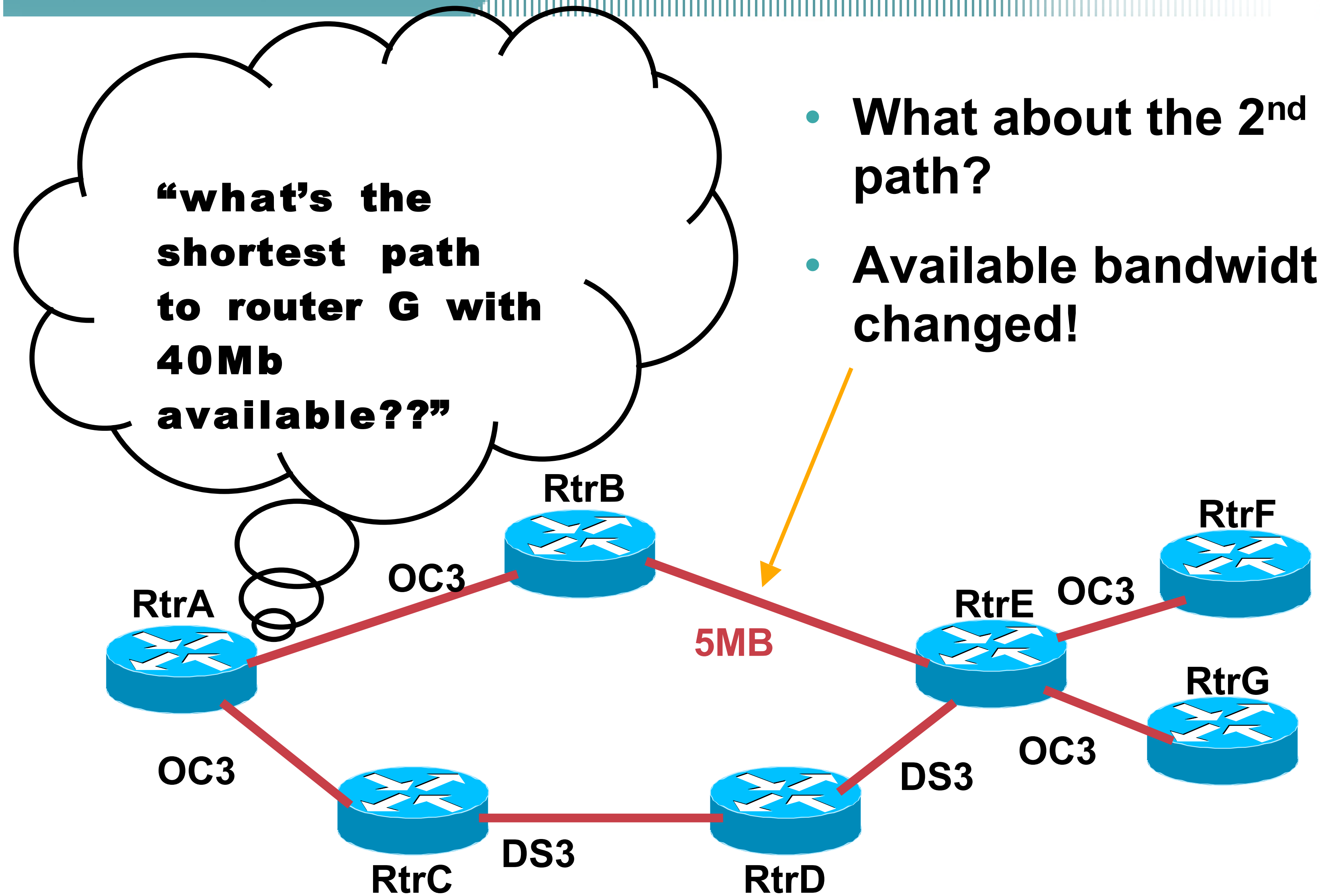
- Constrained SPF – find shortest path to a specific node
- Consider more than just link cost!

Path Calculation

- **“But Wait! There’s nothing different between the two SPF results!”**
- **….but….**

Path Calculation

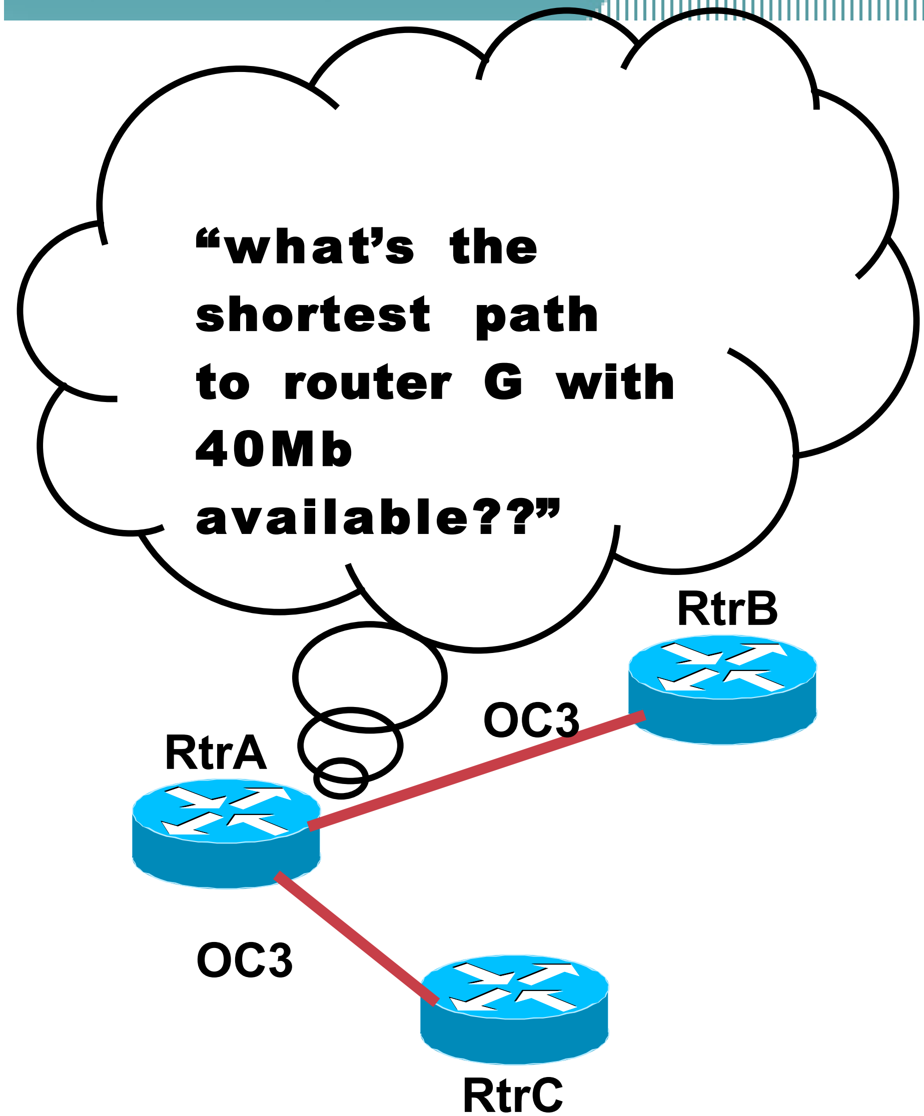
- What about the 2nd path?
- Available bandwidth has changed!



Path Calculation

- What about the 2nd path?
- Available bandwidth has changed!

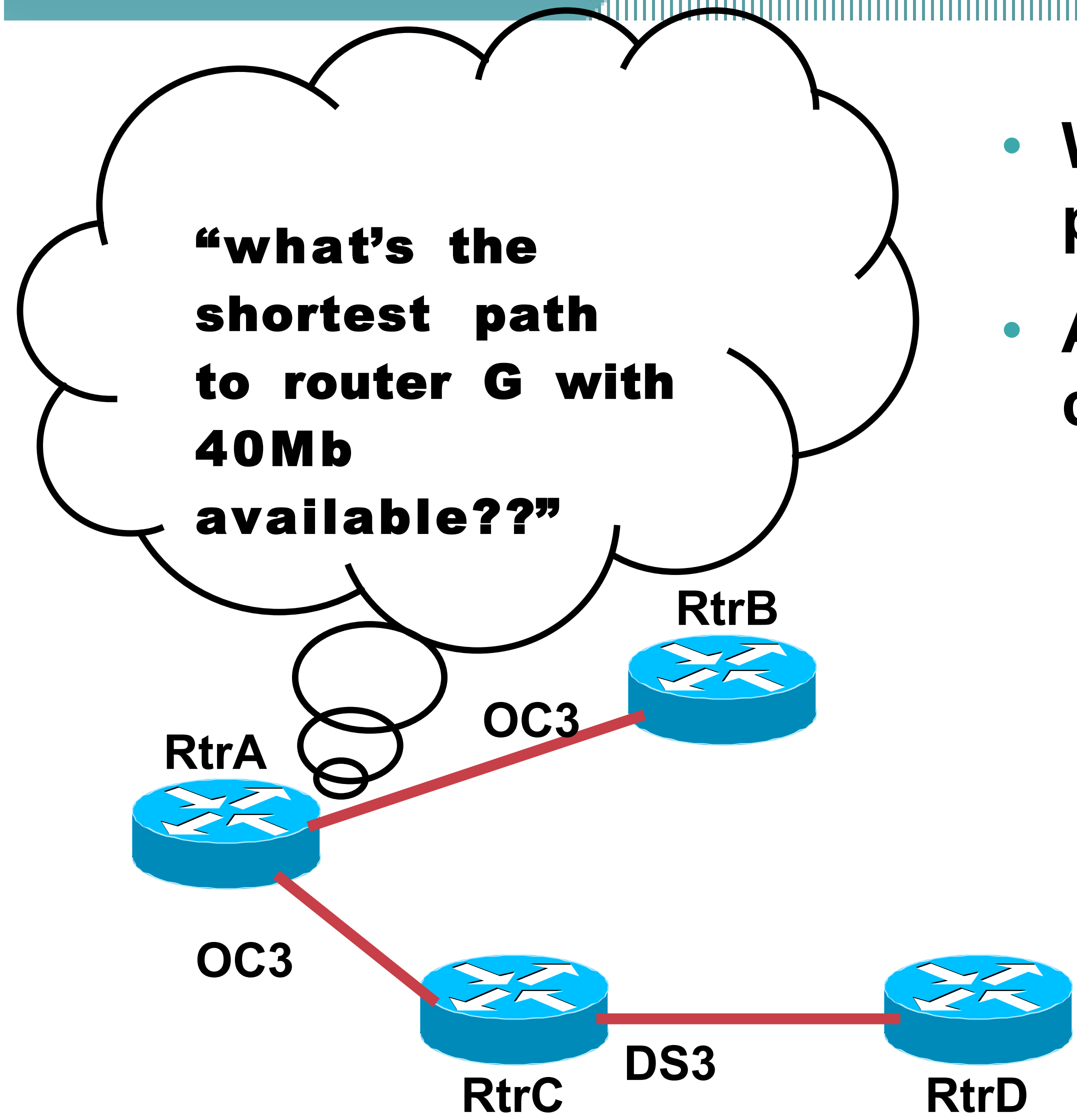
Path Calculation



- What about the 2nd path?
- Available bandwidth has changed!

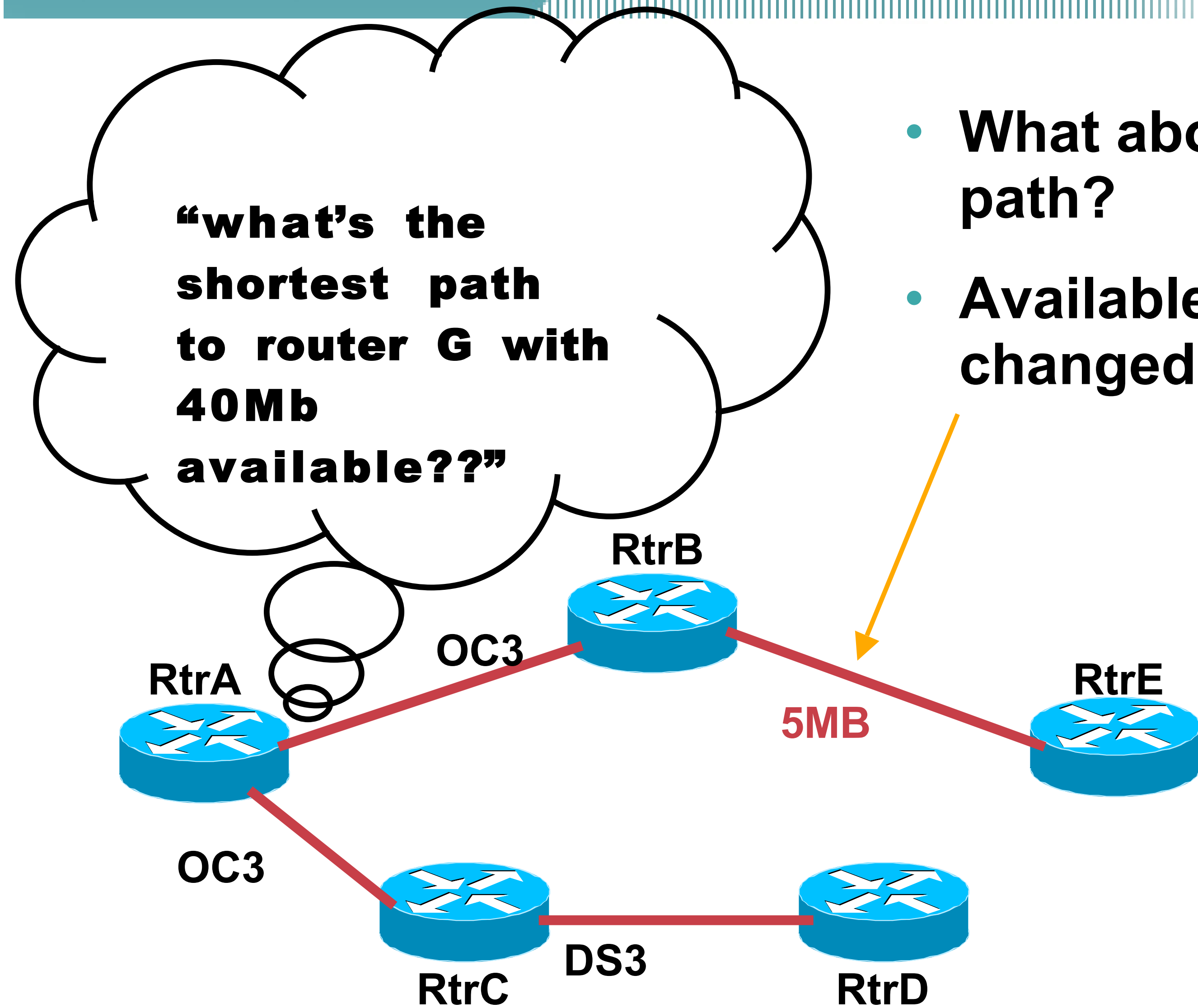
Path Calculation

- What about the 2nd path?
- Available bandwidth has changed!



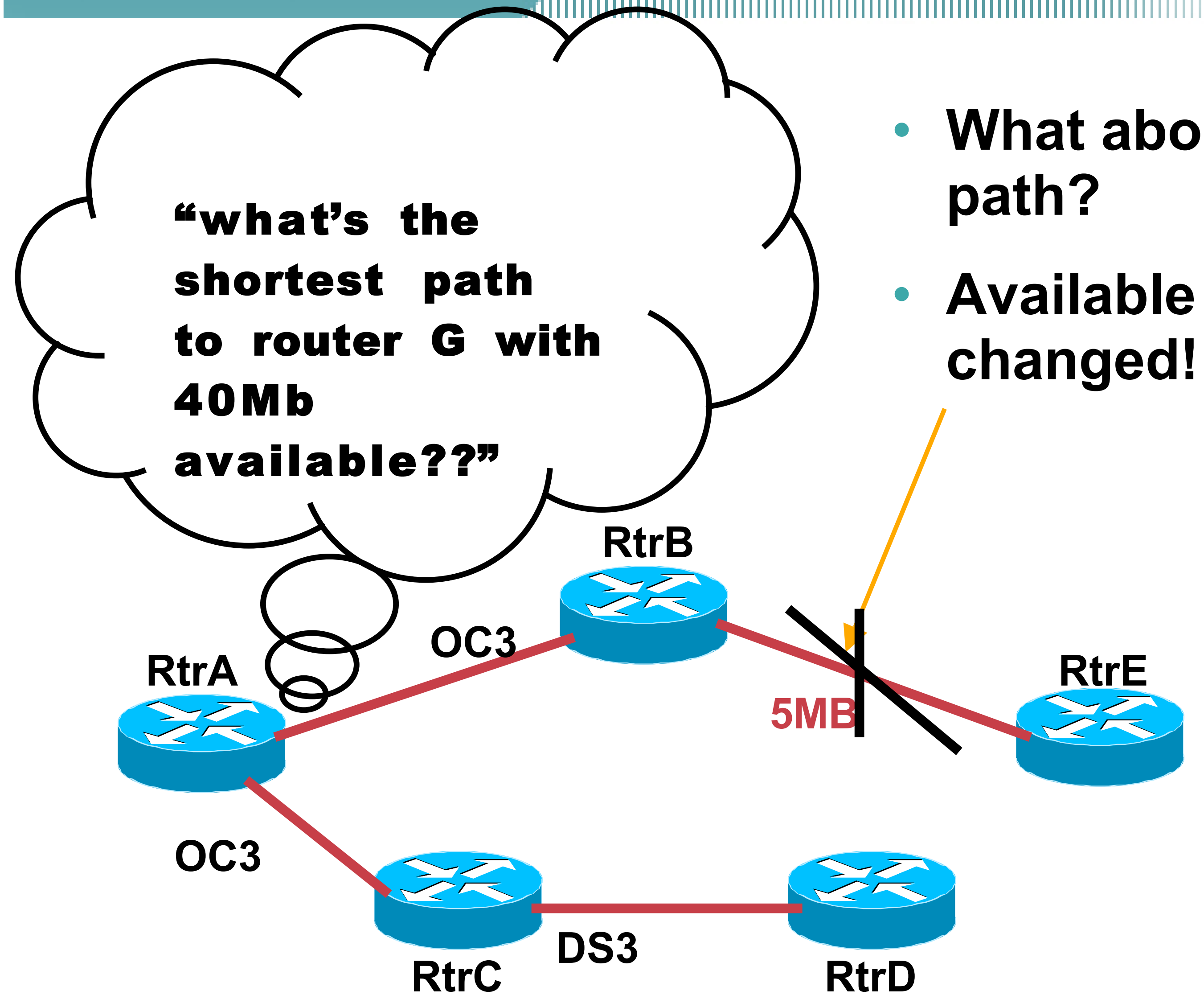
Path Calculation

- What about the 2nd path?
- Available bandwidth has changed!

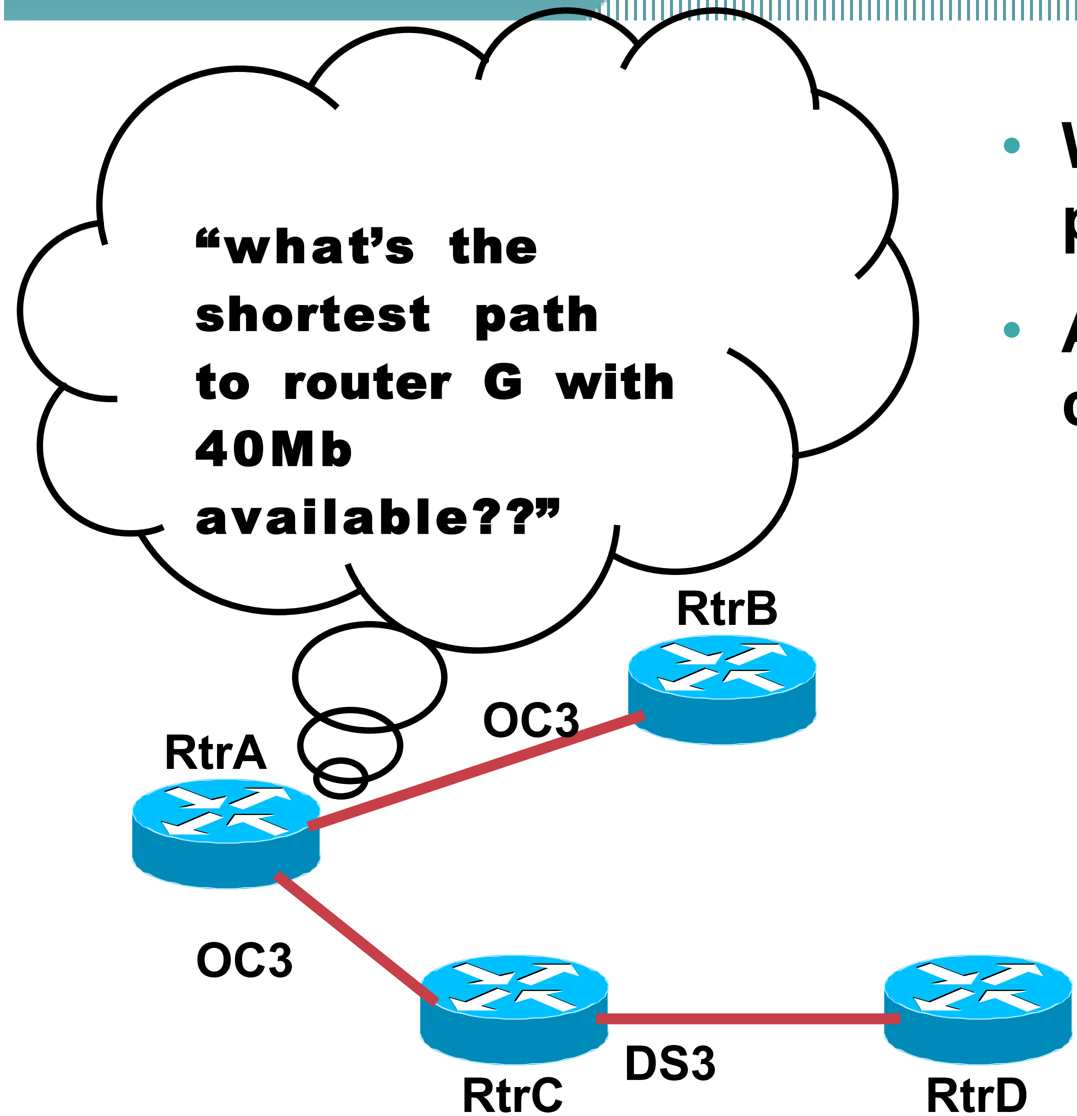


Path Calculation

- What about the 2nd path?
- Available bandwidth has changed!

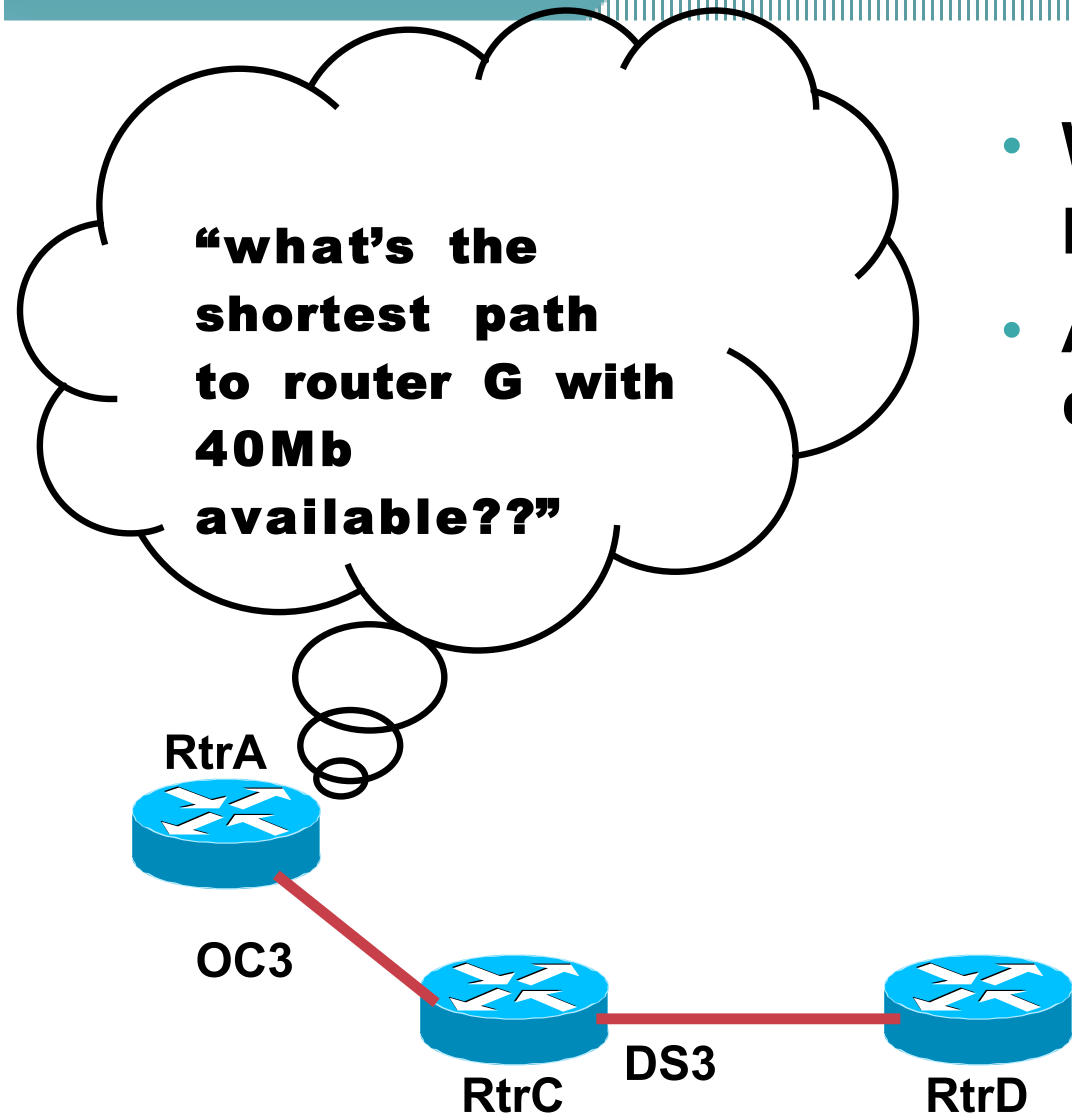


Path Calculation



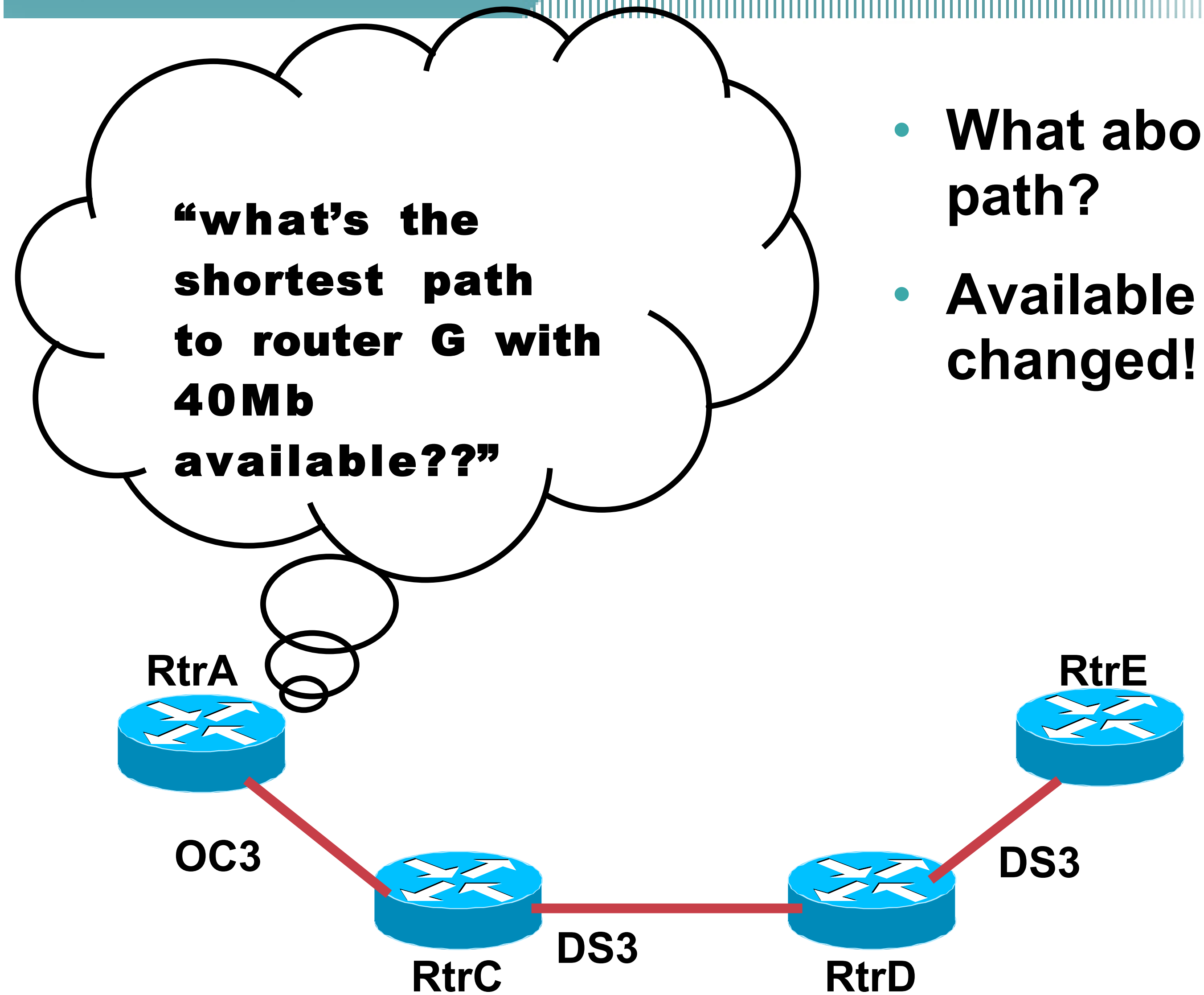
- What about the 2nd path?
- Available bandwidth has changed!

Path Calculation



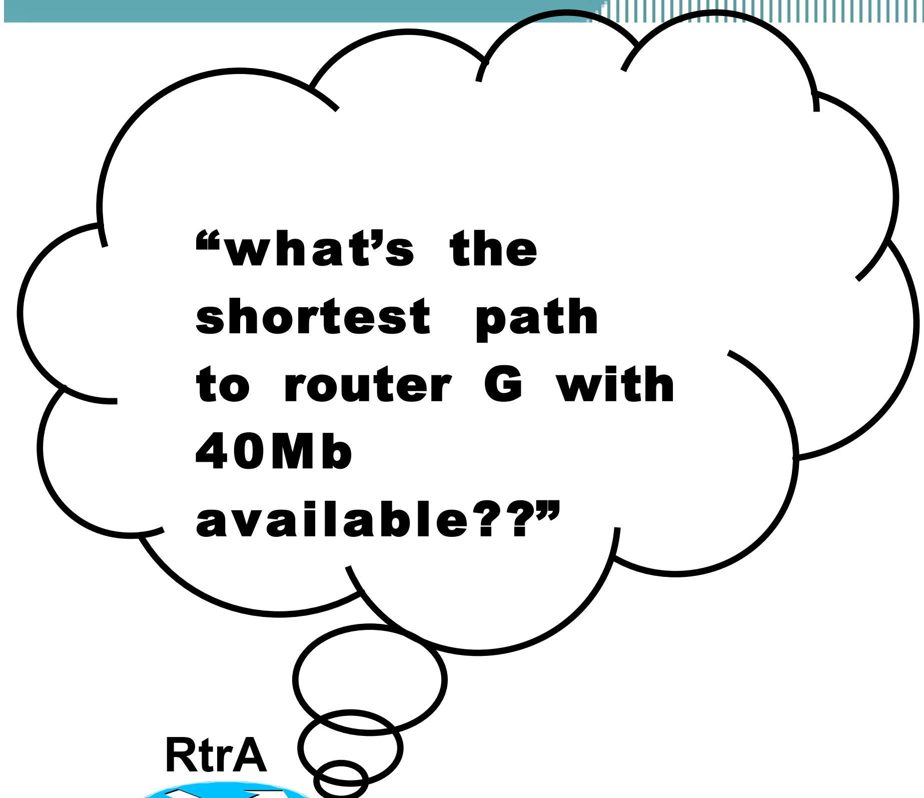
- What about the 2nd path?
- Available bandwidth has changed!

Path Calculation

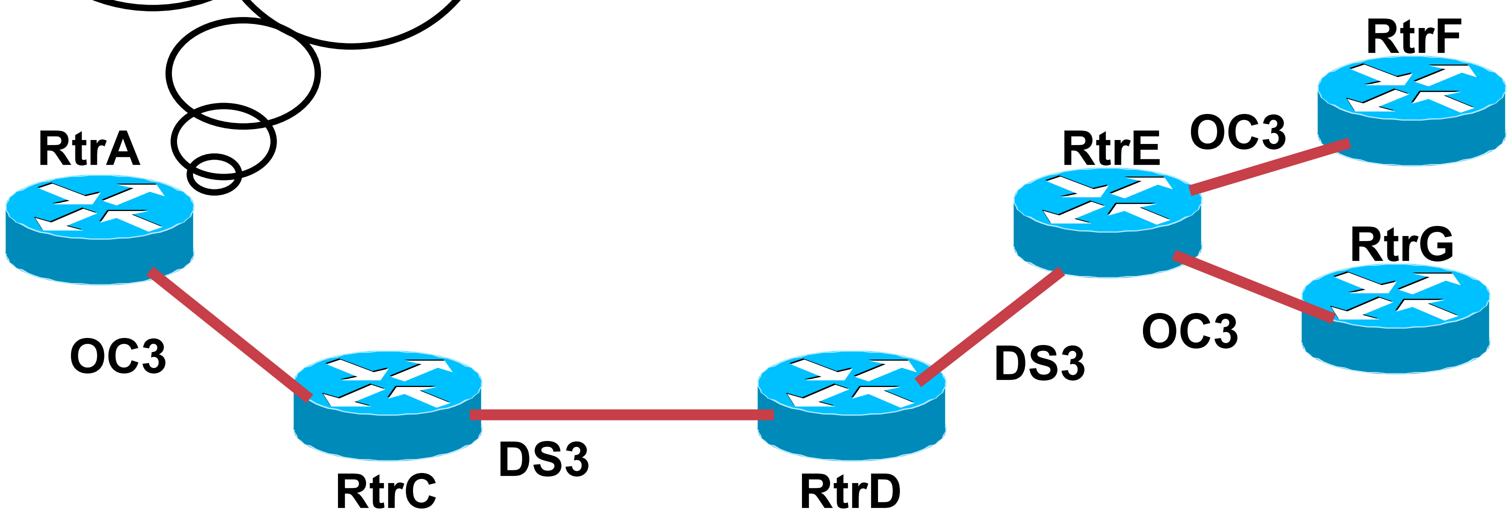


- What about the 2nd path?
- Available bandwidth has changed!

Path Calculation

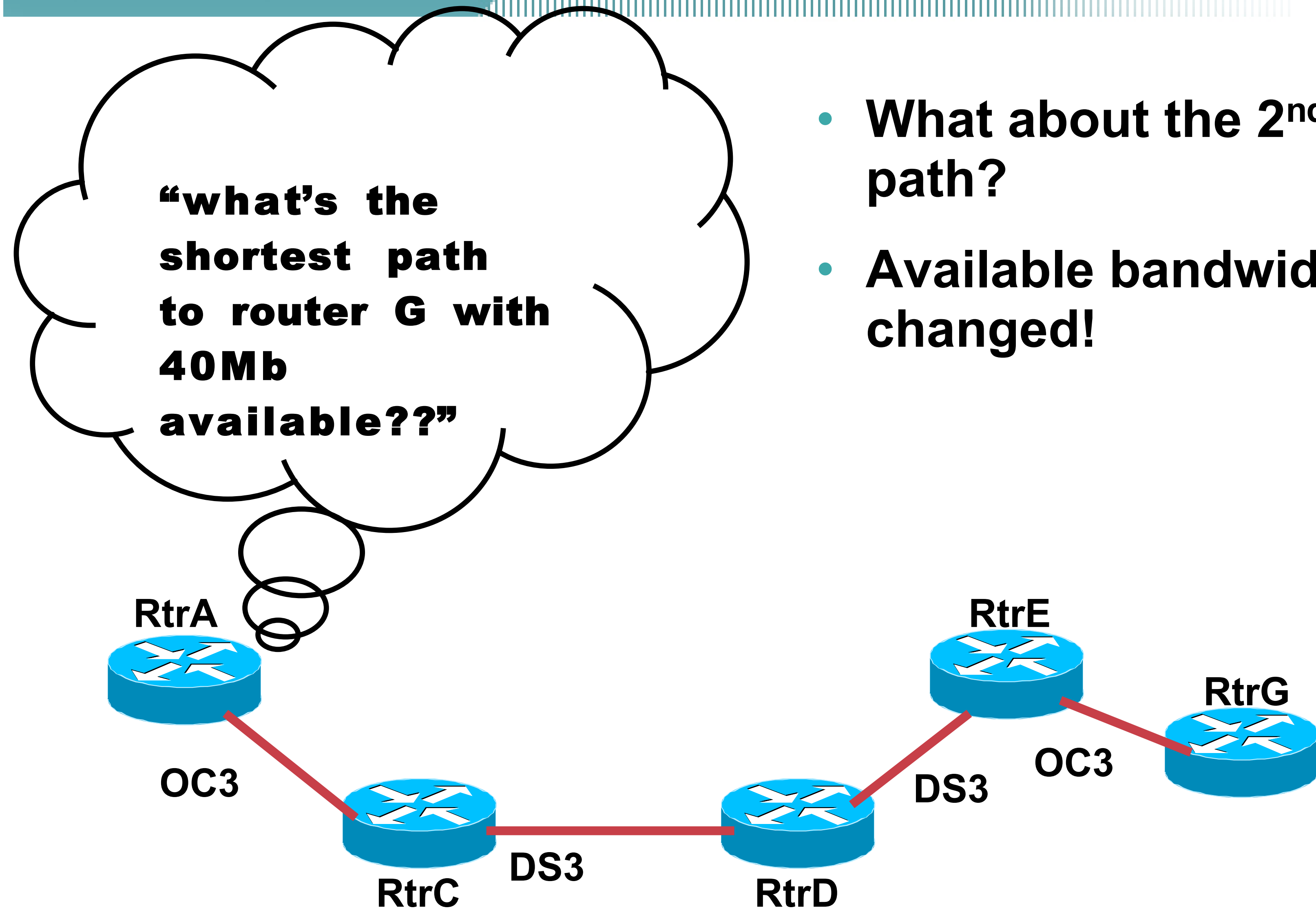


- What about the 2nd path?
- Available bandwidth has changed!



Path Calculation

- What about the 2nd path?
- Available bandwidth has changed!

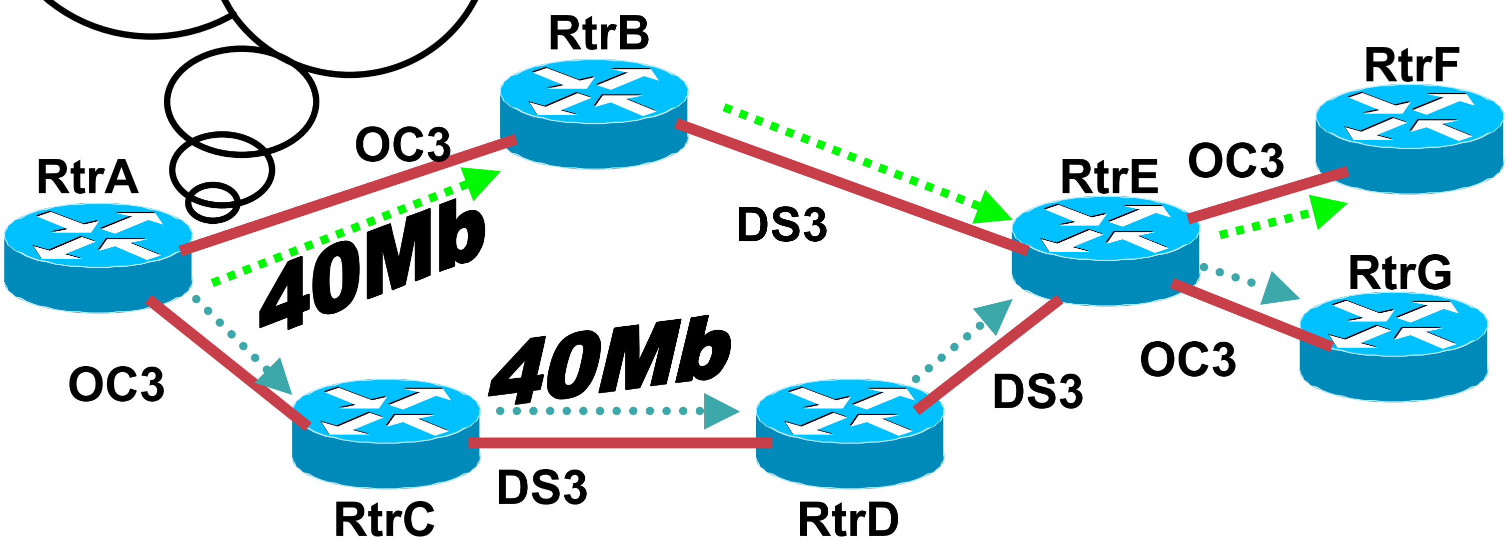


Path Calculation

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

- **End result:**

  -bandwidth used efficiently!



Path Calculation

- What if there's more than one path that meets the minimum requirements (bandwidth, etc.)?
- PCALC algorithm:

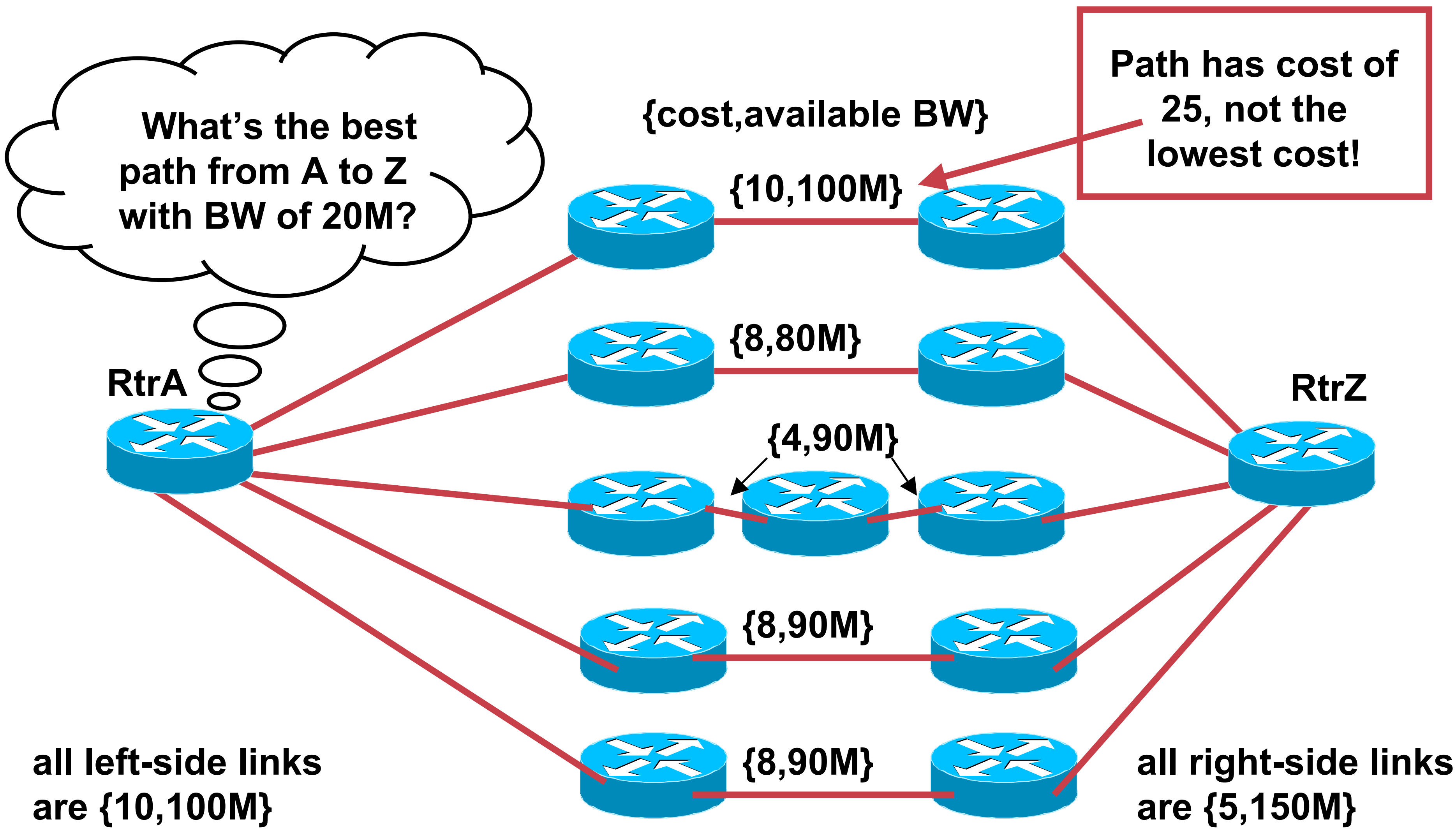
  Find all paths with the lowest IGP cost

  Then pick the path with the highest minimum bandwidth along the path

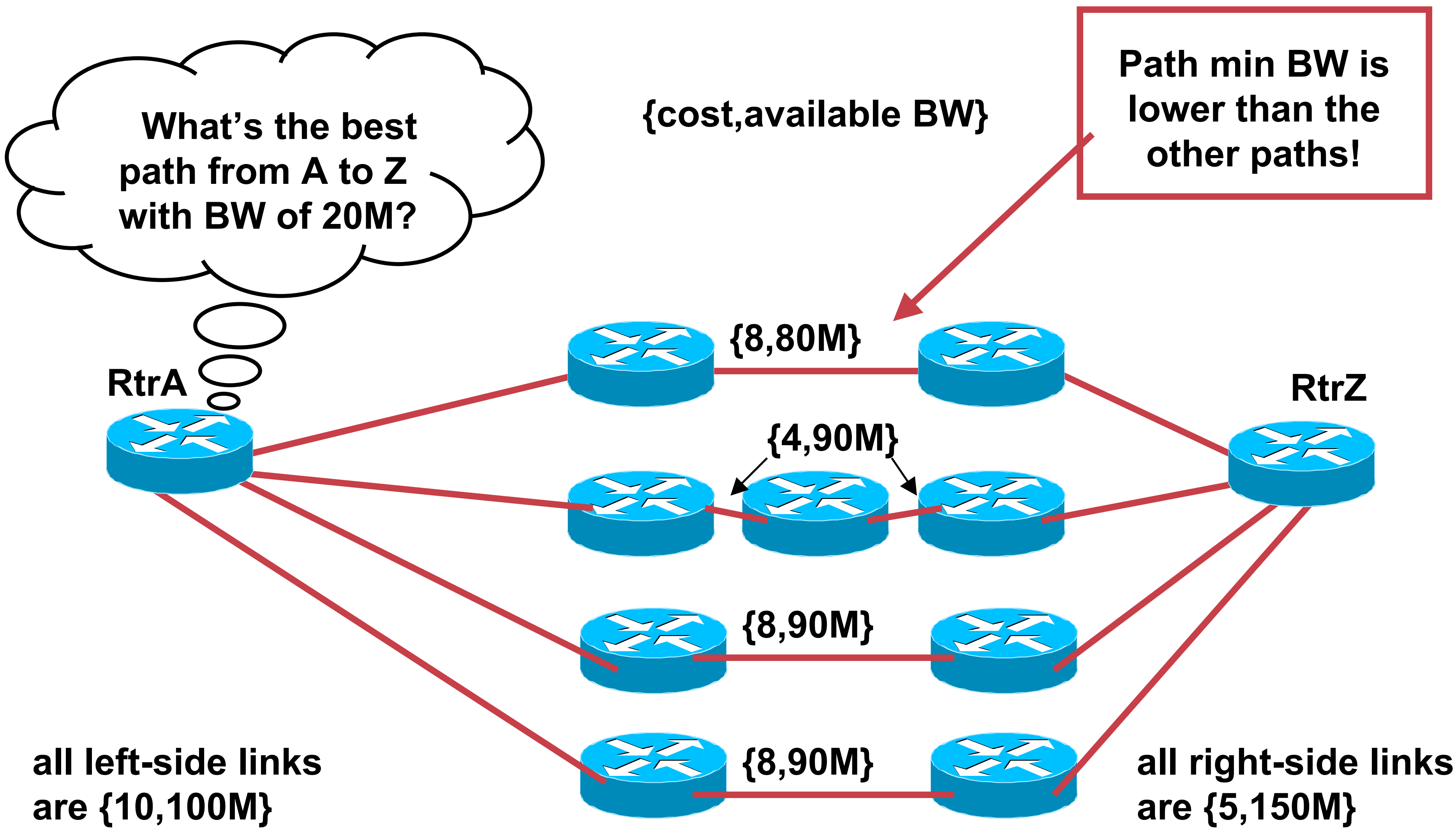
  Then pick the path with the lowest hop count (not IGP cost, but hop count)

  Then just pick one path at "random" (take the top path on the TENT list)

Path Calculation

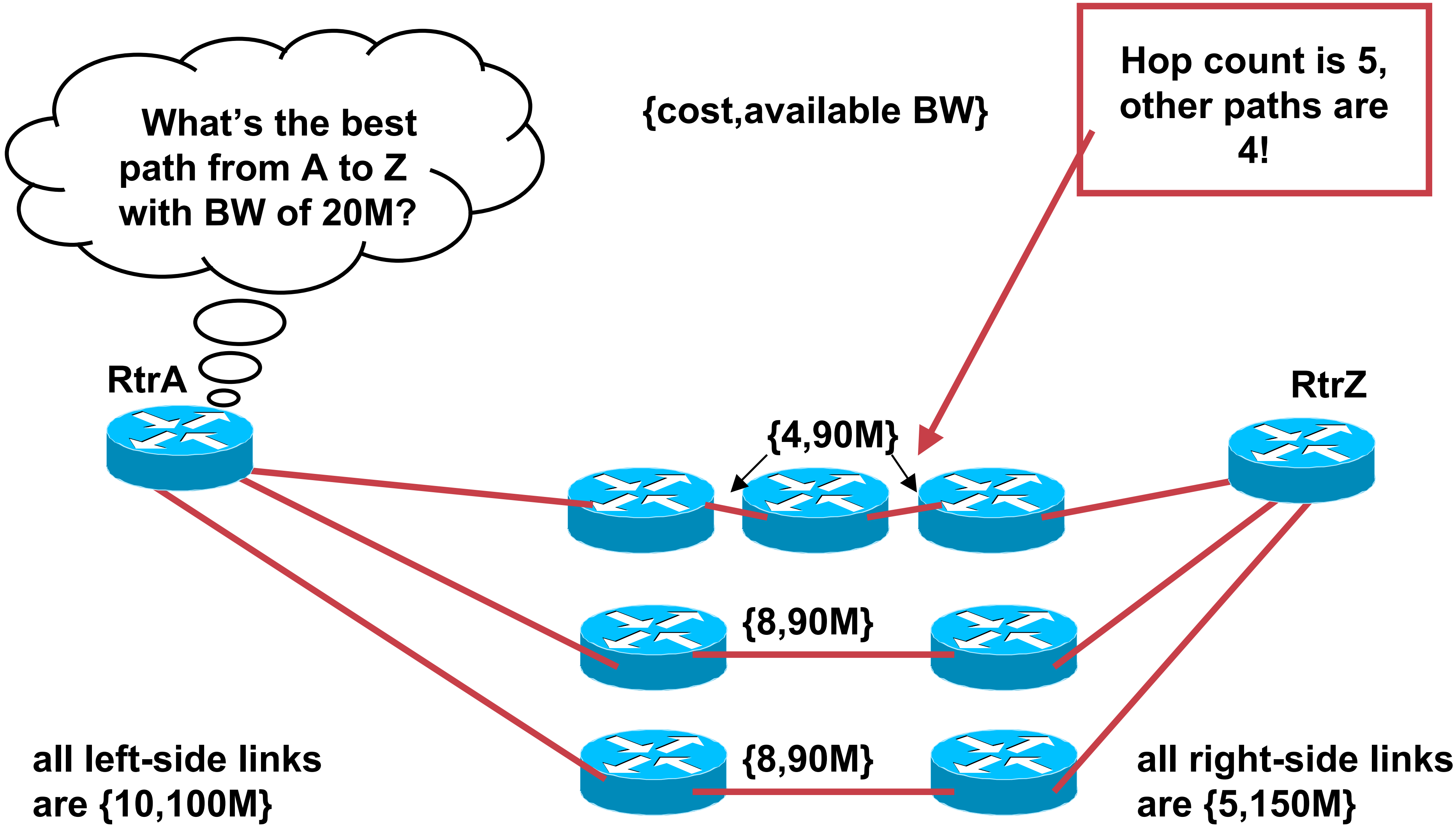


Path Calculation

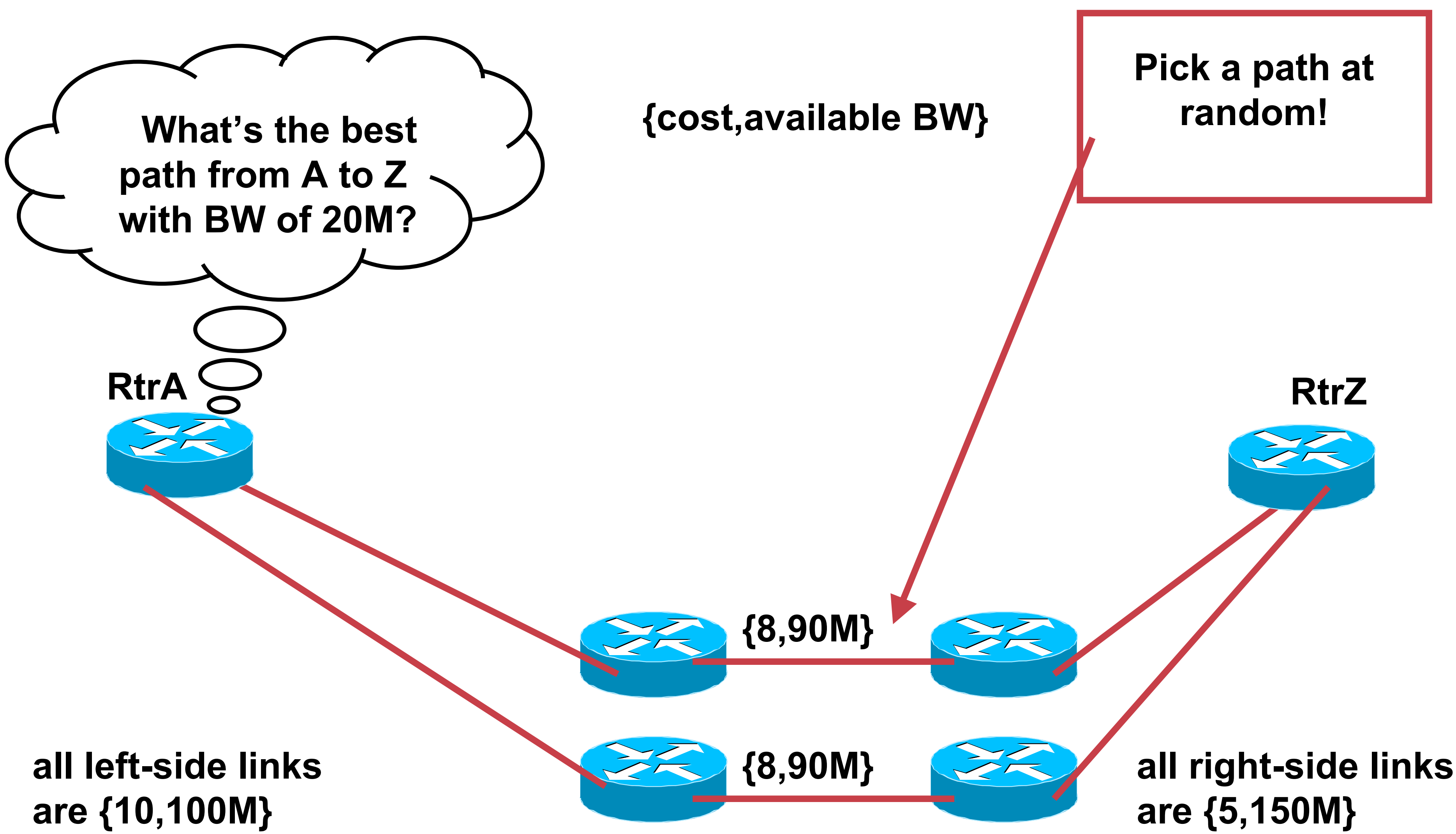


Path Calculation

Cisco.com

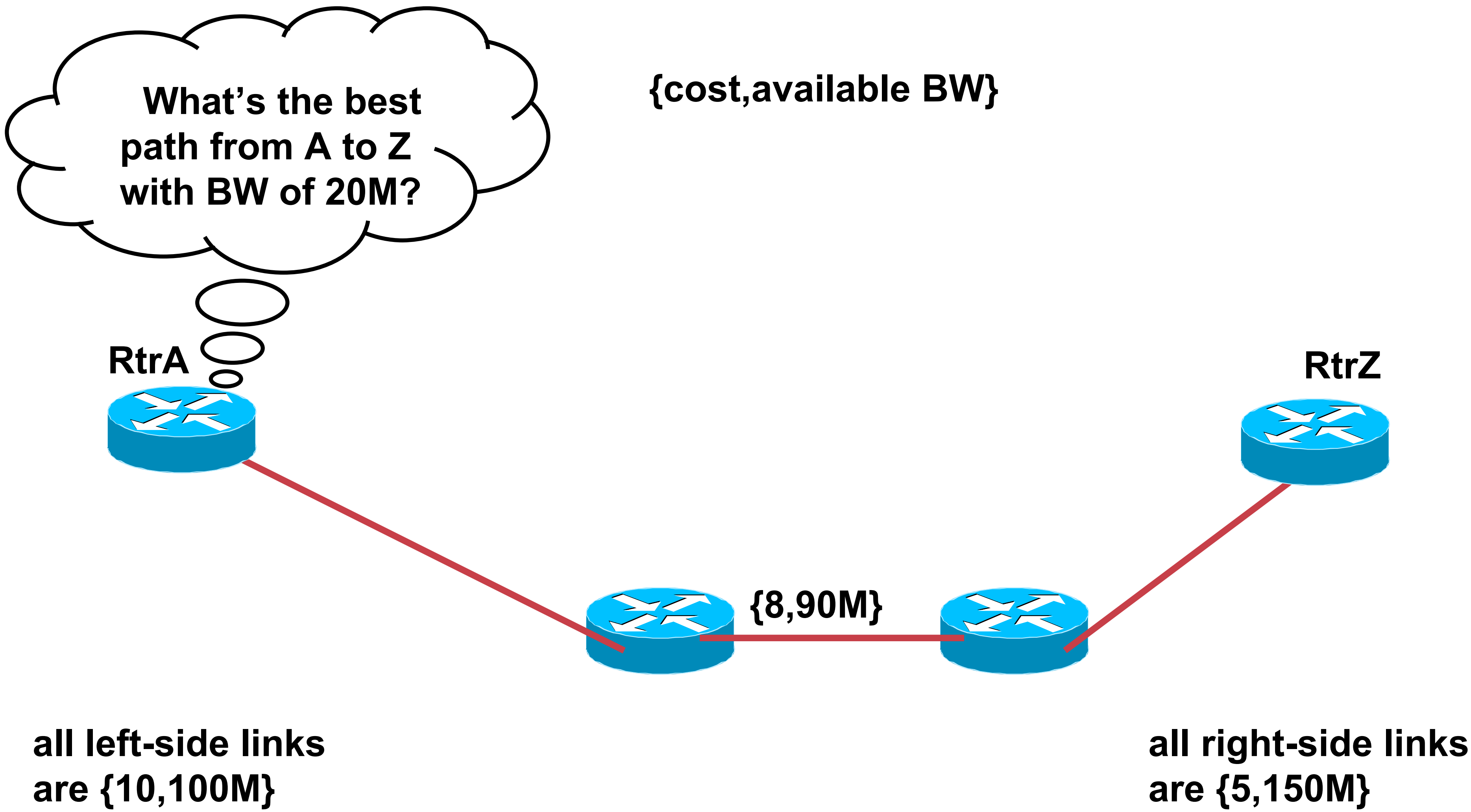


Path Calculation



Path Calculation

Cisco.com



TE Basics

- Information Distribution
- Path Calculation
- Path Setup
- Forwarding Traffic Down Tunnels

Path Setup

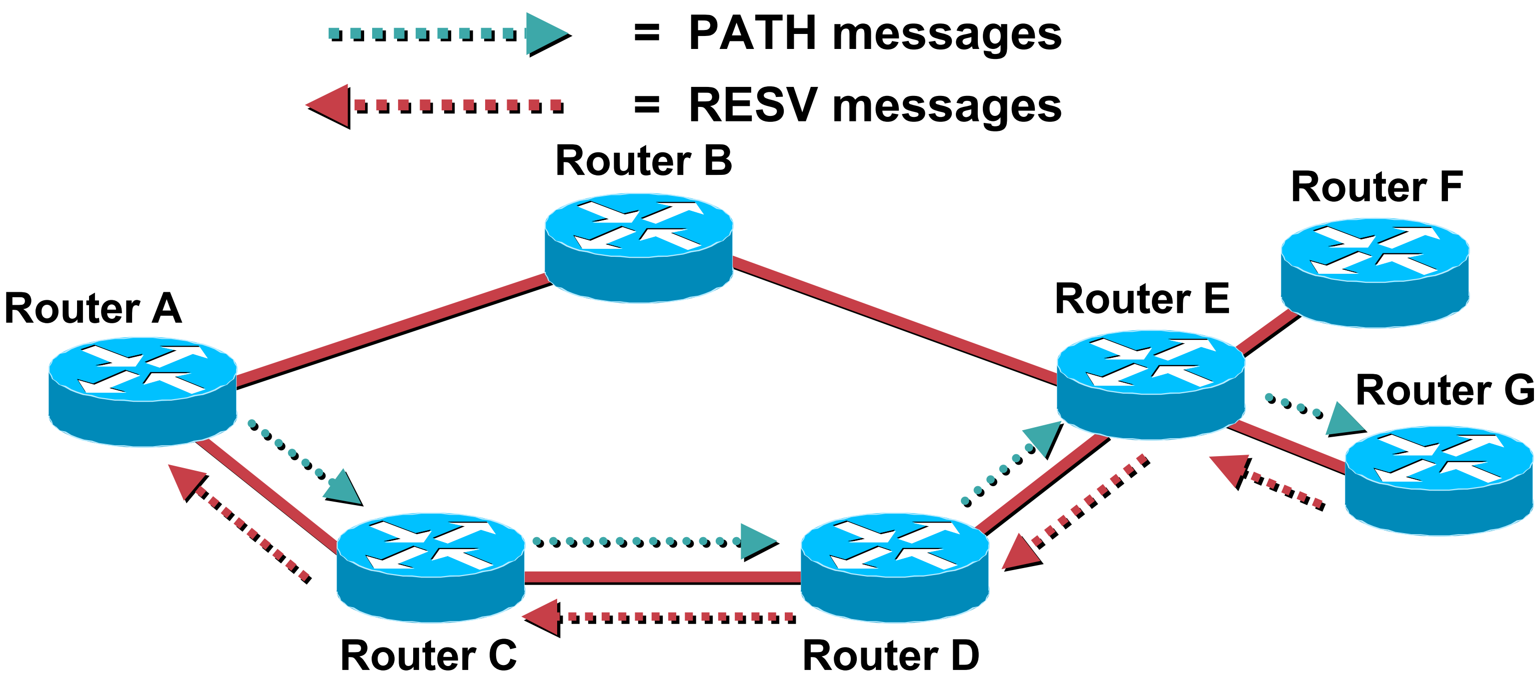
- MPLS-TE uses RSVP
- RFC2205 (base RSVP), RFC 3209 (TE extensions for RSVP)
- CR-LDP is Dead.

Path Setup

- **Once the path is calculated, it is handed to RSVP**
- **RSVP uses PATH and RESV messages to request an LSP along the calculated path**

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 Setup

Cisco.com

- **Errors along the way will trigger RSVP errors**
- **May also trigger re-flooding of TE information if appropriate**

TE Basics

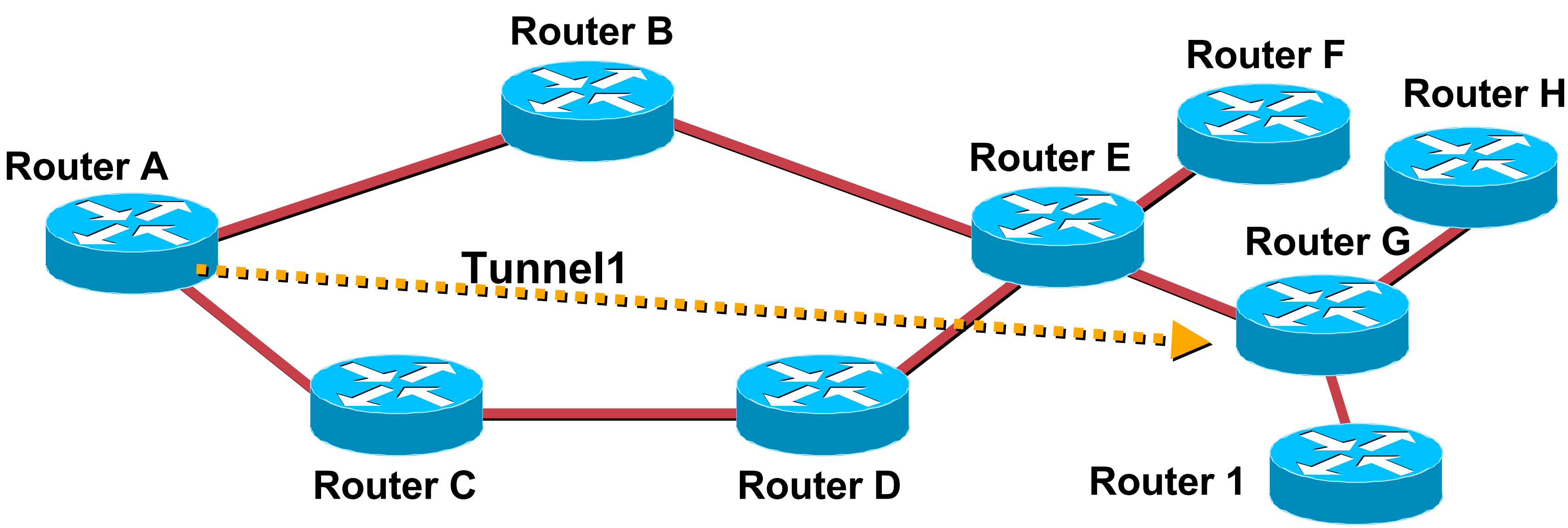
- Information Distribution
- Path Calculation
- Path Setup
- Forwarding Traffic Down Tunnels

Forwarding Traffic Down a Tunnel

- There are four ways traffic can be forwarded down a TE tunnel
  - Static routes
  - Policy routing
  - Auto-route
  - Forwarding-adjacency
- With all but PBR, MPLS-TE gets you unequal cost load balancing

Static Routing

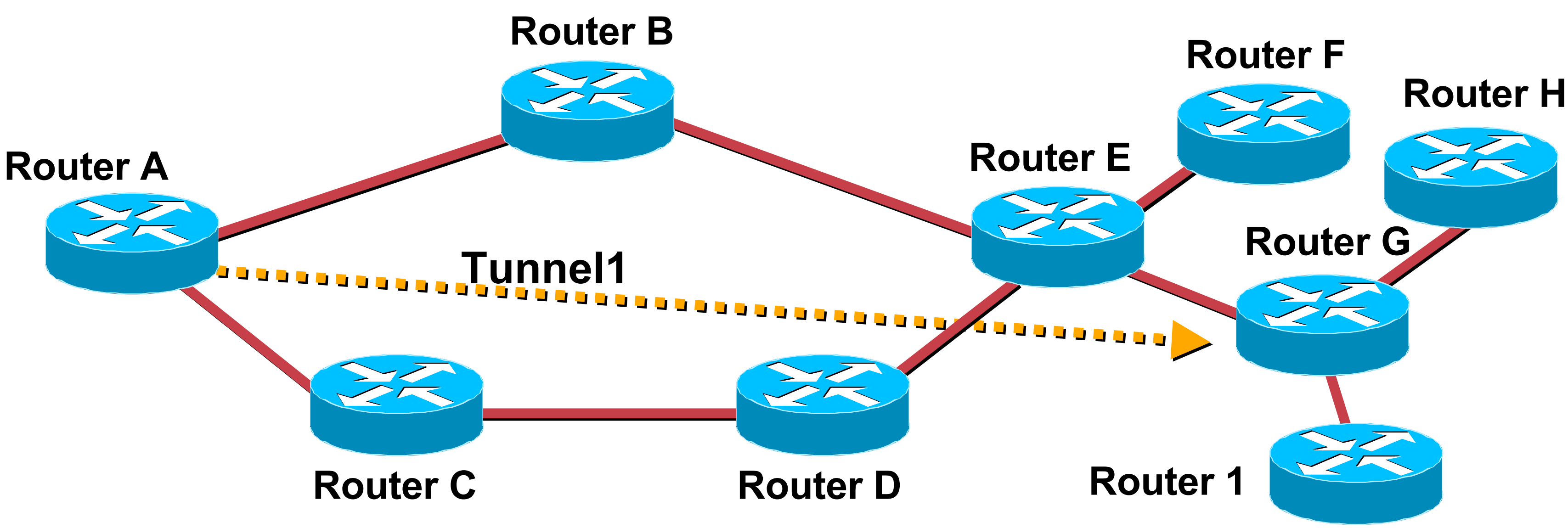

```
RtrA(config)#ip route H.H.H.H
  255.255.255.255 Tunnel1
```



Static Routing

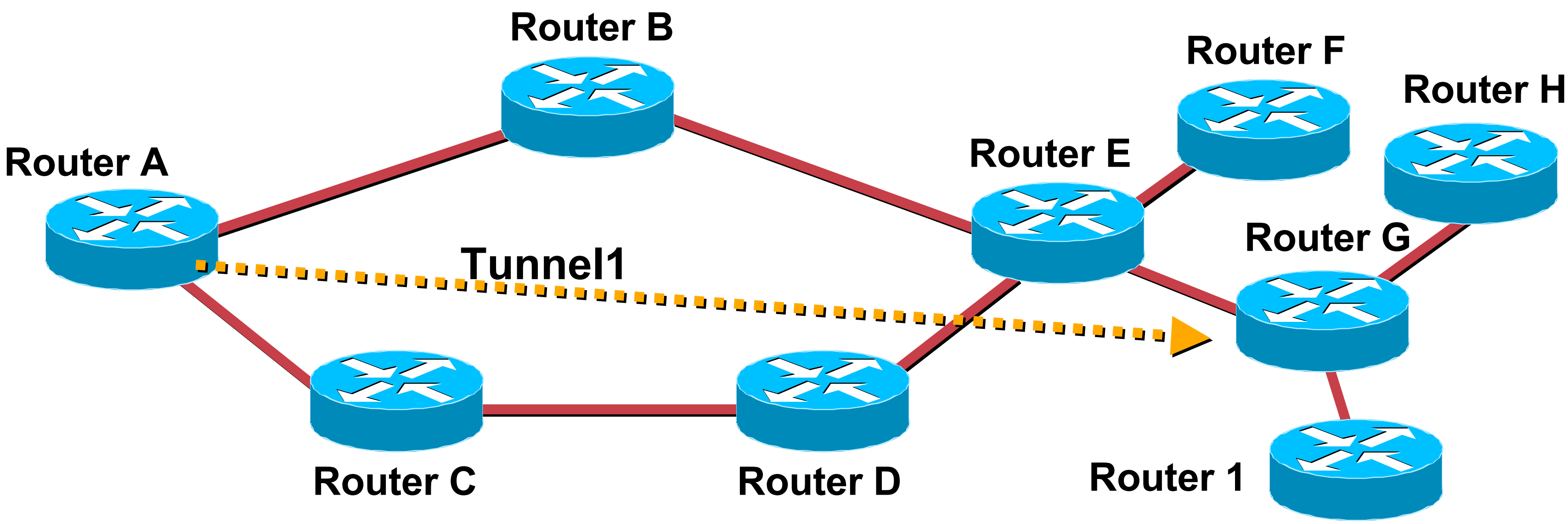
Node	Next-Hop	Cost
B	B	10
C	C	10
D	C	20
E	B	20
F	B	30
G	B	30
H	Tunnel 1	40
I	B	40

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



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
```

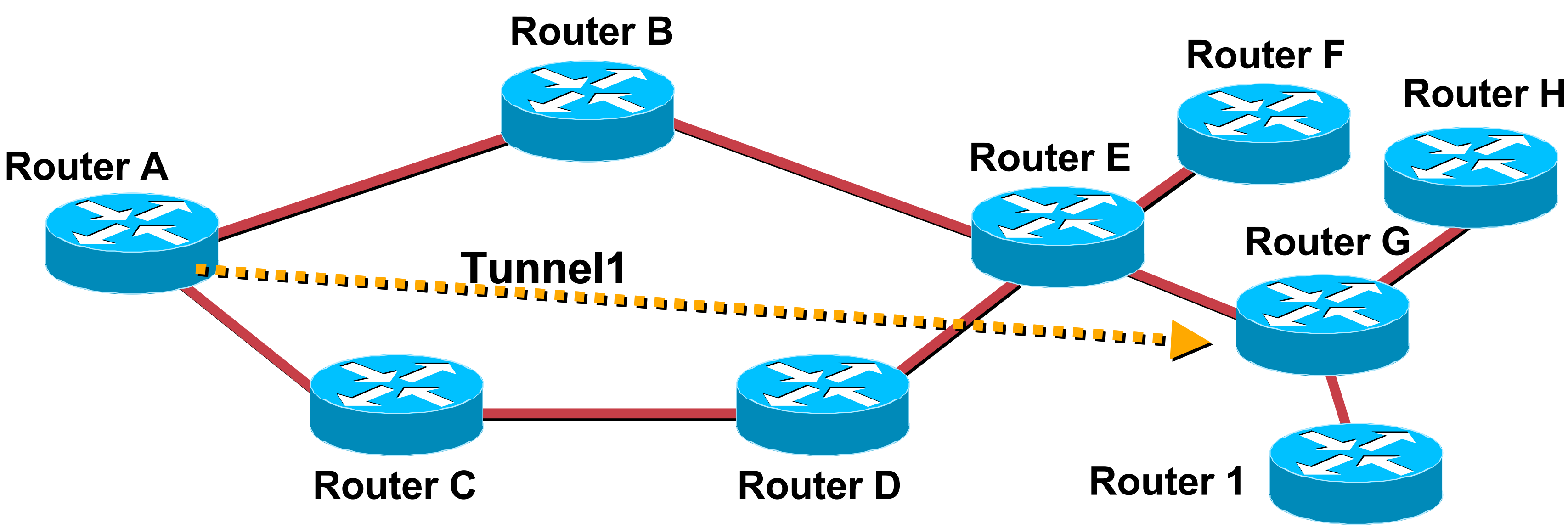


Policy Routing

Cisco.com

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
- Need (12.0(16)ST or 12.2T) or higher for 'set interface tunnel' to work (CSCdp54178)

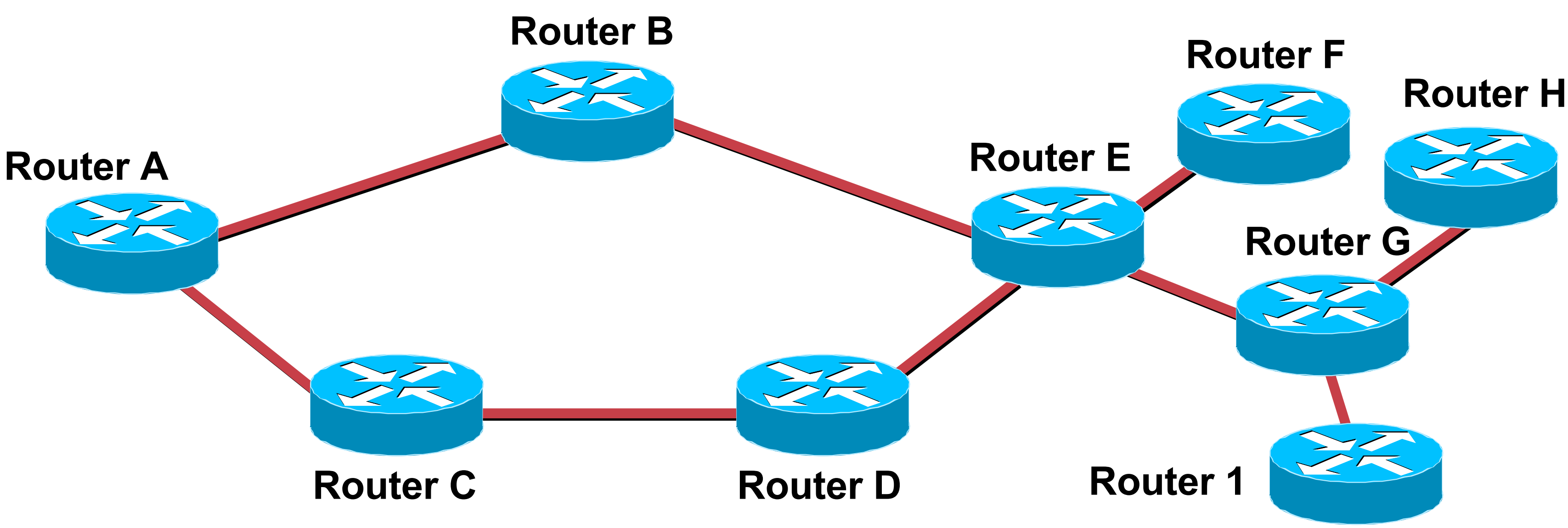


Auto-Route

- Auto-route = "Use the tunnel as a directly connected link for SPF purposes"
- This is not the CSPF (for path determination), but the regular IGP SPF (route determination)
- Behavior is intuitive, operation can be confusing

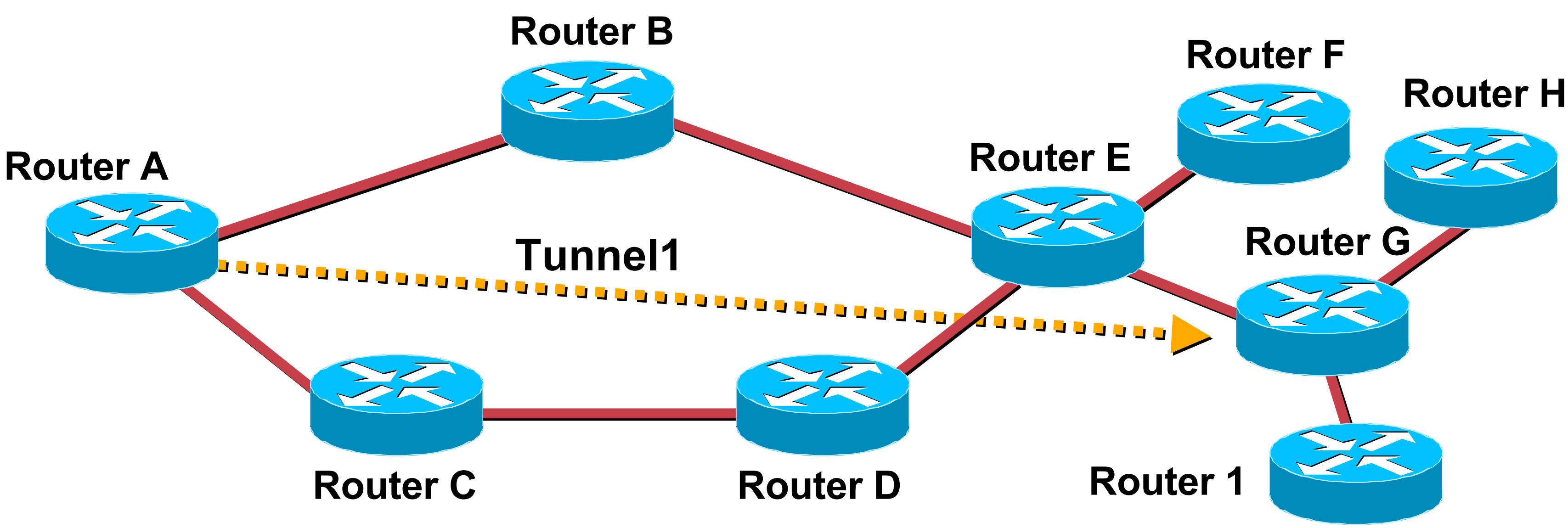
Auto-Route

This Is the Physical Topology



Auto-Route

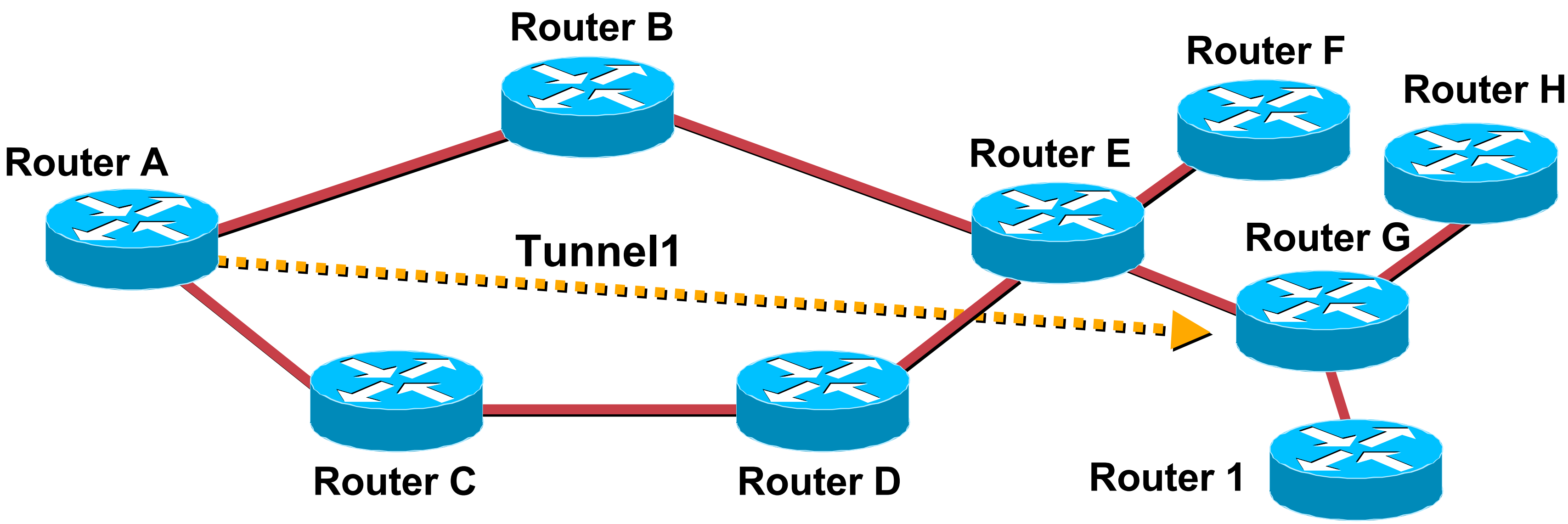
- This is Router A's logical topology
- By default, other routers don't see the tunnel!



Auto-Route

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



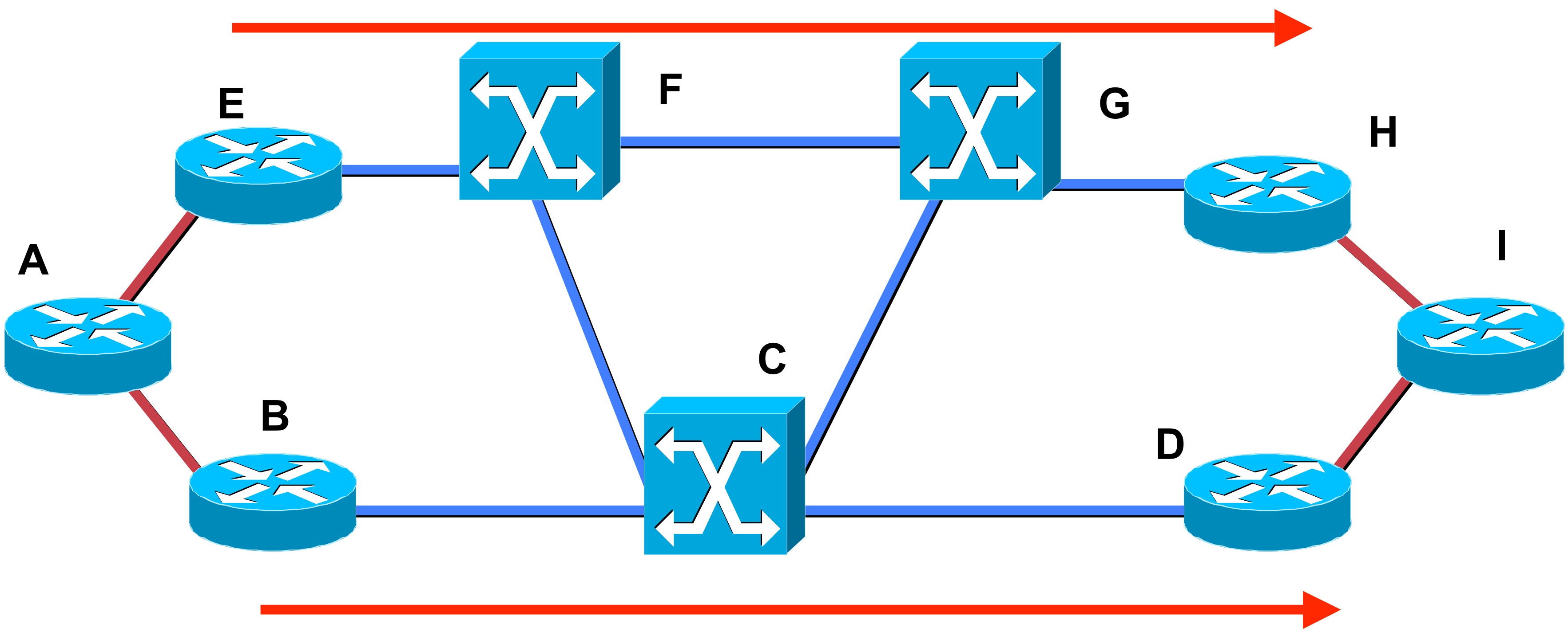
Forwarding Adjacency

Cisco.com

- Autoroute metric change is purely local to the headend
- This makes MPLS TE different from TE with ATM
  - In ATM TE, the TE link (PVC) has its cost and neighbor advertised into the network
  - In MPLS TE, no such thing is done

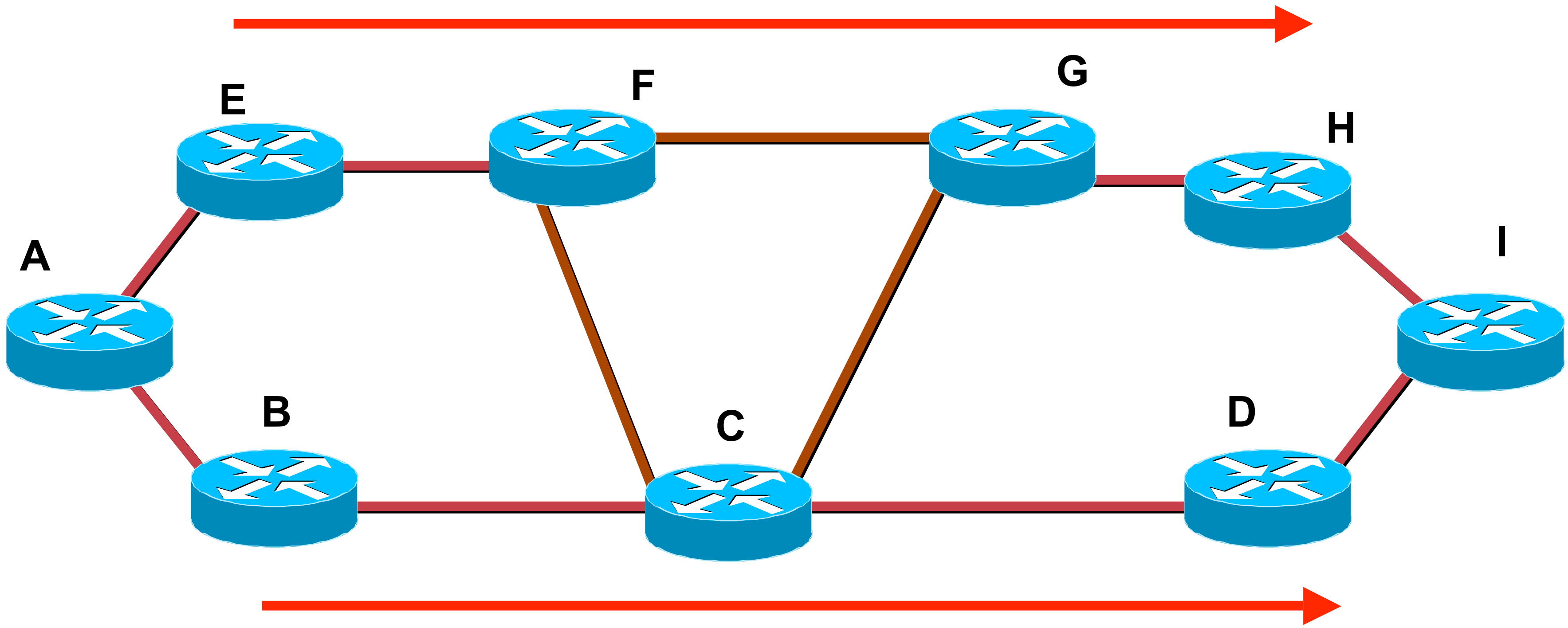
ATM model

Cisco.com



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

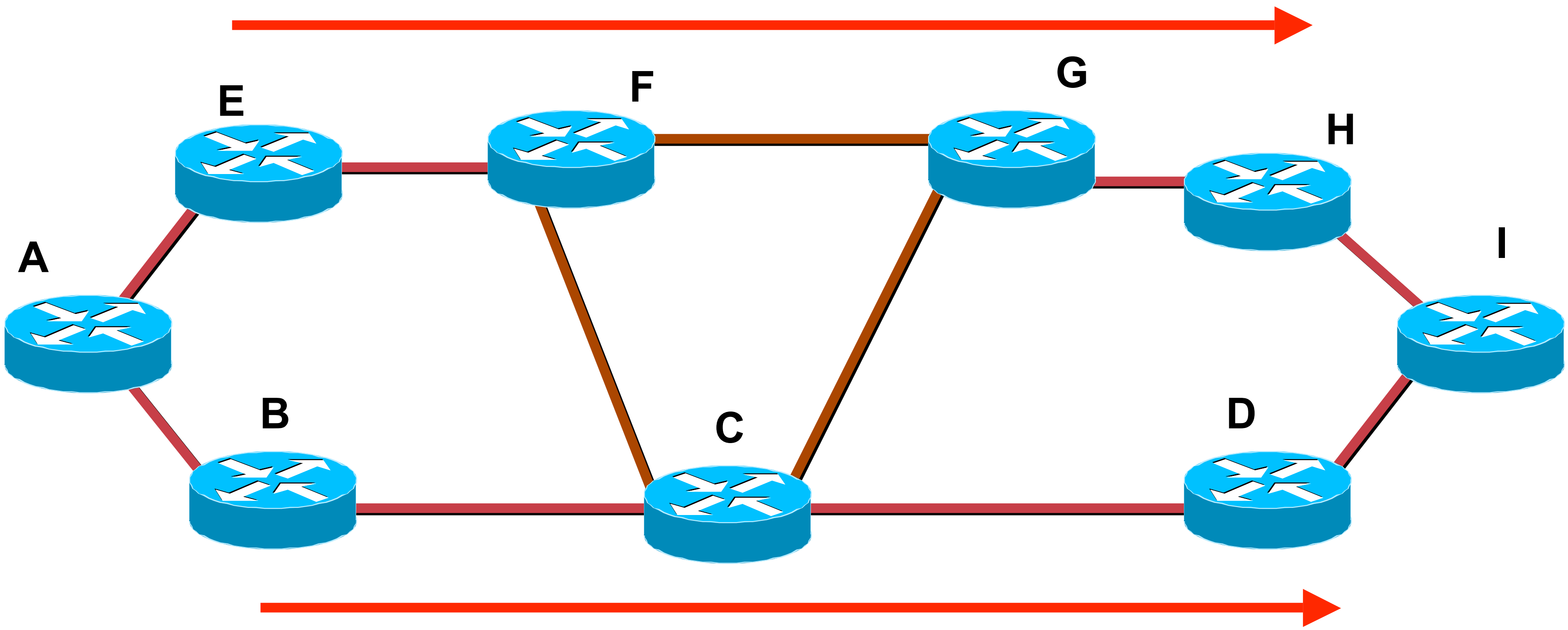
before FA



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

F-A advertises TE tunnels in the IGP

Cisco.com



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

F-A issues

- In order for A to use F-A links, they need to be the best cost IGP path
  - otherwise the physical topo gets used
- F-A configured with

```
tunnel mpls traffic-eng forwarding-adjacency
isis metric <x> level-<y>
```

F-A issues

Cisco.com

- F-A must be bidirectional
- IGP adjacency still not run over TE tunnel
- F-A cost should probably be lower than lowest possible IGP path from head to tail, otherwise it might not always get used

Unequal Cost Load Balancing

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

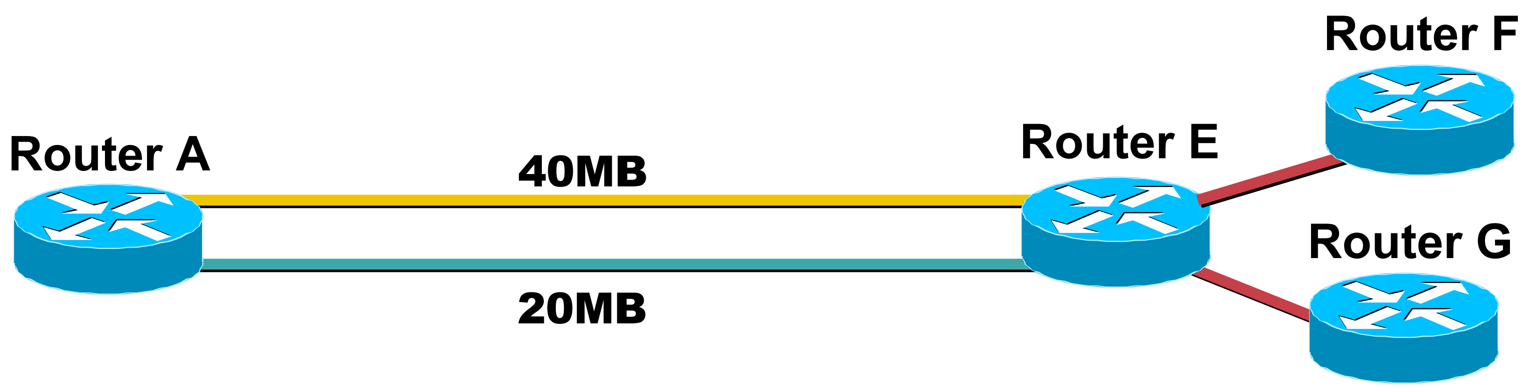
***EIGRP Has ‘Variance’, but That’s Not As Flexible**

Unequal Cost Load Balancing

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

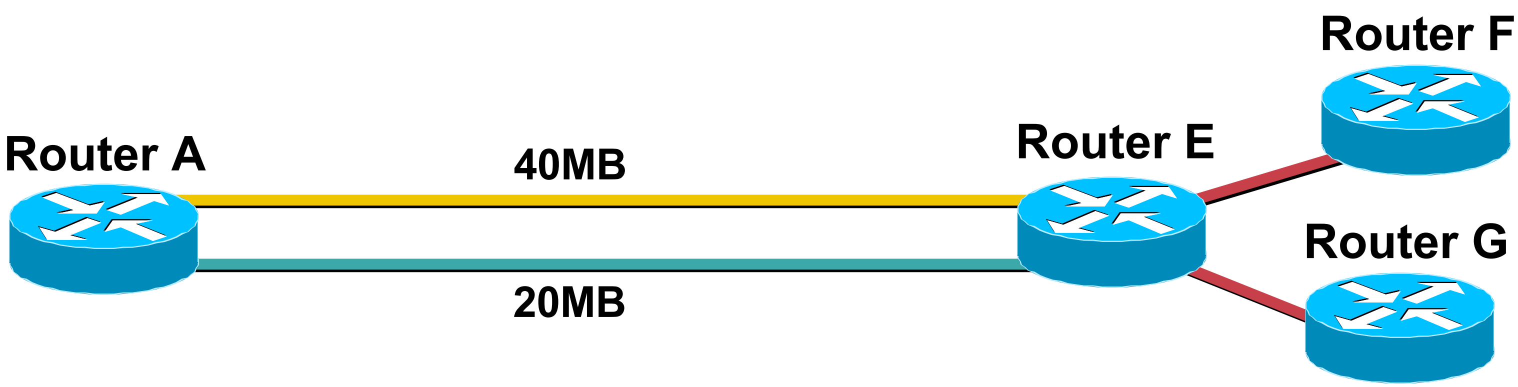
Unequal Cost: Example 1

Cisco.com




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

Unequal Cost: Example 1

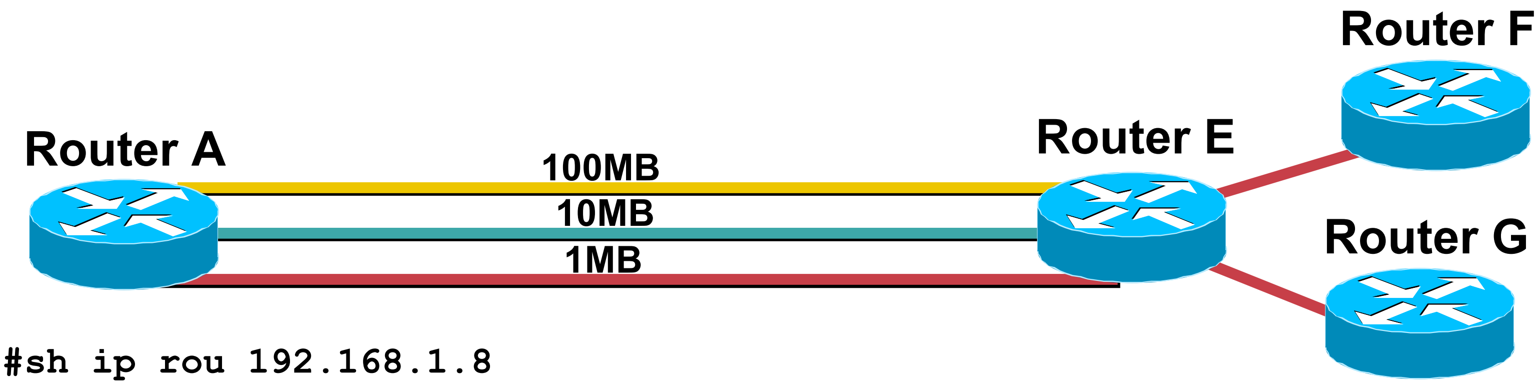



```
gsr1#sh ip cef 192.168.1.8 internal
.........
Load distribution: 0 1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 (refcount 1)
  Hash  OK  Interface                  Address          Packets  Tags imposed
  1     Y   Tunnel0                    point2point            0    {23}
  2     Y   Tunnel1                 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

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

Unequal Cost: Example 2

Cisco.com

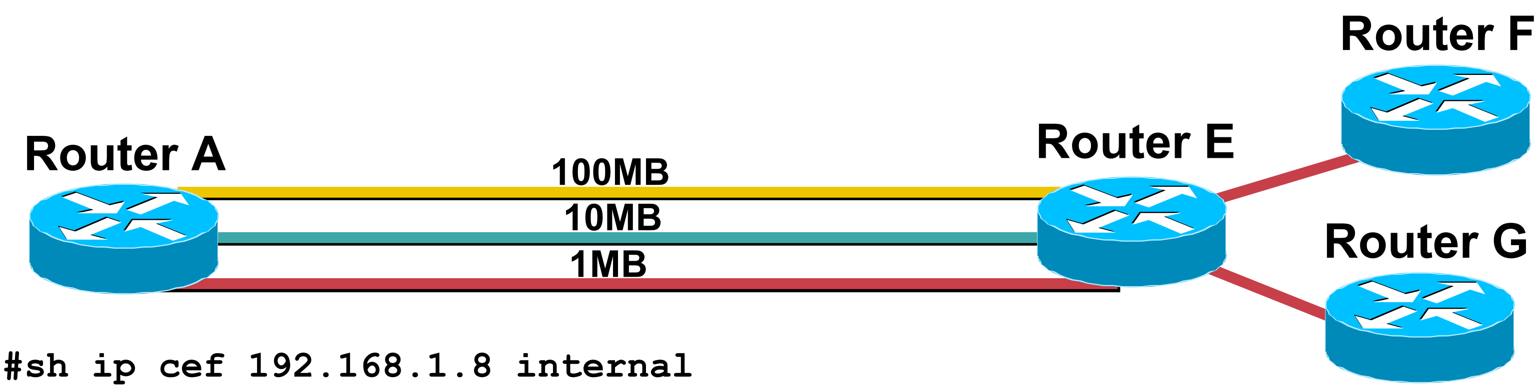



```
gsr1#sh ip rou 192.168.1.8
Routing entry for 192.168.1.8/32
  Known via "isis", distance 115, metric 83, type level-2
  Redistributing via isis
  Last update from 192.168.1.8 on Tunnel2, 00:00:08 ago
  Routing Descriptor Blocks:
  * 192.168.1.8, from 192.168.1.8, via Tunnel0
      Route metric is 83, traffic share count is 100
    192.168.1.8, from 192.168.1.8, via Tunnel1
      Route metric is 83, traffic share count is 10
    192.168.1.8, from 192.168.1.8, via Tunnel2
      Route metric is 83, traffic share count is 1
```

Q: How Does 100:10:1 Fit Into a 16-Deep Hash?

Unequal Cost: Example 2

Cisco.com




```
gsr1#sh ip cef 192.168.1.8 internal
.........
 Load distribution: 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 (refcount 1)

  Hash   OK   Interface                 Address          Packets  Tags imposed
  1      Y    Tunnel0                   point2point            0    {36}
  2      Y    Tunnel1                   point2point            0    {37}
.........
```

A: Any Way It Wants to! 15:1, 14:2, 13:2:1, It Depends on the Order the Tunnels Come Up

Deployment Guideline: Don't Use Tunnel Metrics That Don't Reduce to 16 Buckets!

Forwarding Traffic down a Tunnel

- You can use any combination of auto-route, forwarding-adjacency, static routes, or PBR
- …But simple is better unless you have a good reason
- Recommendation: autoroute, forwarding-adjacency, or statics to BGP next-hops, depending on your needs

Agenda

- How MPLS-TE Works
- Design Guidelines
- Fast ReRoute

Design Guidelines

- **Deployment methodologies**
- **Scalability**

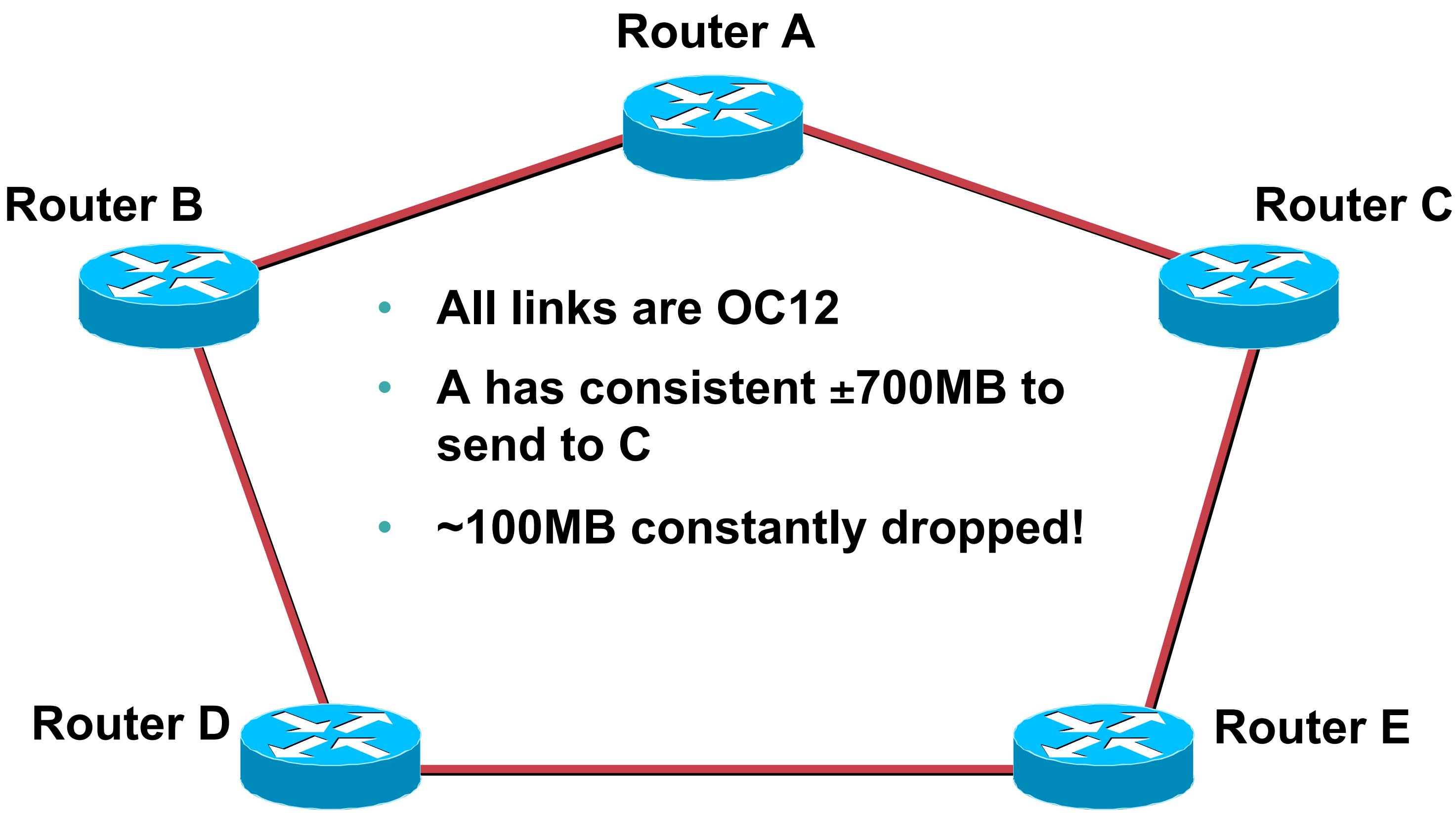
Deployment Methodologies

- **Two ways to deploy MPLS-TE**
  - **As needed to clear up congestion - tactical**
  - **Full mesh between a set of routers - strategic**
- **Both methods are valid, both have their pros and cons**

Tactical

Cisco.com

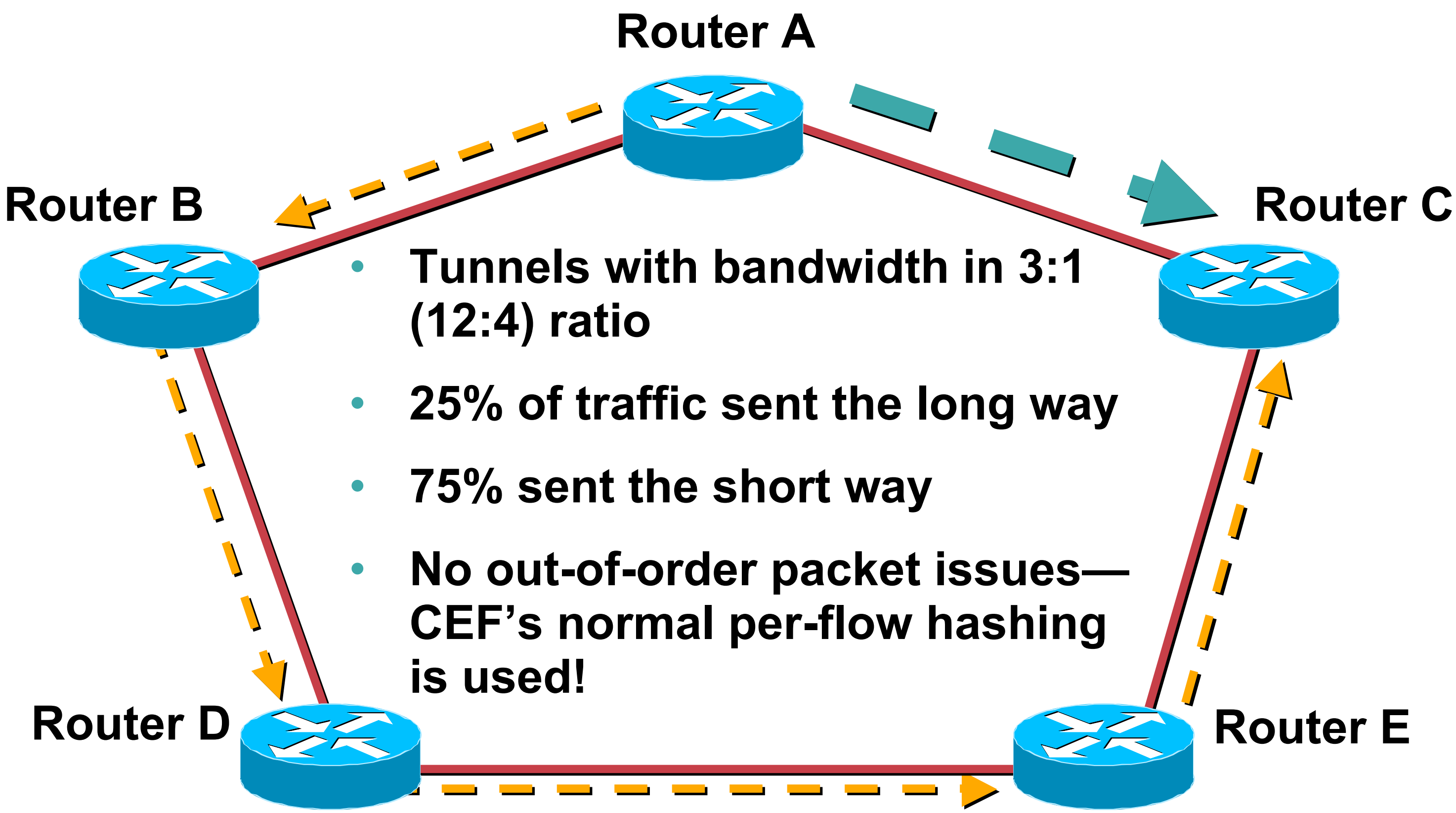
Case Study: A Large US ISP



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

Tactical

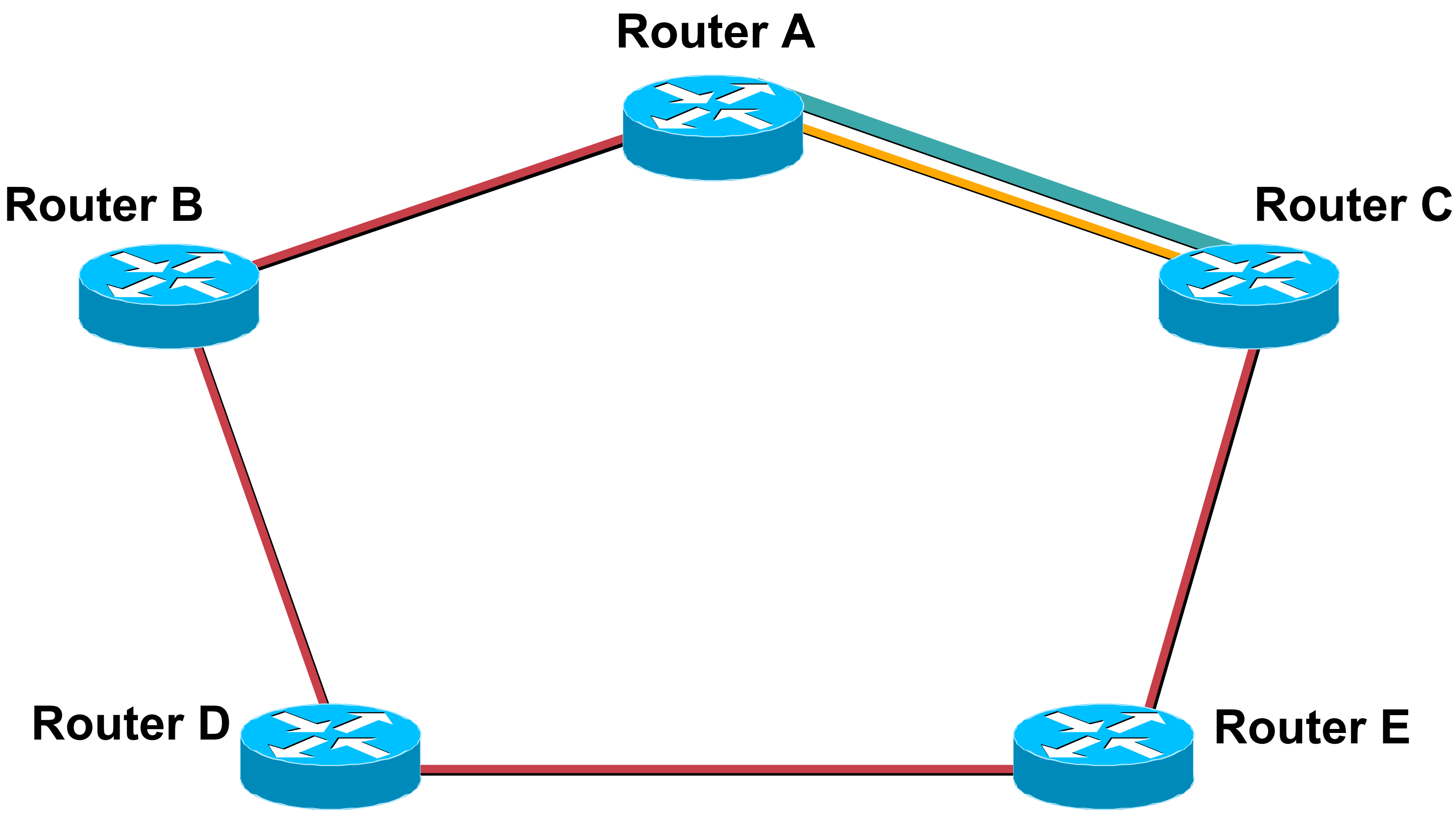
- **Solution: Multiple tunnels, unequal cost load sharing!**



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

Tactical

- **From Router A's perspective, topology is:**



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
- Link protection pretty straightforward, node protection much harder to track

Strategic

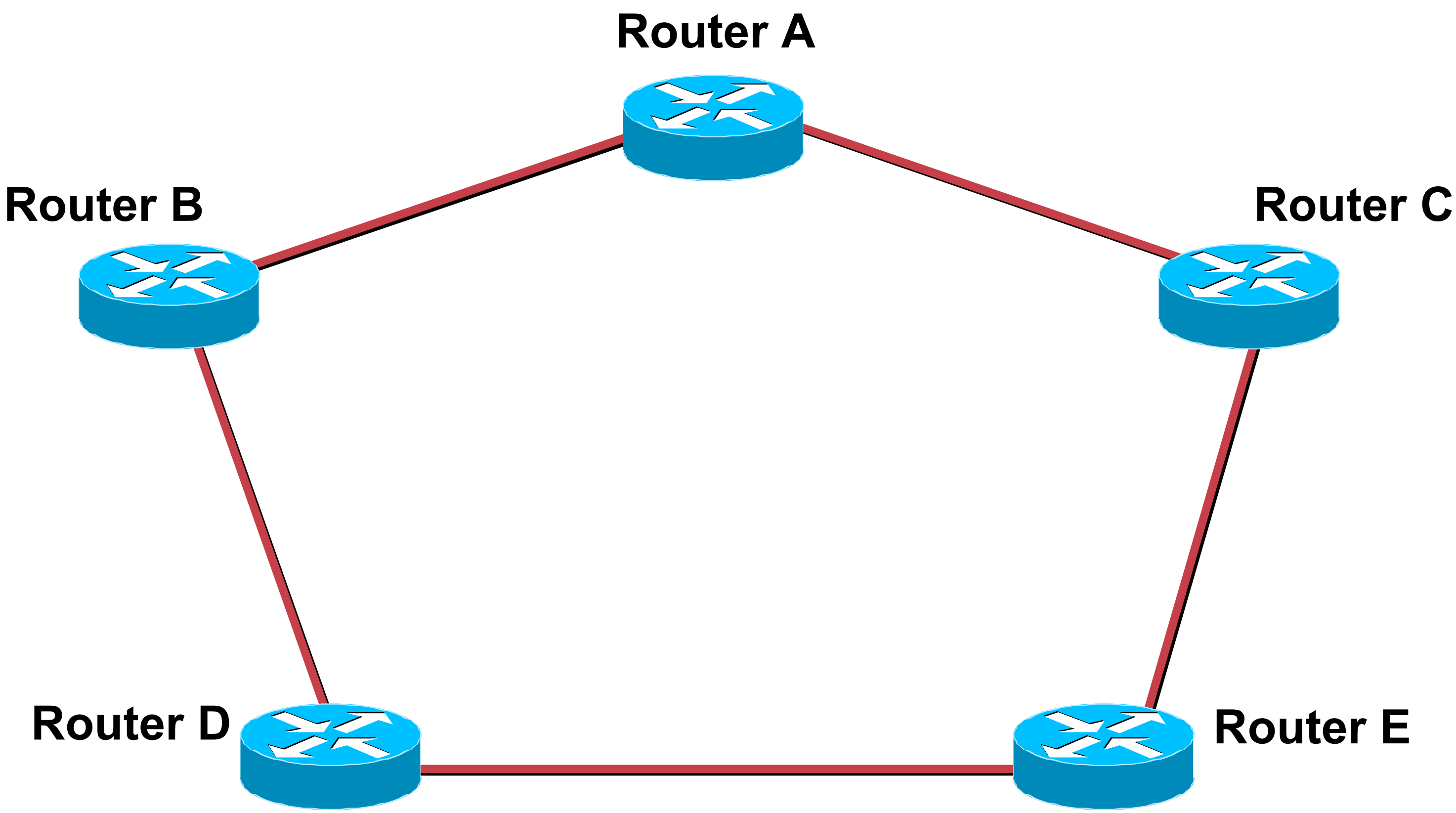
- Rather than tunnels as needed, provision a full mesh of TE tunnels
- Save money by using more of what you already have and thereby deferring upgrades
- Most useful in the core (most expensive links)

Strategic

- Some folks deploy full mesh just to get router-to-router (pop-to-pop) traffic matrix
- Largest TE network ~100 routers full mesh (~10,000 tunnels)
- As tunnel bandwidth is changed, tunnels will find the best path across your network

Strategic

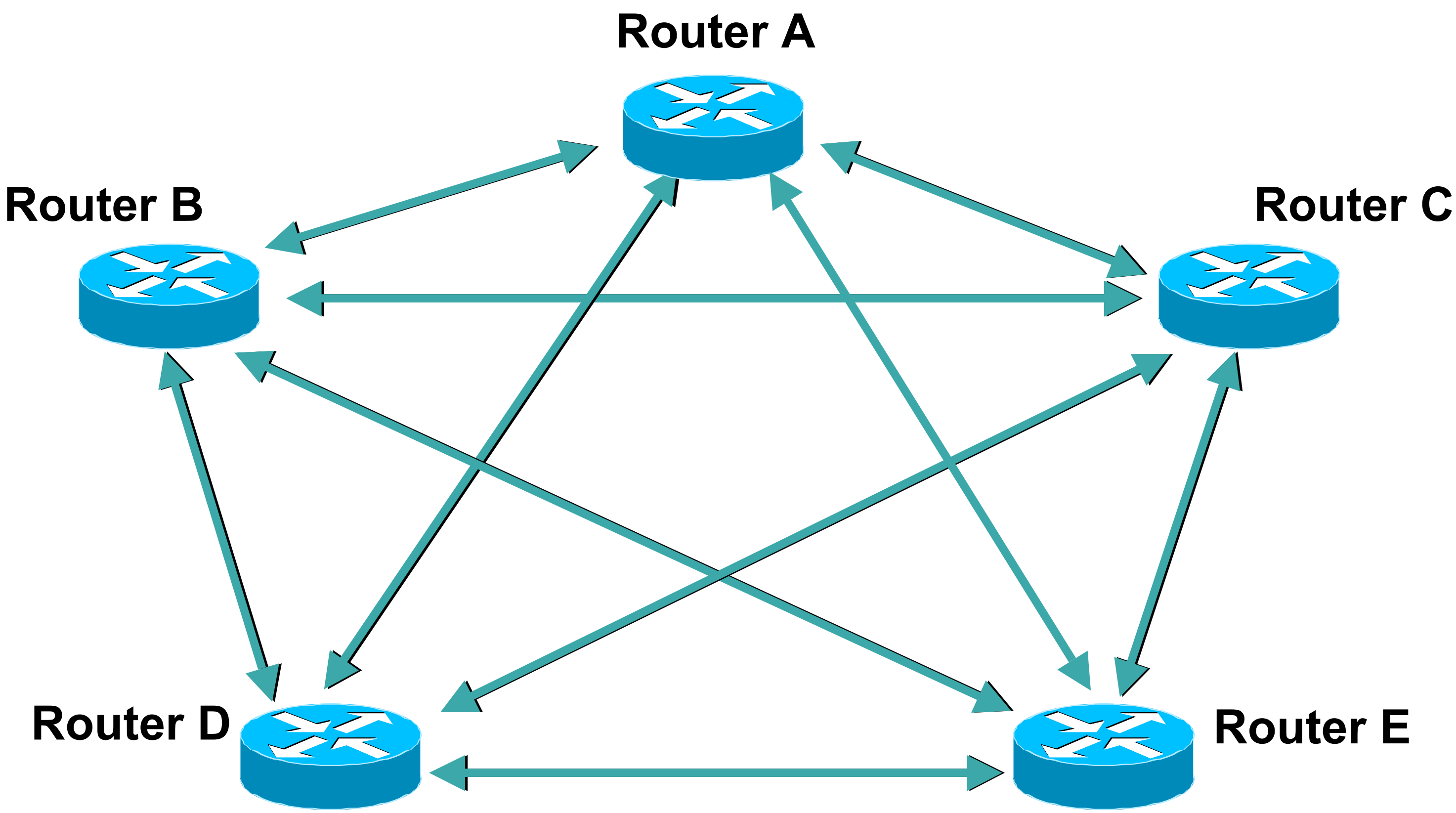
- **Physical topology is:**



Strategic

Cisco.com

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



Strategic

- **Things to remember with full mesh**

  N routers, N*(N-1) tunnels

  Routing protocols not run over TE tunnels— unlike an ATM/FR full mesh!

  Tunnels are unidirectional—this is a good thing

  …Can have different bandwidth reservations in two different directions

Strategic

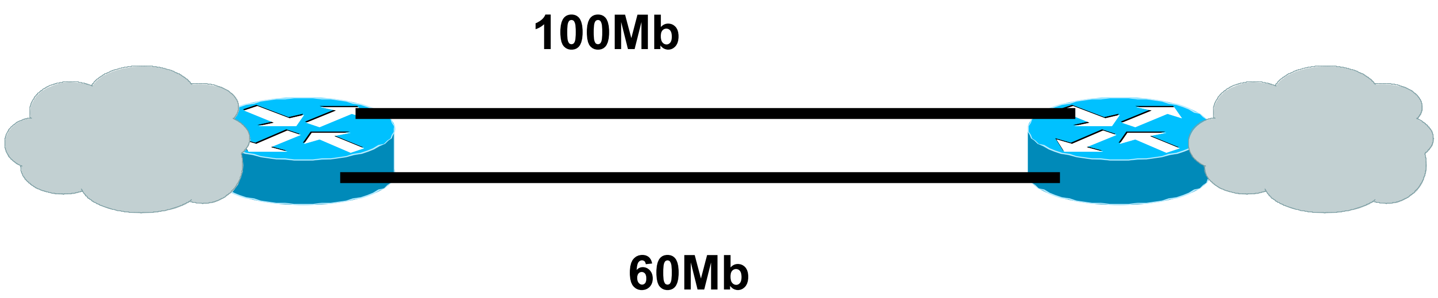
- Two ways to place full mesh tunnels
  - –Online calculation – router calculates the tunnel paths
  - –Offline calculation – an NMS or similar calculates the tunnel paths
  - –Offline is more work, more stuff, but more efficient and therefore saves more money

Strategic Offline Path Calculation

- CSPF is performed for one tunnel at a time
- Demands of multiple tunnels on the same headend are not taken into account
- Demands of multiple tunnels on *different* headends also not taken into account
- This can lead to suboptimality

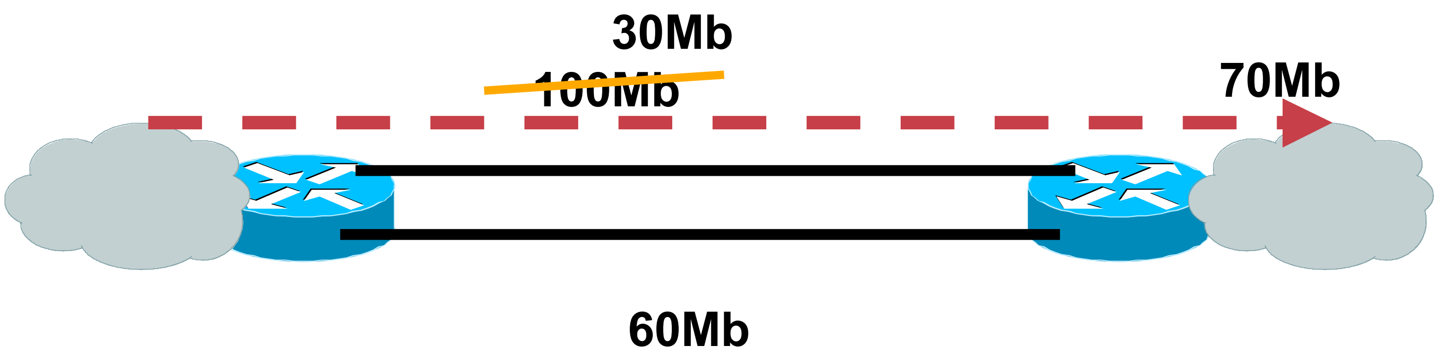
Suboptimal Tunnel Placement

Place two LSPs – one of 50Mb, one of 70Mb.



Suboptimal Tunnel Placement

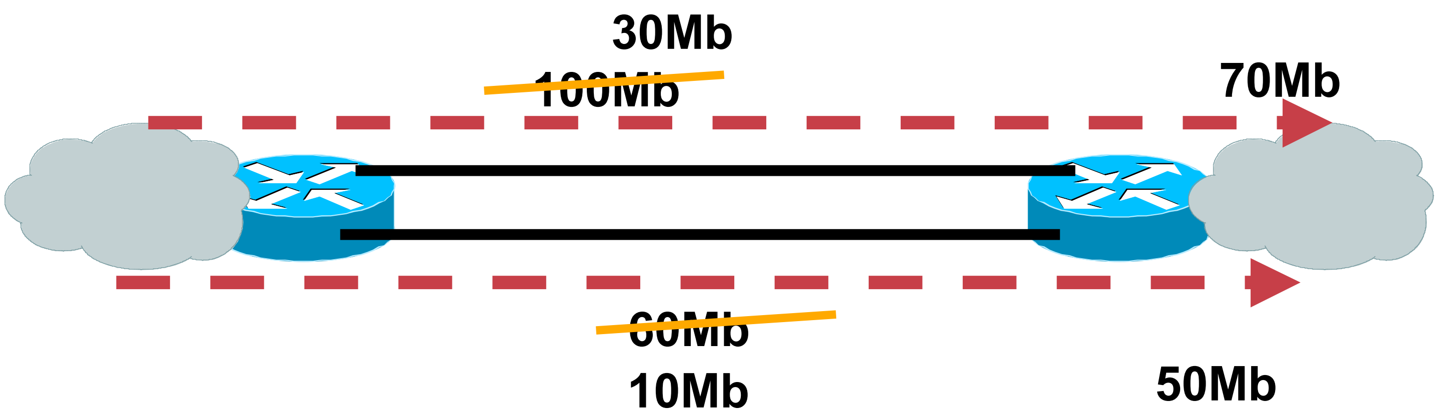
case 1: 70Mb LSP comes up first



Suboptimal Tunnel Placement

Cisco.com

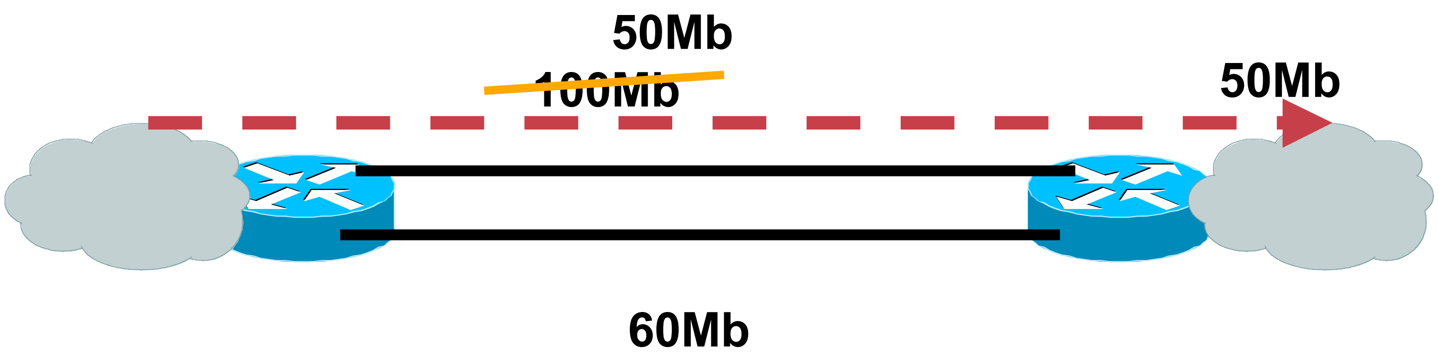
case 1: 70Mb LSP comes up first
then 50Mb LSP comes up



Everything's OK.

Suboptimal Tunnel Placement

case 2: 50Mb LSP comes up first



Where do we put the 70Mb LSP???

Suboptimal Tunnel Placement

- With only 2 tunnels and 2 links, if we change the TE tie-breaker to “link with lowest available BW”, the previous scenario will be OK.
- With more tunnels than links, we’re still potentially out of luck
- If the link have different metrics, we’re still out of luck
- Need offline tool that knows about all resources and all demands
- WANDL makes one, some customers make their own tools, etc.
- See also Tunnel Builder

Deploying and Designing

- Deployment methodologies
- Scalability

Scalability

How Many Tunnels on a Router?

Code	Number of Head-End Tunnels	Number of Mid-Points	Number of
12.OST	600	10,000	5,000

- Tests were done on a GSR
- RSP4, RSP8, VXR300, VXR400 will be similar

Scalability

Cisco.com

- Largest TE network today = 100 routers, ~10,000 tunnels full mesh
- 12.0ST—600 head-ends, 360,000 tunnels full mesh with 10,000 tunnels per midpoint
- 600 = 100*6

  Or (360,000=10,000*36) if you're in marketing
- Bottom line: MPLS-TE is not a gating factor in scaling most networks!

Scalability

- **The 600/10,000/5,000 numbers are probably pessimistic**
- **RFC2961 (RSVP Refresh) will greatly increase these numbers**
- **The bottleneck is sending lots of RSVP messages**

Agenda

Cisco.com

- **How MPLS-TE Works**
- **Design Guidelines**
- **Fast ReRoute**

Fast ReRoute

- Introduction
- Terminology of Protection/Restoration
- MPLS Traffic Engineering Fast Reroute
- Conclusion

Protection/restoration in IP/MPLS networks

Cisco.com

- Many various protection/restoration schemes (co)exist today:
  - Optical protection
  - Sonet/SDH
  - IP
  - MPLS Traffic Engineering Fast Reroute
- The objective is to avoid double protection

Protection/Restoration in IP/MPLS networks

- IP routing protocol typically offers a convergence on the order of seconds (anywhere from a couple of secs to 30-40 secs)
- IP restoration is Robust and protects against link AND node protection
- IP convergence may be dramatically improved and could easily offer a few seconds convergence (1, 2, 3, sub-secs?) using various enhancements:

  fast fault detection,

  fast SPF and LSA propagation triggering,

  priority flooding,

  Incremental Dijsktra,

  Load Balancing

Protection/Restoration in IP/MPLS networks

- Couple of secs may be sufficient for some traffic but others (ex: voice trunking) will require more aggressive target, typically 50 ms.
- Solutions ?
  - Optical protection,
  - Sonet/SDH (GR 253)
  - MPLS protection/restoration

Protection/Restoration in IP/MPLS networks

MPLS Traffic Engineering Protection/Restoration

- Compared to lower layer mechanisms, MPLS offers:
  - A protection against link AND node failures
  - A much better bandwidth usage
  - Finer granularity. Different level of protection may be applied to various classes of traffic.
    - Ex: an LSP carrying VoIP traffic will require a 50ms protection scheme as Internet traffic may rely on IP convergence
  - A more cost effective protection mechanism

Fast ReRoute

Cisco.com

- Introduction
- Terminology of Protection/Restoration
- MPLS Traffic Engineering Fast Reroute
- Conclusion

Protection/Restoration in MPLS networks

Cisco.com

Terminology

- **Protection: a back-up path is pre-established to be used as soon as the failure has been detected**
- **Restoration: putting traffic on an alternate path. The alternate path may or may not be pre-computed.**
- **In Cisco's Local Protection scheme Protection and Restoration are combined**

Protection/Restoration in IP/MPLS networks

Cisco.com

Scope of recovery: local repair versus global repair

- Local (link/node) repair: the recovery is being performed by the node immediately upstream to the failure

    Example

    MPLS local repair FRR (link/node protection)

- Global repair: the recovery is being performed by the head-end (where the LSP is initiated)

Protection/Restoration in IP/MPLS networks (Global Repair)

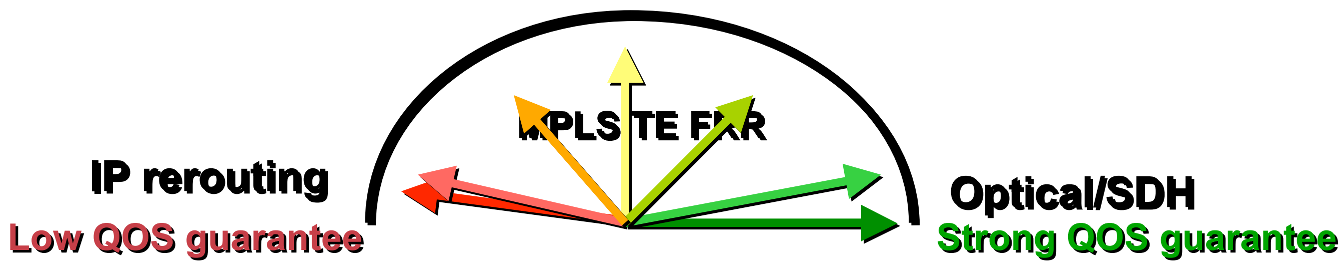
- Slower than local repair (propagation delay of the FIS may be a non negligible component)
- Examples of global repair mechanisms

  IP is a global repair mechanism using restoration. TTR is typically O(s)

  MPLS TE Path protection is a global repair mechanism

Protection/Restoration in IP/MPLS networks

- **Path mapping: refers to the method of mapping traffic from the faulty working path onto the protected path (1:1, M:N)**
- **QOS of the protected path: does the protected path offer an equivalent QOS as the working path during failure ?**



Fast ReRoute

- Introduction
- Terminology of Protection/Restoration
- MPLS Traffic Engineering Fast Reroute
- Conclusion

Terminology

Terminology

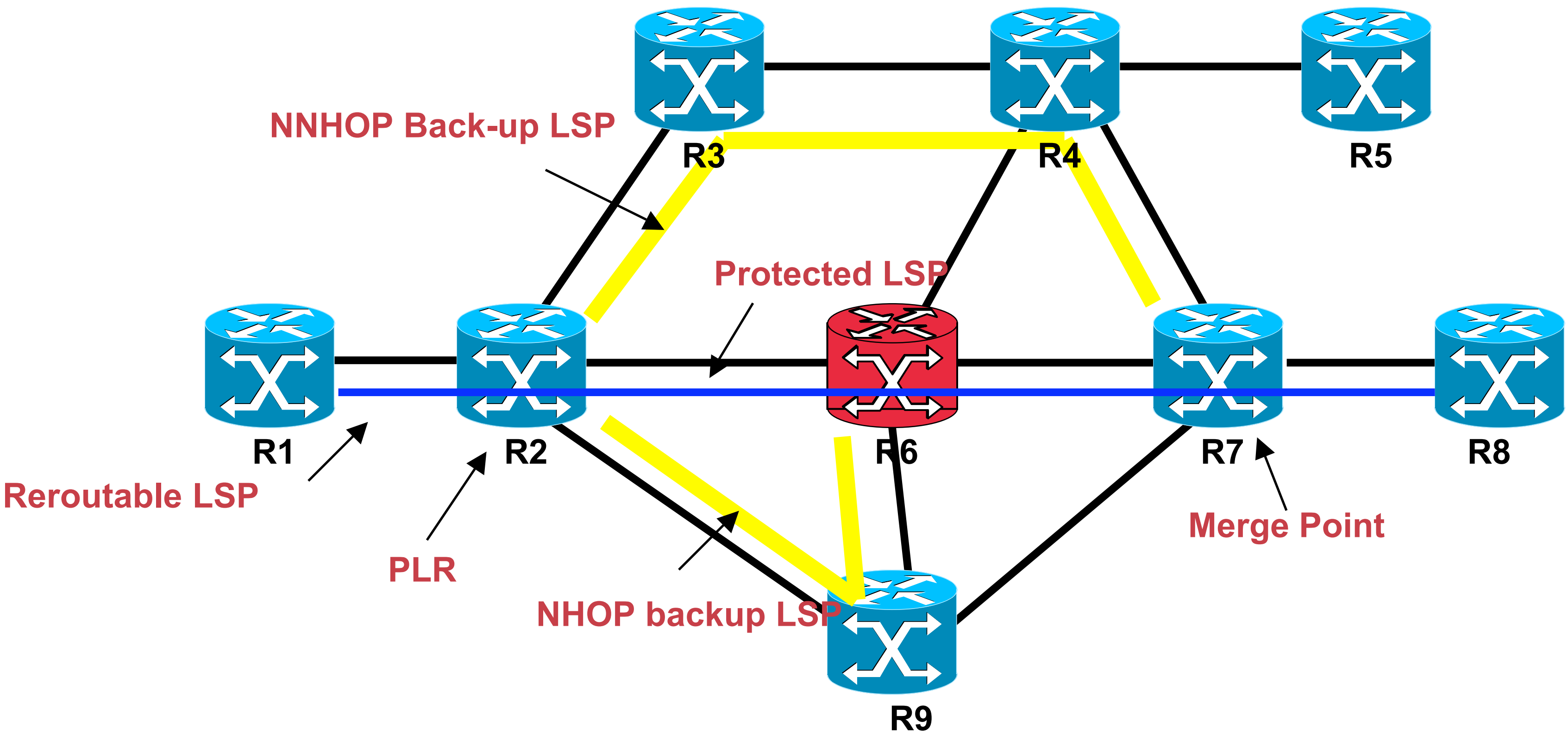
- **Reroutable LSP**: TE LSP for which a local protection is desired
- **Protected LSP**: an LSP is being protected at a HOP H if and only if it does have a backup tunnel associated at hop H.
- **PLR**: Point of local repair (head-end of the backup tunnel)
- **Backup tunnel/LSP**: TE LSP used to backup the protected LSP

Terminology

Terminology (cont)

- **Merge point**: Tail-end of the backup tunnel
- **NHOP backup tunnel**: a Backup Tunnel which bypasses a single link of the Primary Path.
- **NNHOP backup tunnel**: a Backup Tunnel which bypasses a single node of the Primary Path.

Terminology



Cisco.com

MPLS TE LSP rerouting (Global restoration)

MPLS TE rerouting

Cisco.com

Global restoration:

- **Headend LSP Reroute**
- **Path Protection (Hot Standby LSP)**

MPLS TE rerouting

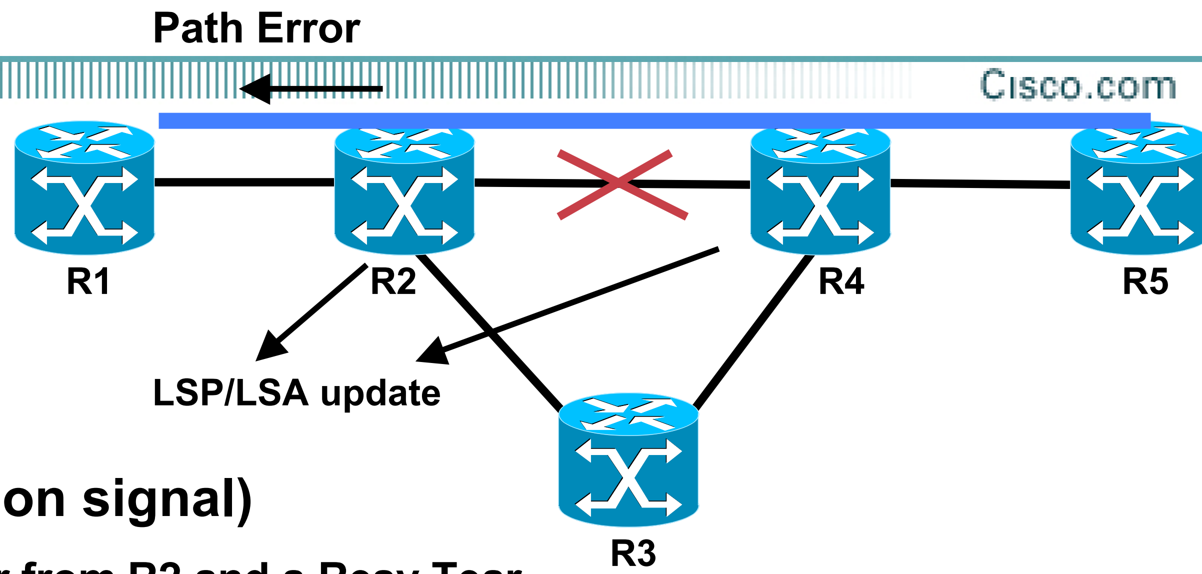
TE LSP rerouting (Global restoration)

- **Controlled by the head-end of a trunk via the resilience attribute of the trunk**
- **Fallback to either (pre)configured or dynamically computed path. Pre-configured path may be either pre-established, or established "on demand"**

```
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



Cisco.com

- **The FIS (failure indication signal)**

  *** R1 receives a Path Error from R2 and a Resv Tear**

  *** R1 will receive a new LSA/LSP indicating the R2-R4 is down and will conclude the LSP has failed (if R1 is in the same area as the failed network element)**

  Which one on those two events will happen first ? It depends of the failure type and IGP tuning

- **An optimisation of the Path Error allows to remove the failed link from the TE database to prevent to retry the same failed link (if the ISIS LSP or the OSPF LSA has not been received yet).**

  mpls traffic-eng topology holddown sigerr <seconds>

MPLS TE rerouting

- **Use RSVP pacing to limit the loss of RSVP message in case of rerouting of several TE LSP:**

  ip rsvp msg-pacing [period msec [burst msgs [max_size qsize]]]

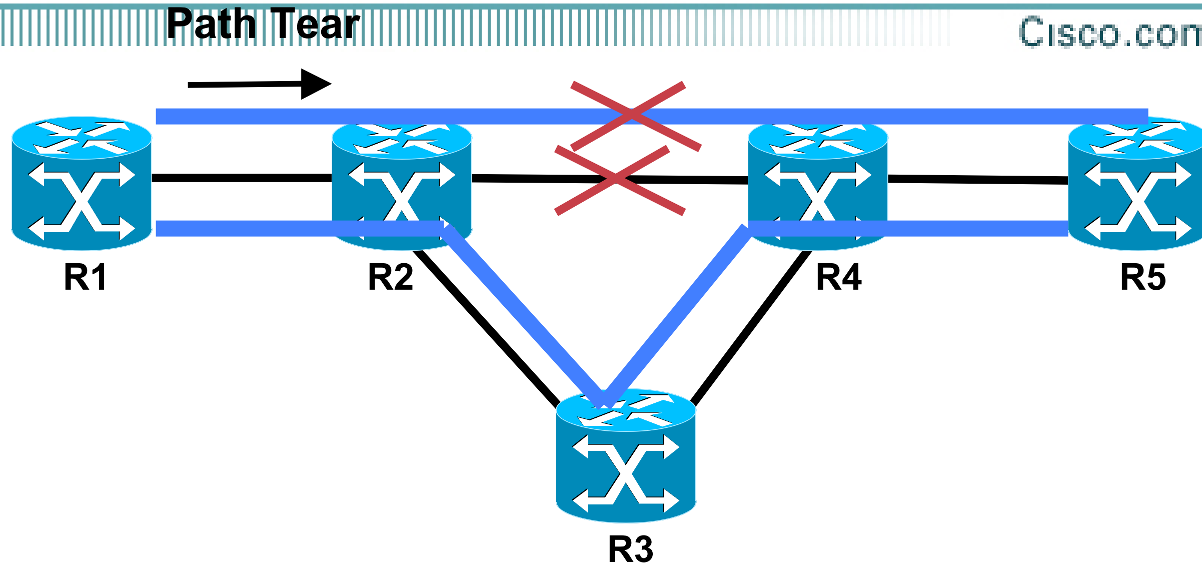
- **ISIS scanner (controls the propagation of TE information from ISIS to the TE database) may be used to speed-up convergence:**

  mpls traffic-eng scanner [interval <1-60>] [max-flash <0-200>]

  Interval: 5 seconds

  Max-flash: 15 updates

MPLS TE rerouting



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

- R1 clears 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)
- PATH Protection time = O(s).
- Fault restoration TTR = O(s).

Restoration: the head must recalculate a Path (CSPF), signal the LSP and reroute the traffic

MPLS TE Path Protection

MPLS TE Path Protection (hot standby LSP)

- **MPLS TE Path Protection is a global repair mechanism using protection switching**
- **The idea is to be able to set up a primary LSP AND a back-up LSP (pre-signalled) so once the failure has been detected and signalled (by the IGP or RSVP signalling) to the head-end the traffic can be switched onto the back-up LSP**
- **No path computation and signalling of the new LSP once the failure has been detected and propagated to the head-end (compared to LSP reroute)**

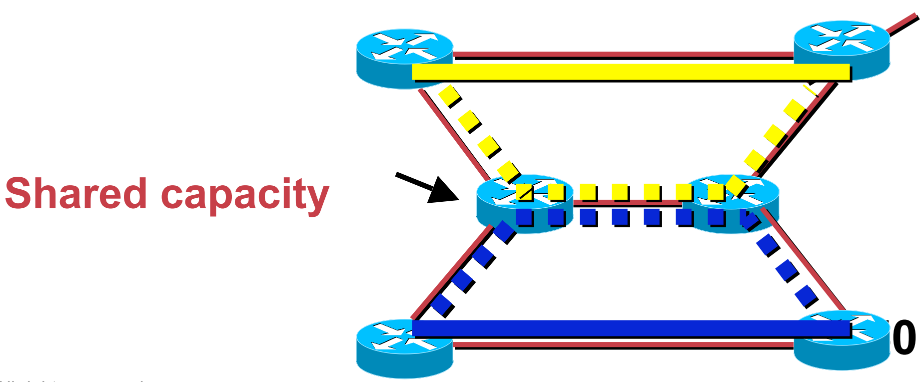
MPLS TE Path Protection

- **By configuration the TE back-up LSP attributes may or may not be different to the primary TE LSP:**
  - **The bw of the back-up LSP maybe some % of the primary bw**
  - **RCA of the back-up LSP may or may not be taken into account**
- **Diversely routed paths are calculated by the CSPF on the head-end (they may be link, node or SRLG diverse)**

MPLS TE Path Protection

- Limitation of MPLS TE Path protection
  - The fault propagation may be unacceptable especially for very sensitive traffic,
  - The number of states in the network is doubled !!
  - CSPF is likely to be highly inefficient in term of bandwidth usage.

    →primary diversely routed paths may share backup bandwidth (under the assumption of single network element failure)



MPLS TE Path Protection

- Path protection may be an attractive solution if and only if:
  - Just a few LSPs require protection
  - A few hundreds of msecs convergence time is acceptable

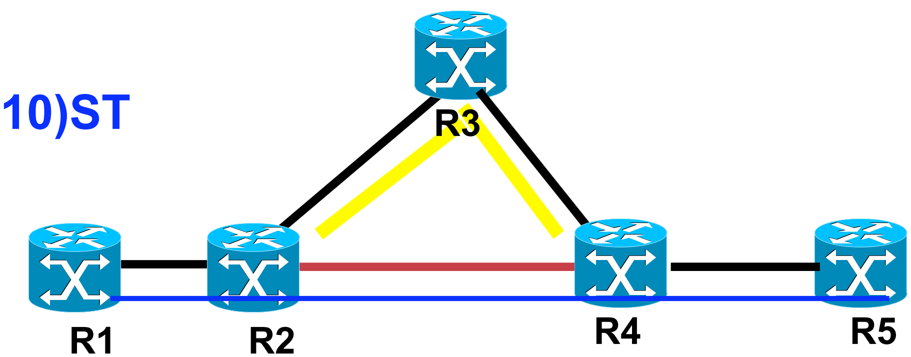
Principles of MPLS TE Fast Reroute (local protection)

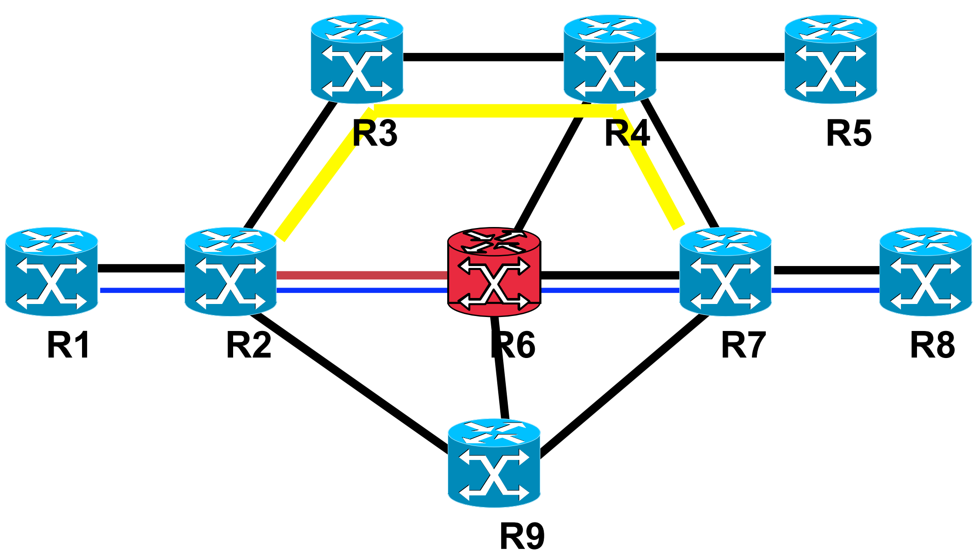
MPLS TE FRR – Local protection

MPLS Fast Reroute local repair

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

12.0(10)ST





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

12.0(22)S

MPLS TE FRR – Local protection

- **MPLS Fast Reroute link and node protection is:**
  - **LOCAL (compared to IGP or Path protection which are global protection/restoration mechanisms) which allows to achieve the 50msecs convergence time**
  - **Uses Protection (Meaning pre-signalled backup)**
  - **reoptimisation with Make before break to find a more optimal path**

MPLS TE FRR – Local protection

- **A key principle of Local repair is to guarantee a very fast traffic recovery with or without QOS guarantee (bandwidth guarantee) during a transient phase while other mechanisms (reoptimisation) are used over a longer time scale.**

MPLS TE FRR Local repair

- Controlled by the PLR
  - local repair is configured on a per link basis
  - the resilience attribute of a trunk allows to control whether local repair should be applied to the trunk (tunn mpls traff fast-reroute).

→"Local Protection Desired" bit of the SESSION_ATTRIBUTE object flag is set.

  - Just the reroutable LSPs will be backed-up (fine granularity)

- Uses nested LSPs (stack of labels)

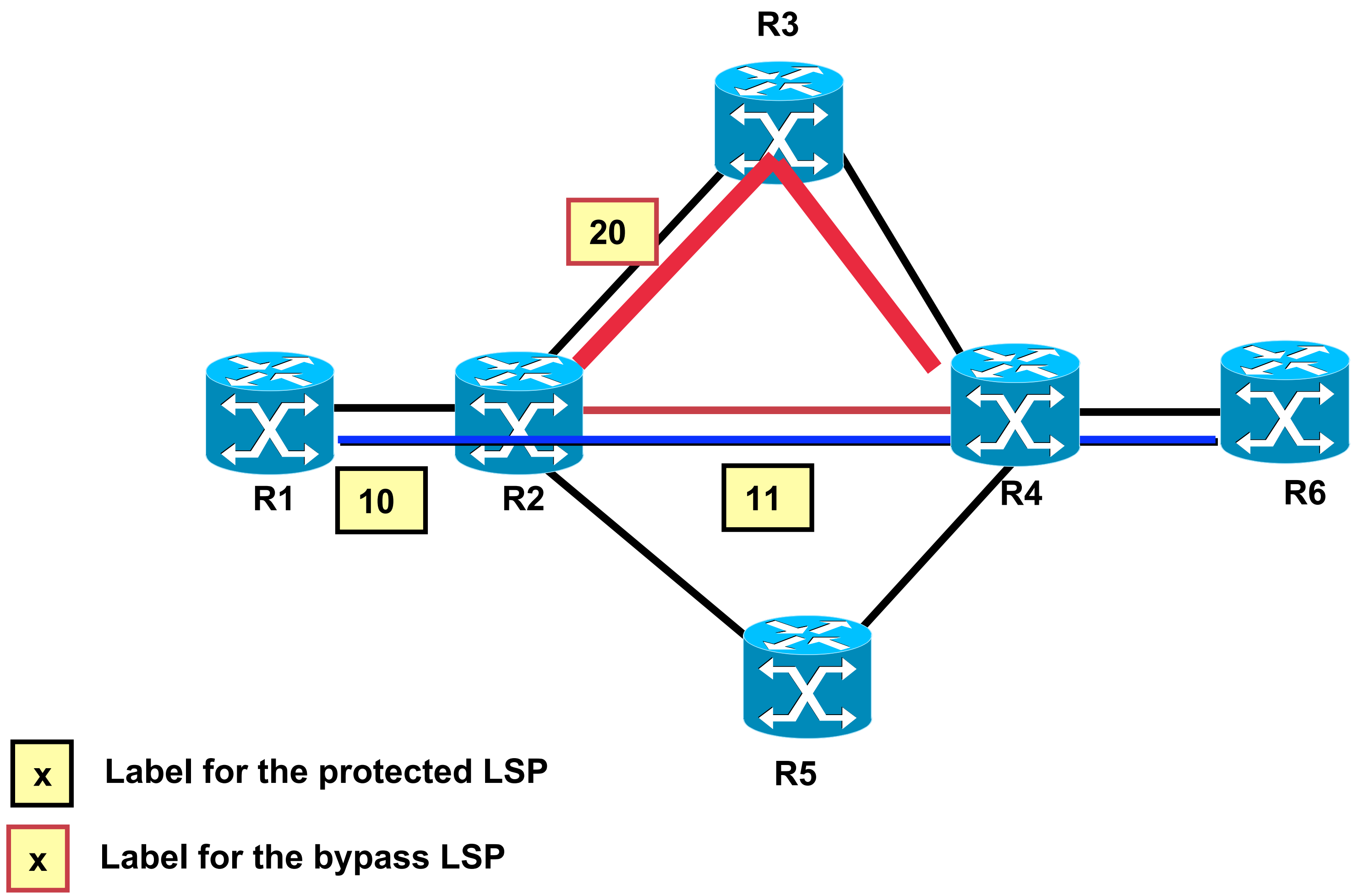
  1:N protection is KEY for scalability. N protected LSP will be backed-up onto the SAME backup LSP

Cisco.com

MPLS TE Fast Reroute
Link Protection
(local protection)

MPLS TE FRR – Link Protection

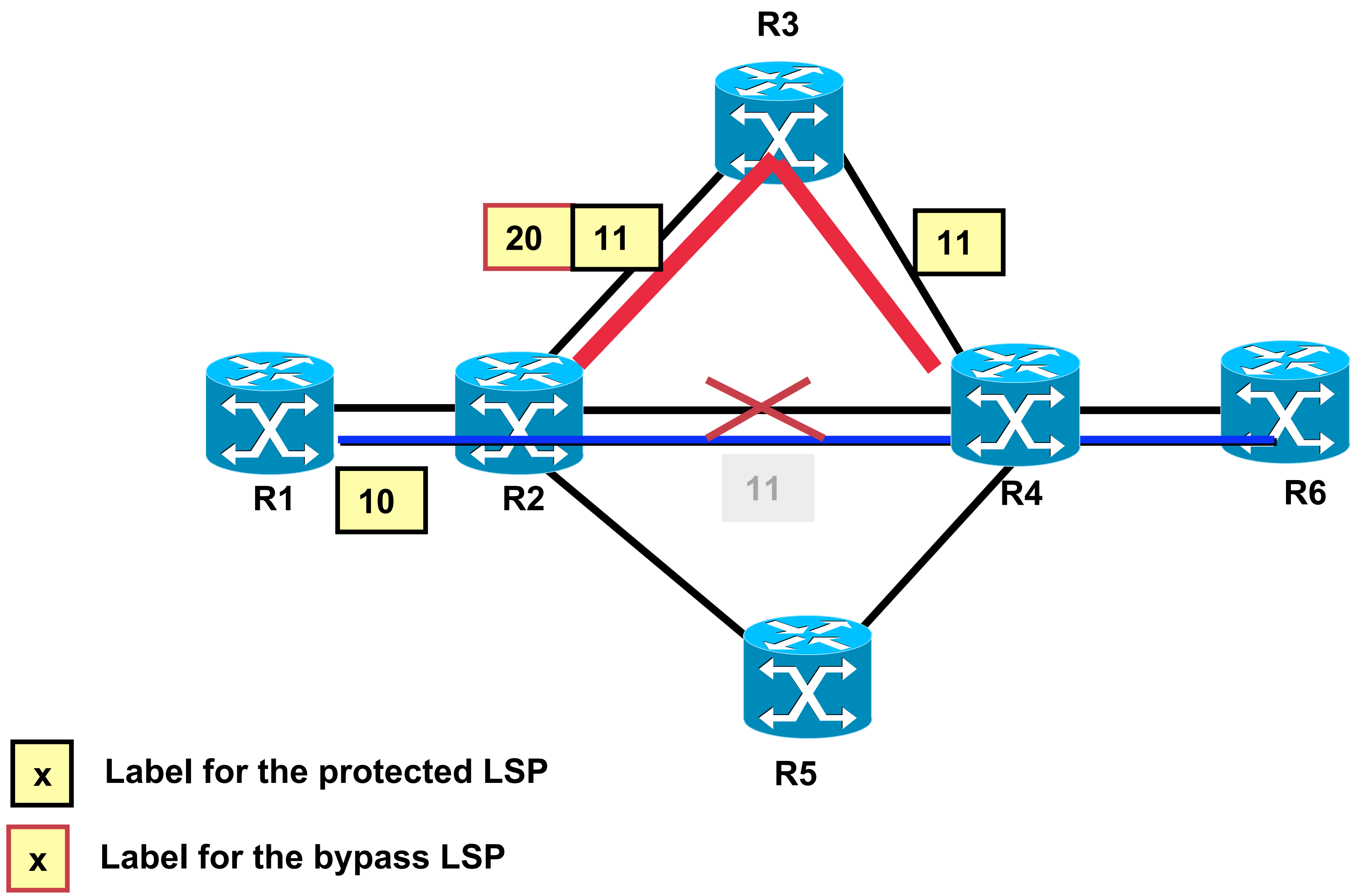
- **Backup labels (NHOP Backup Tunnel)**



MPLS TE FRR – Link Protection

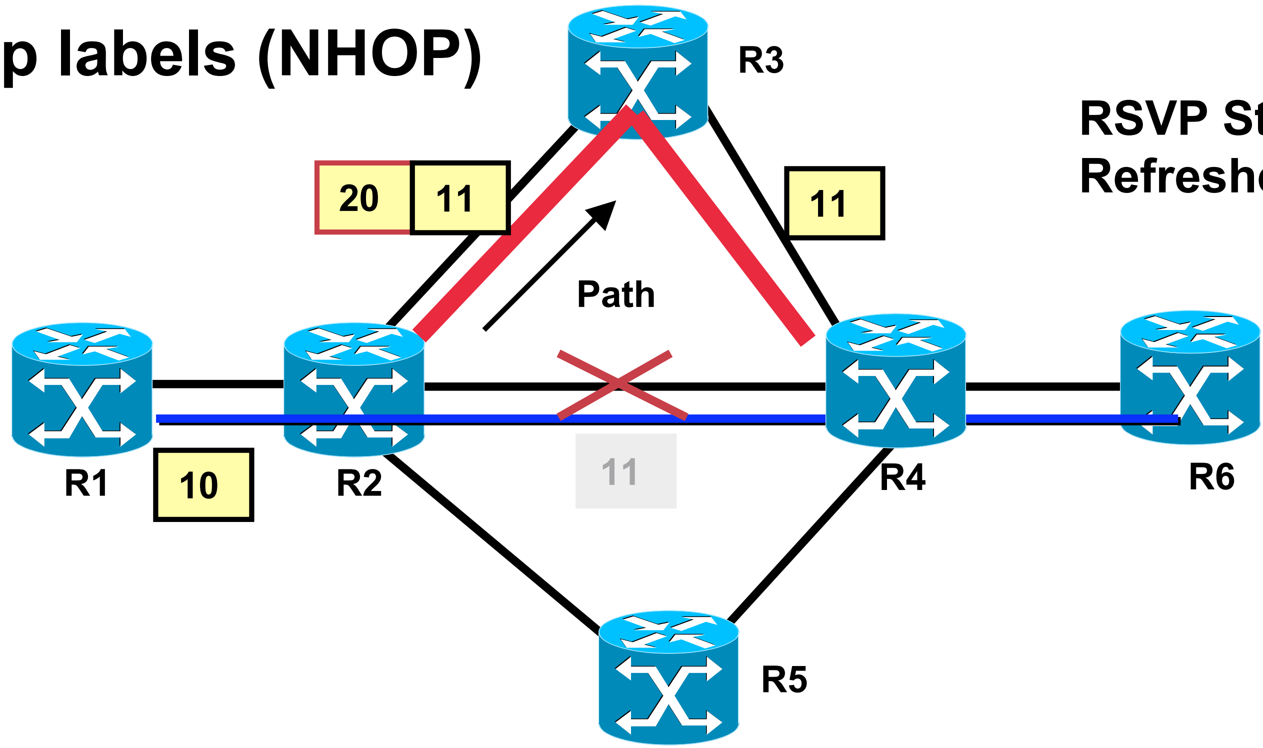
Cisco.com

- **Backup labels (NHOP Backup Tunnel)**



MPLS TE FRR – Link Protection

- **Backup labels (NHOP)**



2 remarks:

*** The path message for the old Path are still forwarded onto the Back-Up LSP**

*** Modifications have been made to the RSVP code so that**

- R2 could receive a Resv message from a different interface than the one used to send the Path message

- R4 could receive a Path message from a different interface (R3-R4 in this case)

MPLS TE FRR – Link Protection

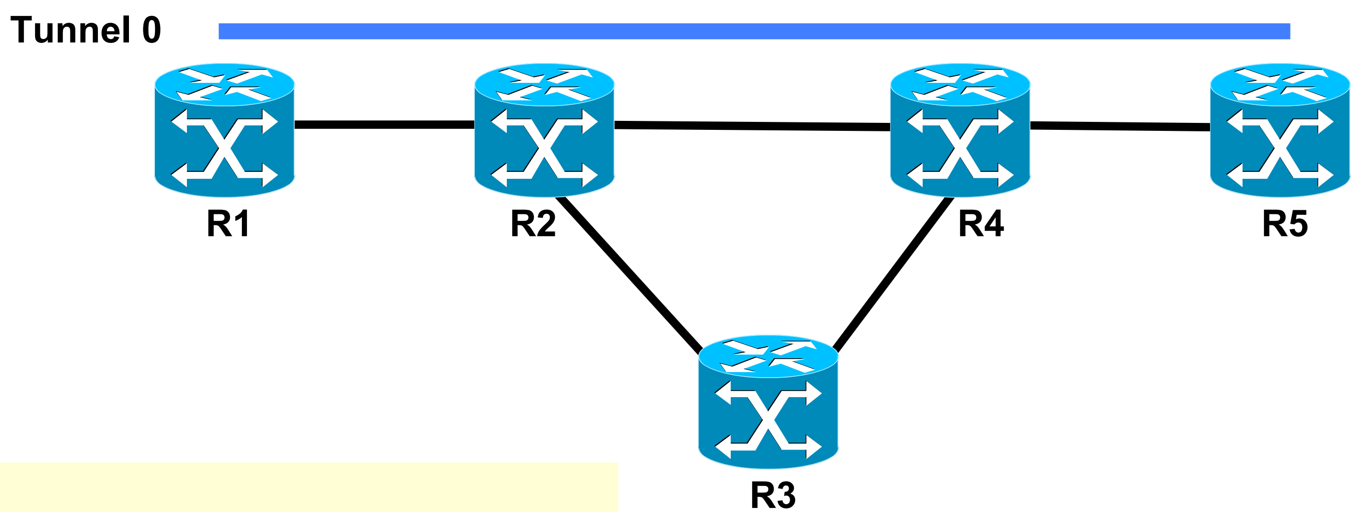
Cisco.com

- The PLR SHOULD send a PathErr message with error code of "Notify" (Error code =25) and an error value 3 ("Tunnel locally repaired").

→ This will trigger the head-end reoptimisation

- Then the TE LSP will be rerouted over an alternate Path (may be identical) using Make Before Break.

MPLS TE FRR - Link Protection - Configuration



- ***On R1***

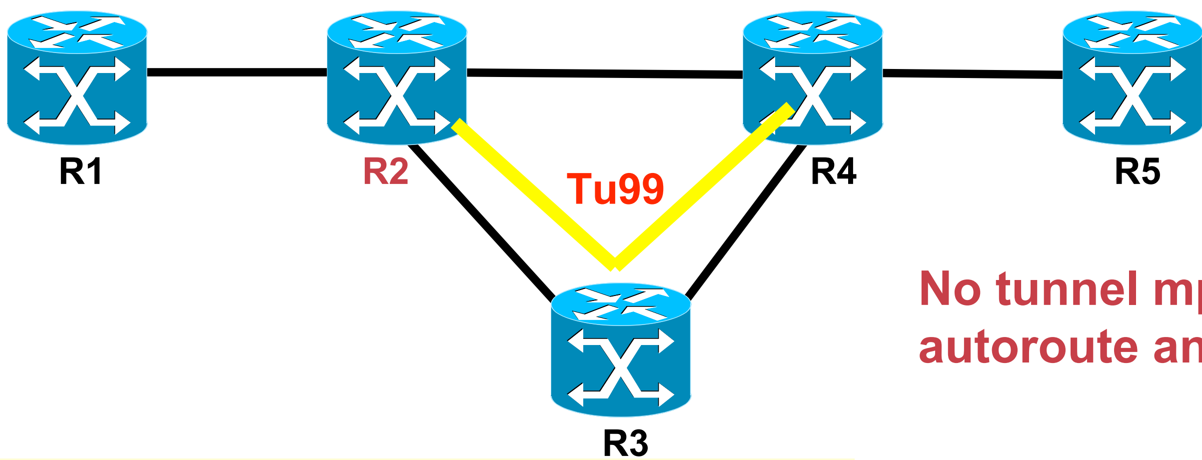
```
!
interface Tunnel0
 ip unnumbered Loopback0
 no ip directed-broadcast
 tunnel destination 10.0.1.102
 tunnel mode mpls traffic-eng
 tunnel mpls traffic-eng autoroute announce
 tunnel mpls traffic-eng priority 3 3
 tunnel mpls traffic-eng bandwidth 10000
 tunnel mpls traffic-eng path-option 1 dynamic
 tunnel mpls traffic-eng record-route
 tunnel mpls traffic-eng fast-reroute
```

Tunnel 0 is configured as fast reroutable

"Local Desired Protection" flag set in the SESSION_ATTRIBUTE object

MPLS TE FRR - Link Protection - Configuration

Cisco.com



No tunnel mpls traffic-eng autoroute announce !

A Back-Up Tunnel Tu99 explicitly routed is configured on R2

```
interface Tunnel99
 ip unnumbered Loopback0
 no ip directed-broadcast
 tunnel destination 10.0.1.100
 tunnel mode mpls traffic-eng
 tunnel mpls traffic-eng priority 1 1
 tunnel mpls traffic-eng bandwidth 10000
 tunnel mpls traffic-eng path-option 1 explicit name secours
 tunnel mpls traffic-eng record-route
```

Use also:

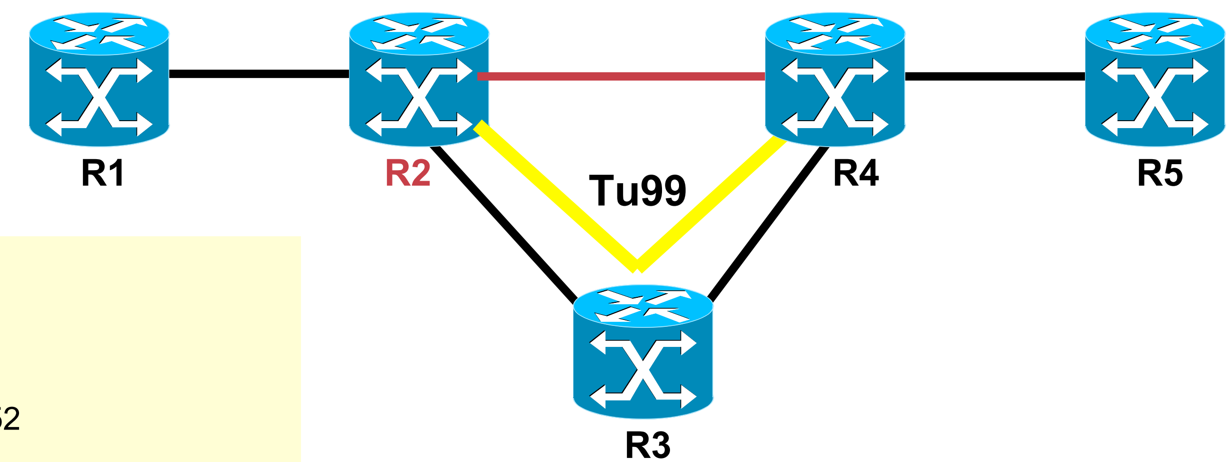
Router (cfg-ip-expl-path)# exclude-address a.b.c.d

Where a.b.c.d is a link address or a router ID to exclude a node

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

MPLS TE FRR - Link Protection - Configuration

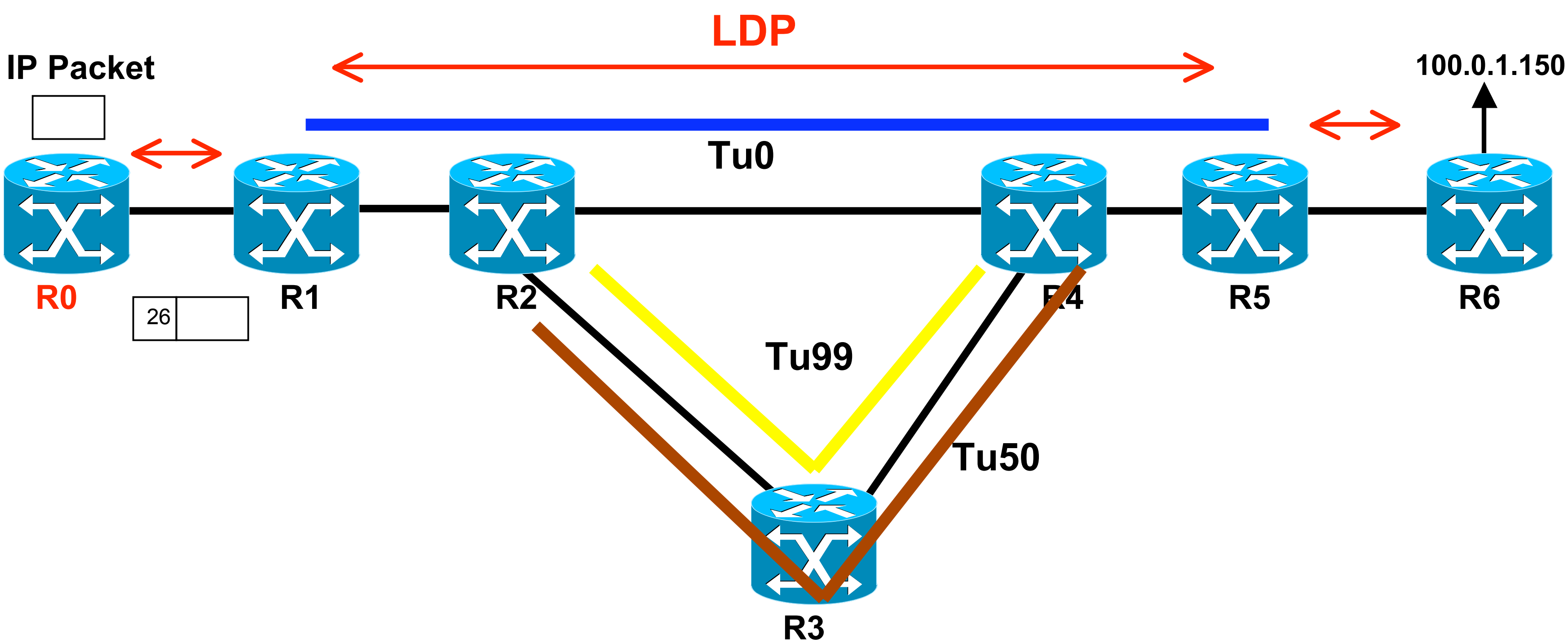
Cisco.com



On R2

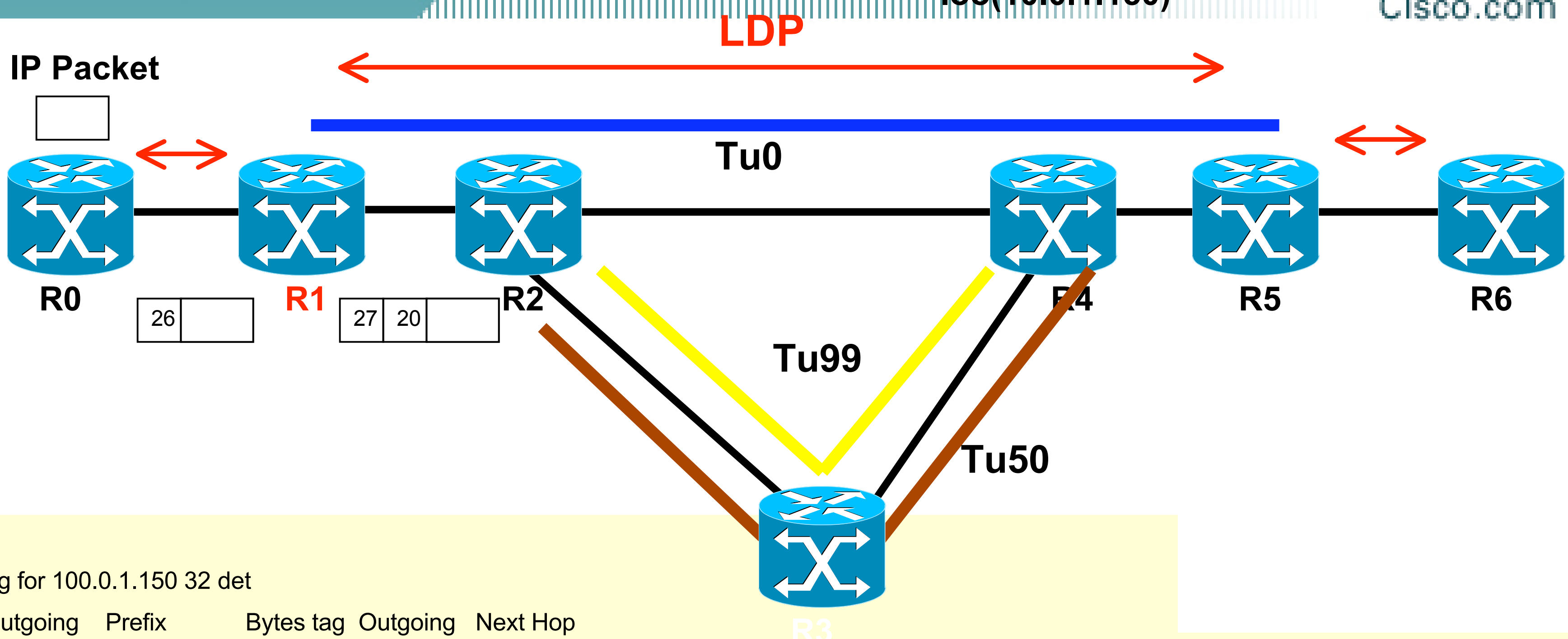
```
interface POS4/0
 description Link to R4
 ip address 10.1.13.2 255.255.255.252
 no ip directed-broadcast
 ip router isis
 encapsulation ppp
 mpls traffic-eng tunnels
 mpls traffic-eng backup-path Tunnel99
 tag-switching ip
 no peer neighbor-route
 crc 32
 clock source internal
 pos ais-shut
 pos report lrdi
 ip rsvp bandwidth 155000 155000
```

MPLS TE FRR - Link Protection



MPLS TE FRR - Link Protection

Traffic is running from R0's loo to R6's loo(10.0.1.150)



On R1

Show tag for 100.0.1.150 32 det

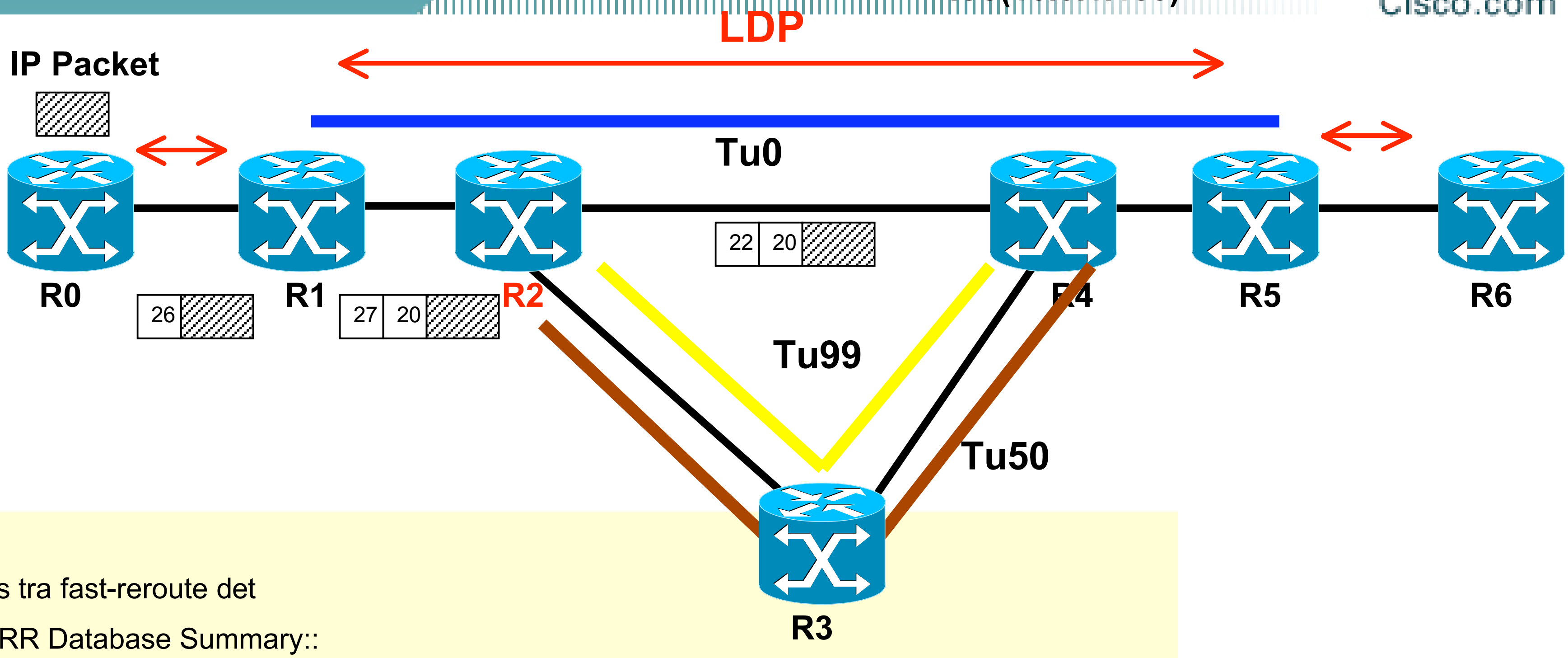
```
Local  Outgoing    Prefix            Bytes tag  Outgoing   Next Hop
tag    tag or VC   or Tunnel Id      switched   interface
26     20          10.0.1.150/32     0          Tu0        point2point
       MAC/Encaps=4/12, MTU=4466, Tag Stack{27 20}, via PO0/0
       0F008847 0001B00000014000
       Fast Reroute Protection via {UnknownIF, outgoing label 27}
   Per-packet load-sharing, slots: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
```

```
sh tag for
Local  Outgoing    Prefix            Bytes tag  Outgoing   Next hop
tag    tag or VC   or Tunnel Id      switched   interface
...
26     20          [T] 10.0.1.150/32  0         Tu0        point2point
...

[T]    Forwarding through a TSP tunnel.
       View additional tagging info with the 'detail' option
```

MPLS TE FRR - Link Protection

Traffic is running from R0's loo to R6's loo(10.0.1.150)



On R2

sh mpls tra fast-reroute det

LFIB FRR Database Summary::

Total Clusters: 1

Total Groups: 1

Total Items: 1

Link 10:: PO4/0 (Up, 1 group)

Group 16:: PO4/0->Tu99 (Up, 1 member)

Transit Item 810D60 (complete) [FRR OutLabel: 22]

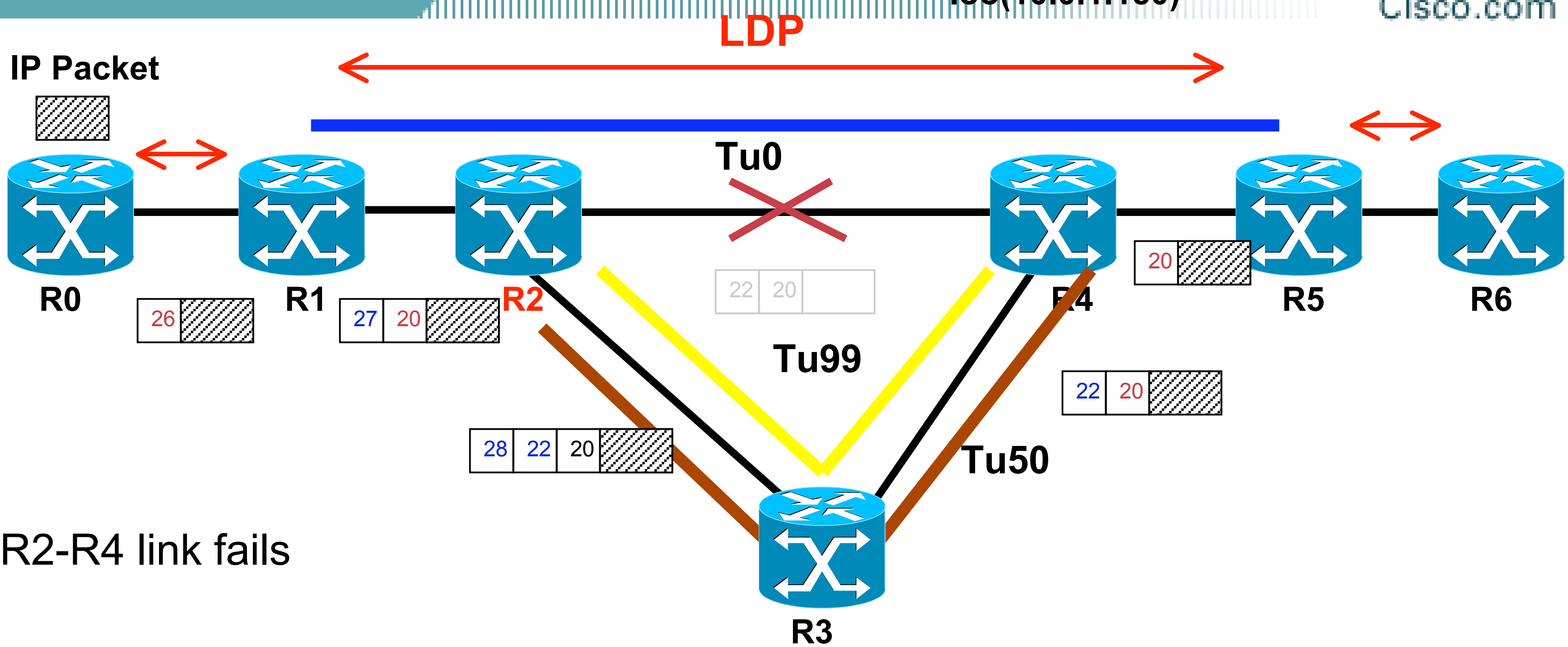
Key {incoming label 27}

sh tag for

Local tag	Outgoing tag or VC	Prefix or Tunnel Id	Bytes tag switched	Outgoing interface	NextHop
...					
27	22	10.0.1.127 0 [1]	16896	PO4/0	point2point
...					

MPLS TE FRR - Link Protection

Traffic is running from R0's loo to R6's loo(10.0.1.150)



t0: R2-R4 link fails

t1:

Data plane: R2 will immediately swap 27 <-> 22 (as before) and Push 28 (This is of course done for all the protected LSPs crossing the R2-R4 link)

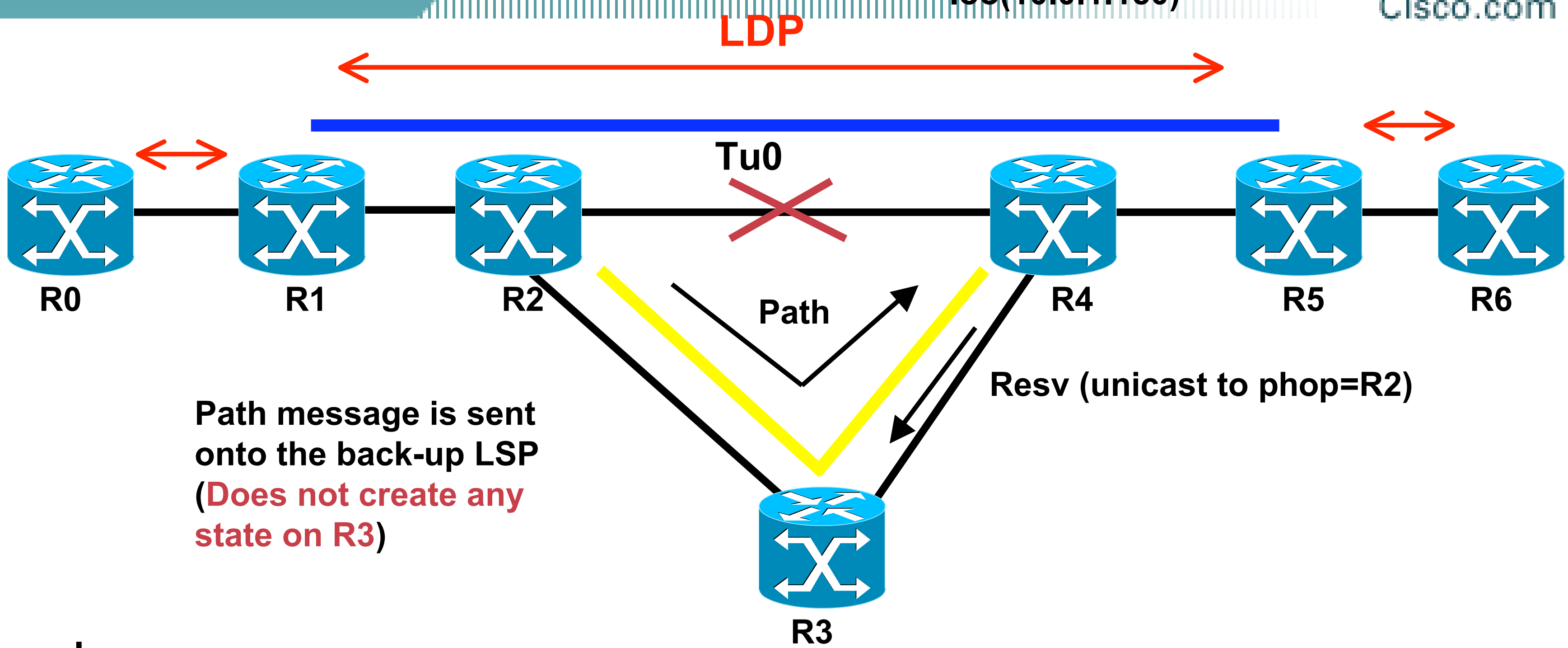
Control Plane registers for a link-down event. Once the RSVP process receives this event, it will send out an RSVP PATH ERR msg (O(s))

t2: R3 will do PHP

t3: R4 receives an identical labeled packet as before (Global Label Allocation needed)

MPLS TE FRR - Link Protection

Traffic is running from R0's loo to R6's loo(10.0.1.150)



2 remarks:

* The path message for the old Path are still forwarded onto the Back-Up LSP

* Modifications have been made to the RSVP code so that

  - R2 could receive a Resv message from a different interface than the one used to send the Path message

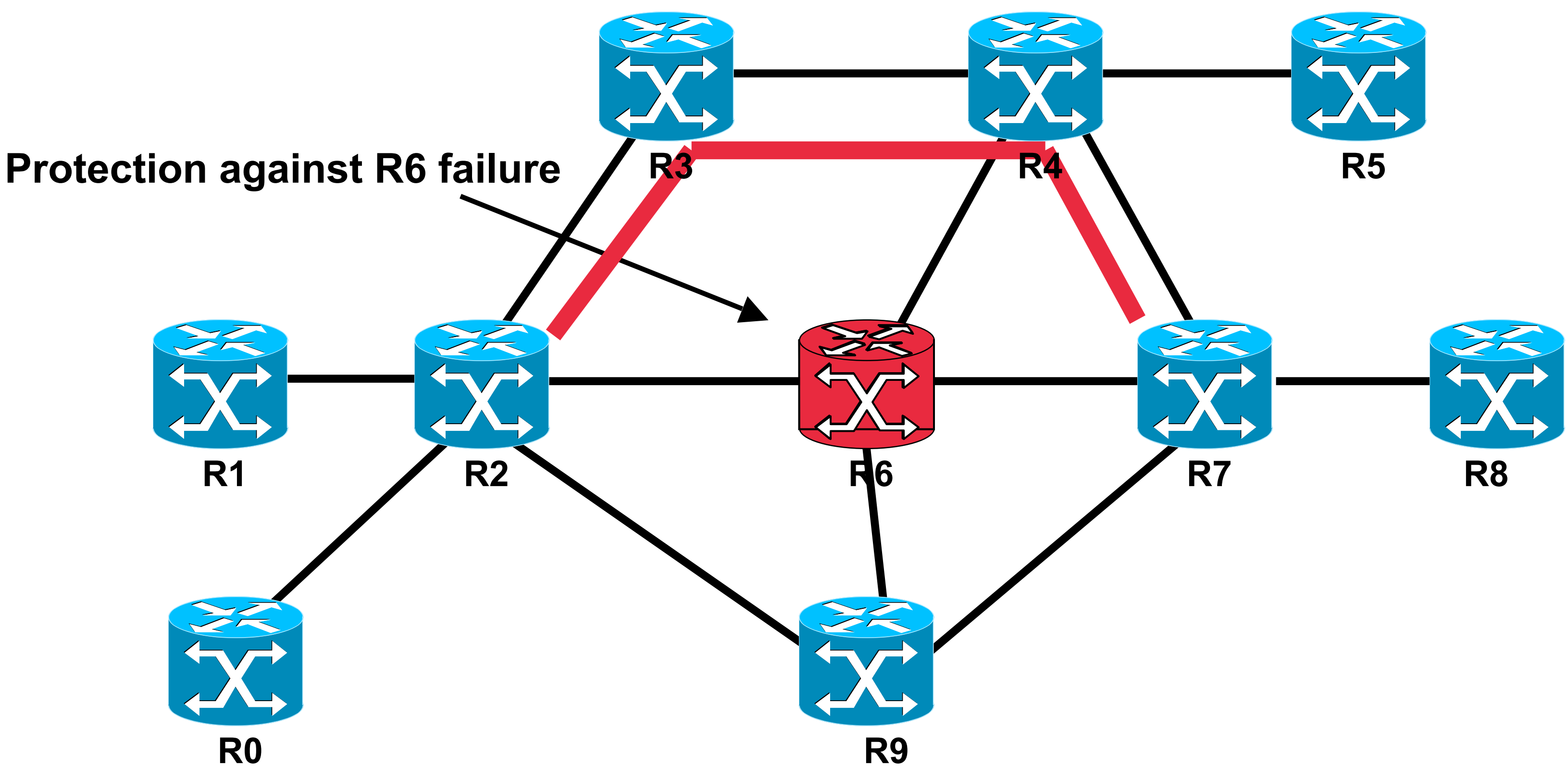
  - R4 could receive a Path message from a different interface (R3-R4 in this case)

Cisco.com

MPLS TE Fast Reroute Node Protection (local protection)

MPLS TE FRR – Node Protection

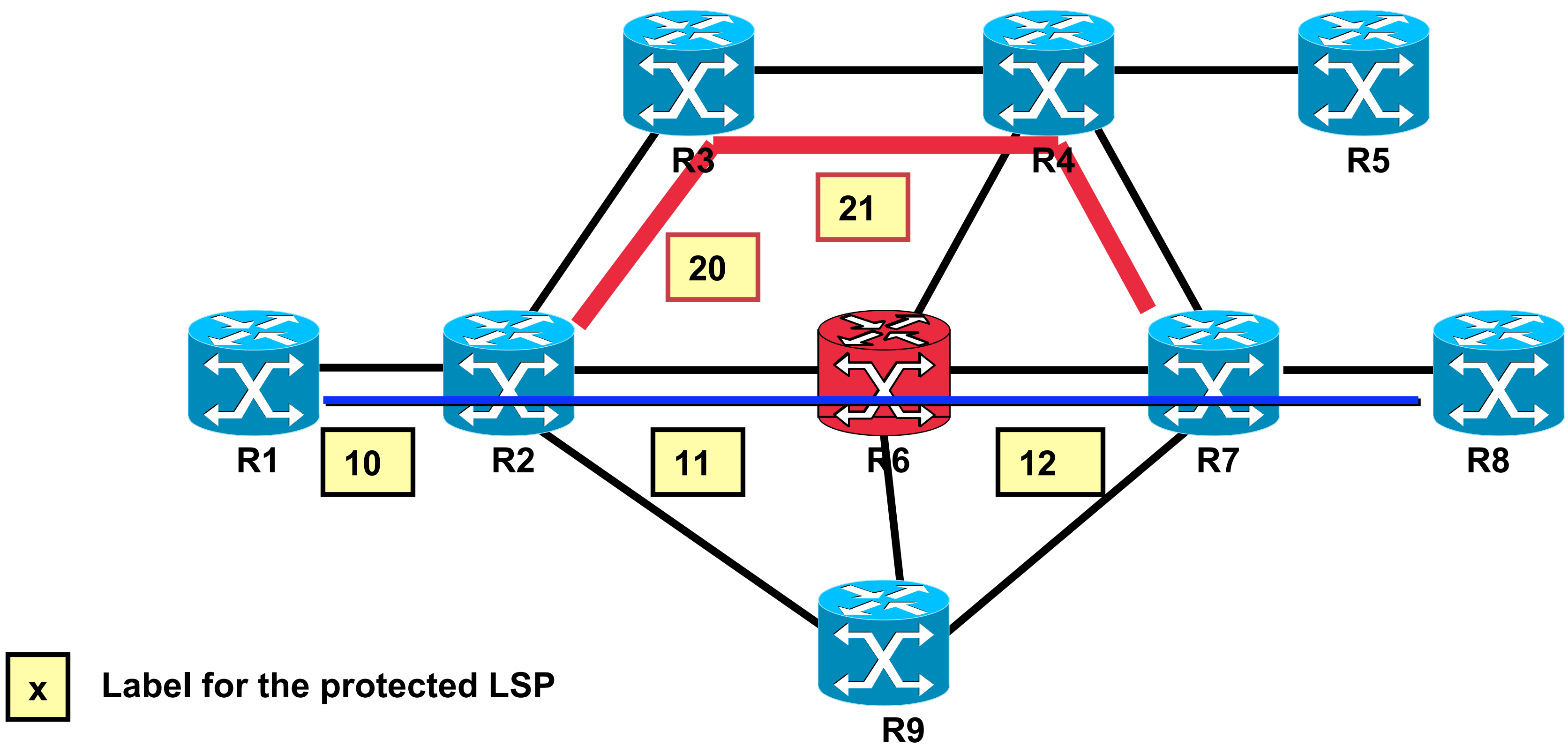
- **Node protection allows to configure a back-up tunnel to the next-next-hop ! This allows to protect against link AND node failure**



MPLS TE FRR – Node Protection

Cisco.com

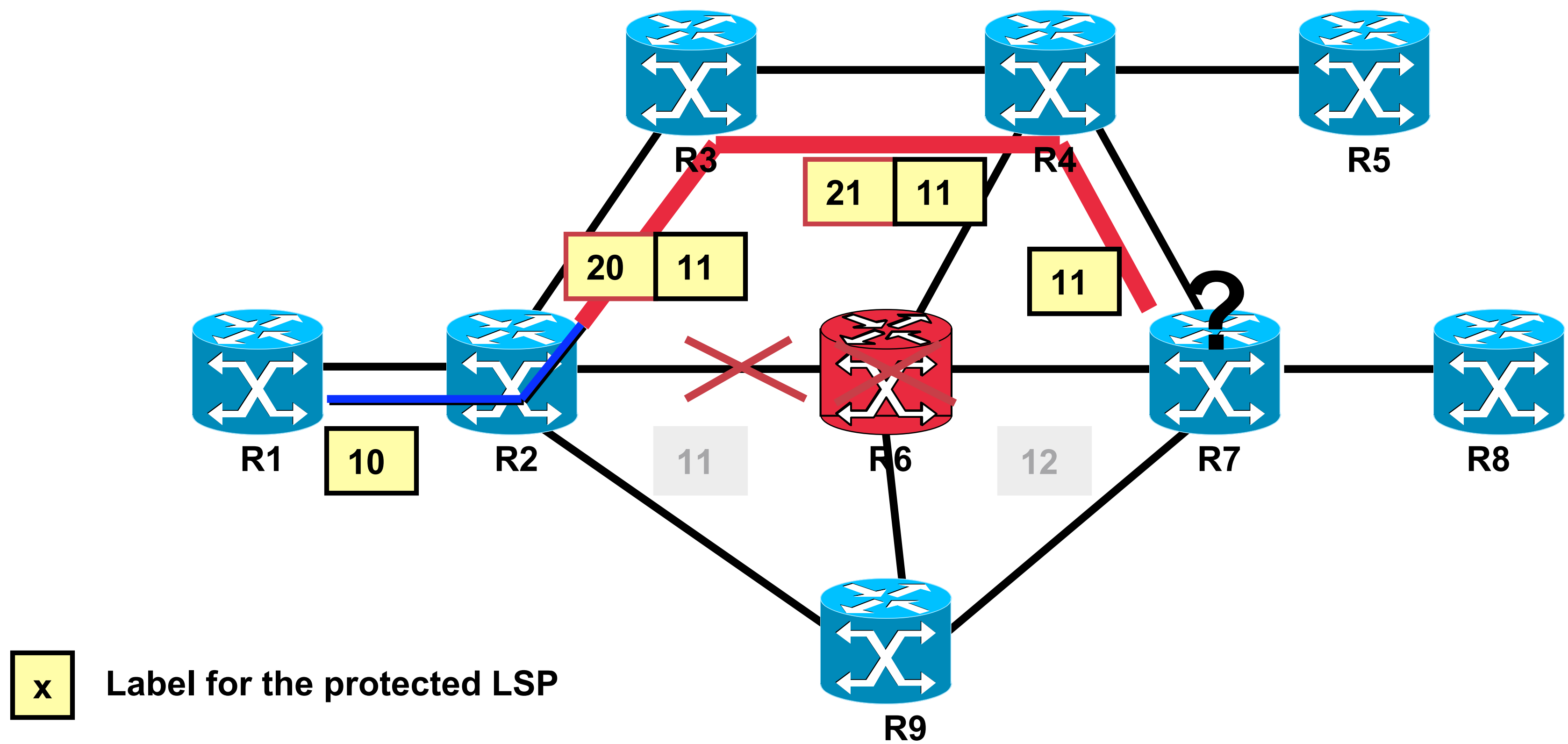
- **Backup labels**



MPLS TE FRR – Node Protection

Cisco.com

- **Backup labels**

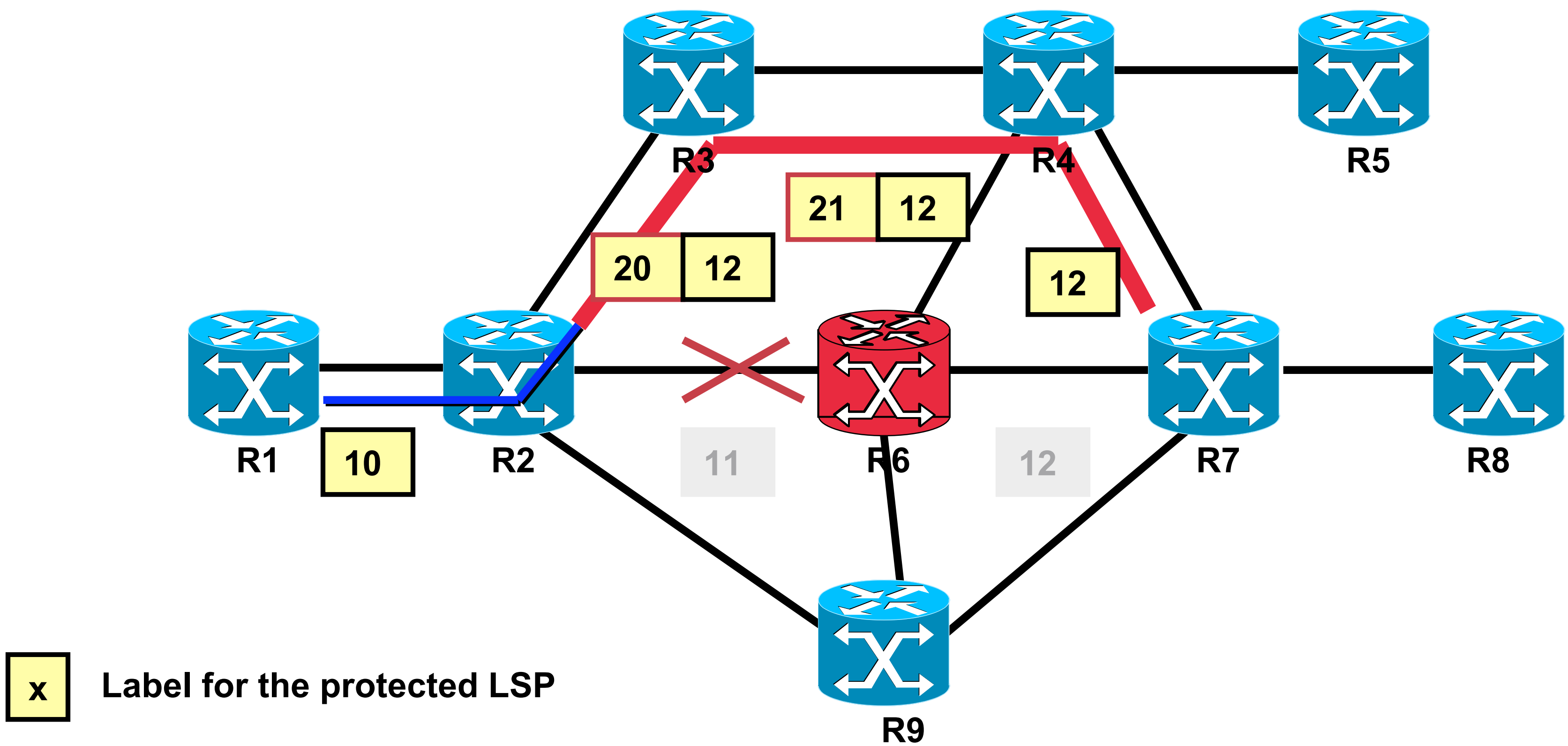


- **The PLR learns the label to use from the RRO object carried in the Resv message when the reroutable LSP is first established – With global label space allocation on the MP**

MPLS TE FRR – Node Protection

Cisco.com

- **Backup labels**



- **The PLR swaps 10 <-> 12, pushes 20 and forward the traffic onto the backup tunnel**

MPLS TE FRR

Cisco.com

- **Path state maintenance**
  - **As in the case of NHOP backup tunnel, the Path messages are sent onto the backup tunnel to refresh the downstream states**

MPLS TE Fast Reroute

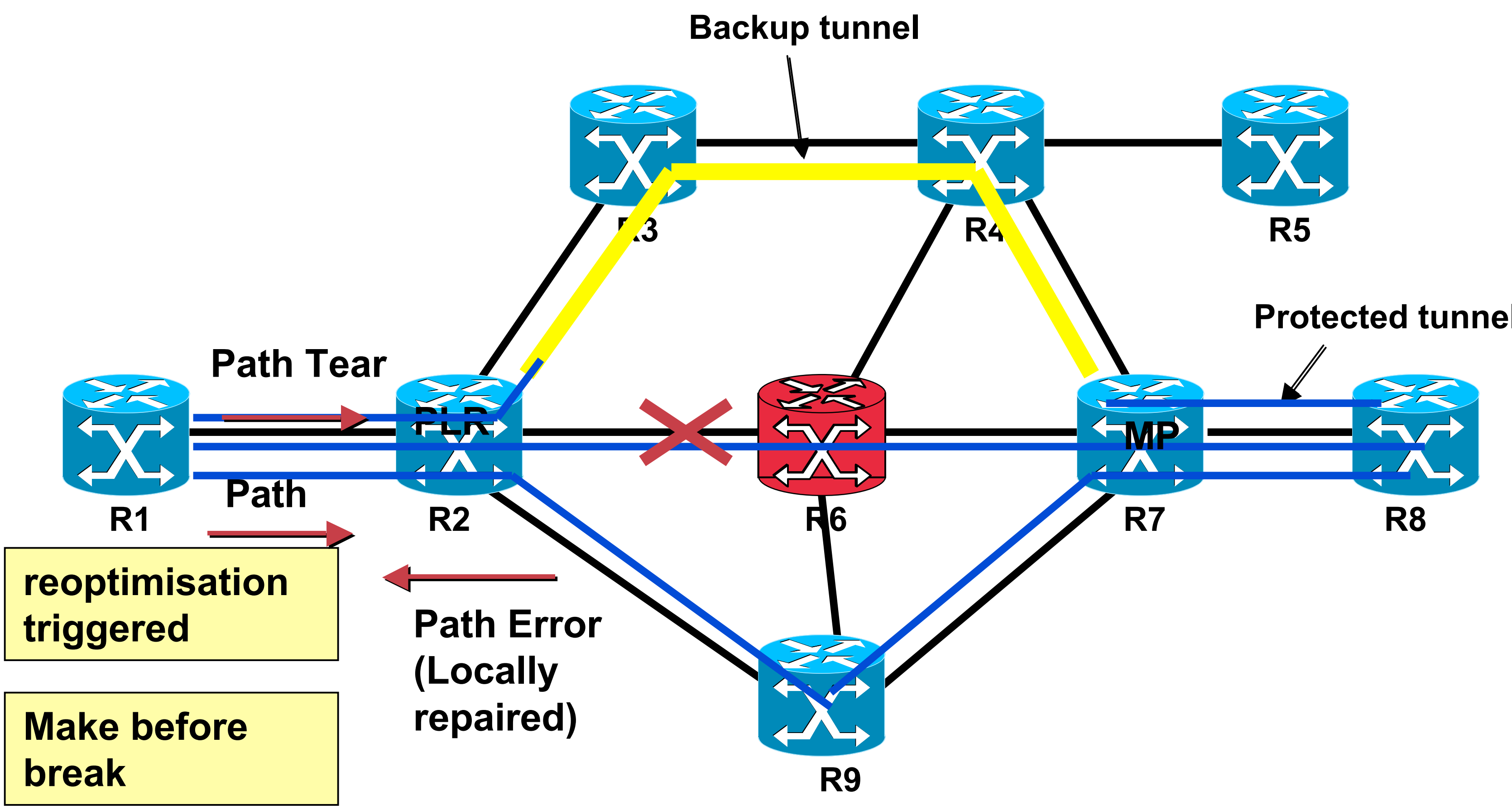
Cisco.com

- When the failure occurs, the PLR also updates:
  - The ERO object,
  - The PHOP object,
  - The RRO object
- The Point of Local Repair SHOULD send a PathErr message with error code of "Notify" (Error code =25) and an error value field of ss00 cccc cccc cccc where ss=00 and the sub-code = 3 ("Tunnel locally repaired").

→ This will trigger the head-end reoptimization

MPLS TE FRR

Cisco.com



MPLS TE FRR – Node Protection

- **The number of back-up tunnels for an interface is no longer limited to one !**

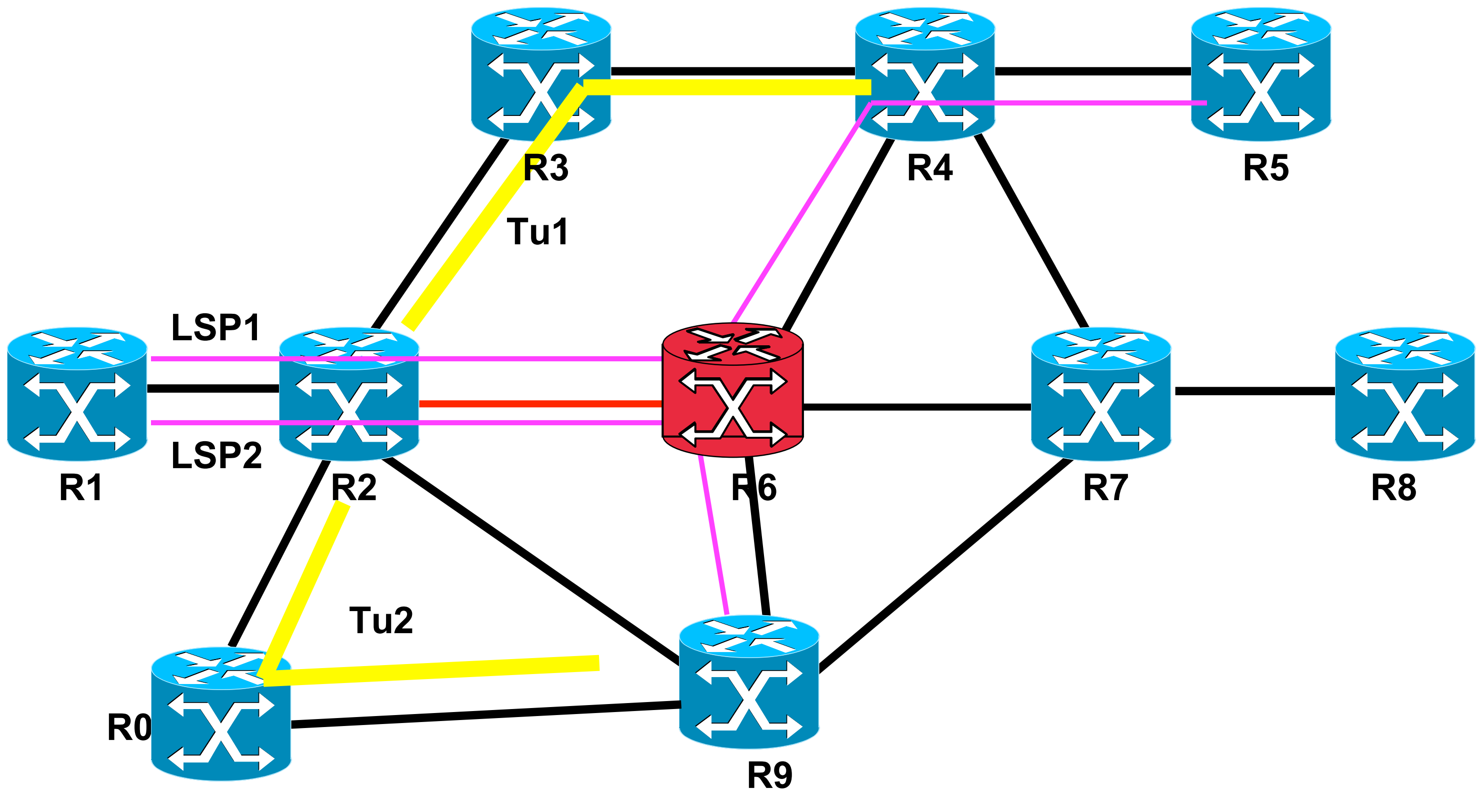
On R2

```
interface POS4/0
 description Link to R4
 ip address 10.1.13.2 255.255.255.252
 no ip directed-broadcast
 ip router isis
 encapsulation ppp
 mpls traffic-eng tunnels
 mpls traffic-eng backup-path Tunnel10
 mpls traffic-eng backup path Tunnel15
 tag-switching ip
 no peer neighbor-route
 crc 32
 clock source internal
 pos ais-shut
 pos report lrdi
 ip rsvp bandwidth 155000 155000
```

- ***Which is mandatory for Node protection …***

MPLS TE FRR – Node Protection

- **Back-up tunnel selection for a given LSP**



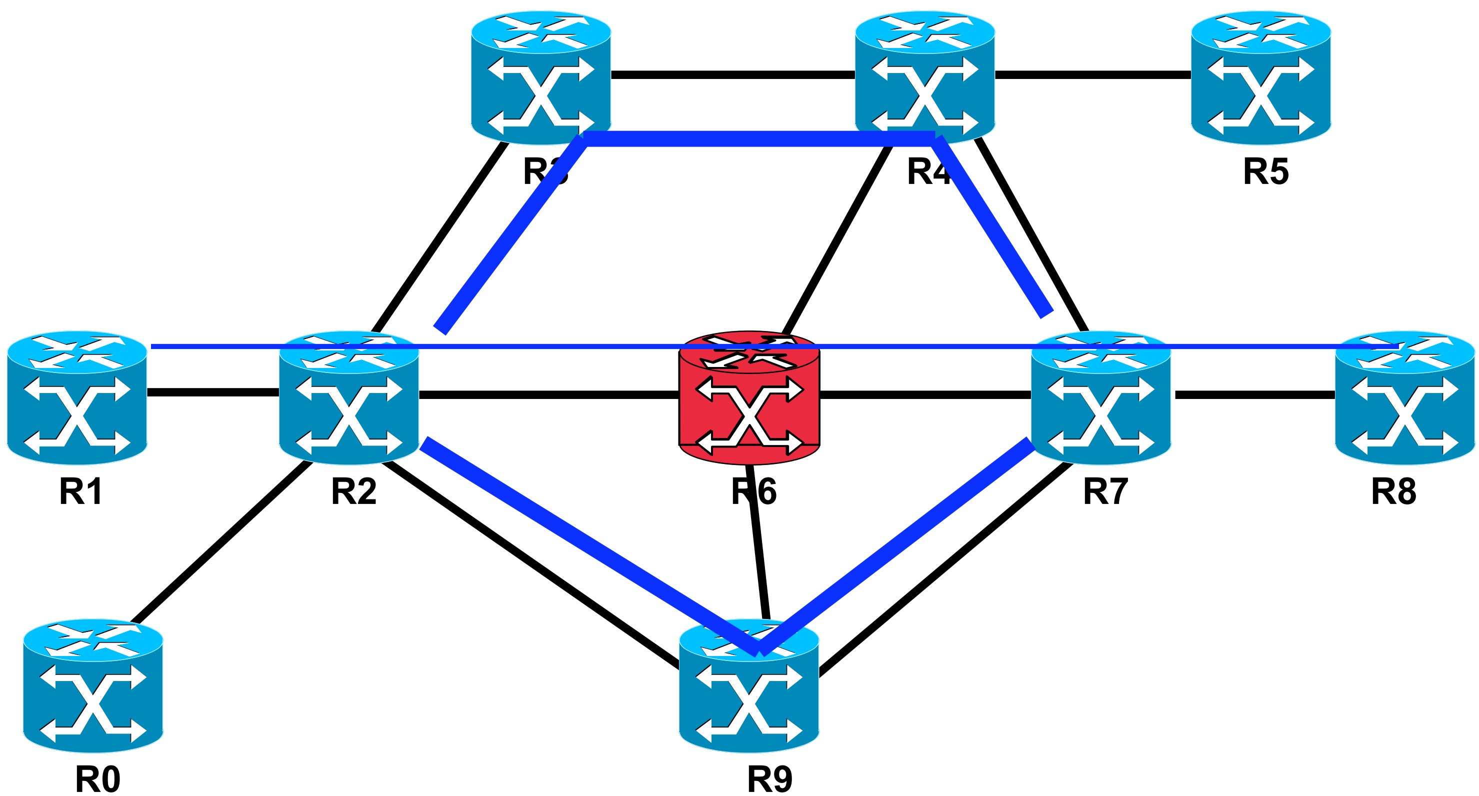
- **Tu1 is chosen for LSP1**
- **Tu2 is chosen for LSP2**

MPLS TE FRR – Node Protection

- One may combine tunnels terminating on the next hop and next-next-hop
- This allows to increase redundancy
- In case of unavailability of a back-up tunnel the other one is used (order of preference is determined by the tunnel ID number)
- Load balancing of LSPs between back-up tunnels terminating on the same NNHOP.

MPLS TE FRR – Node Protection

- **Load balancing: Multiple back-up tunnels to the same destination may be created.**

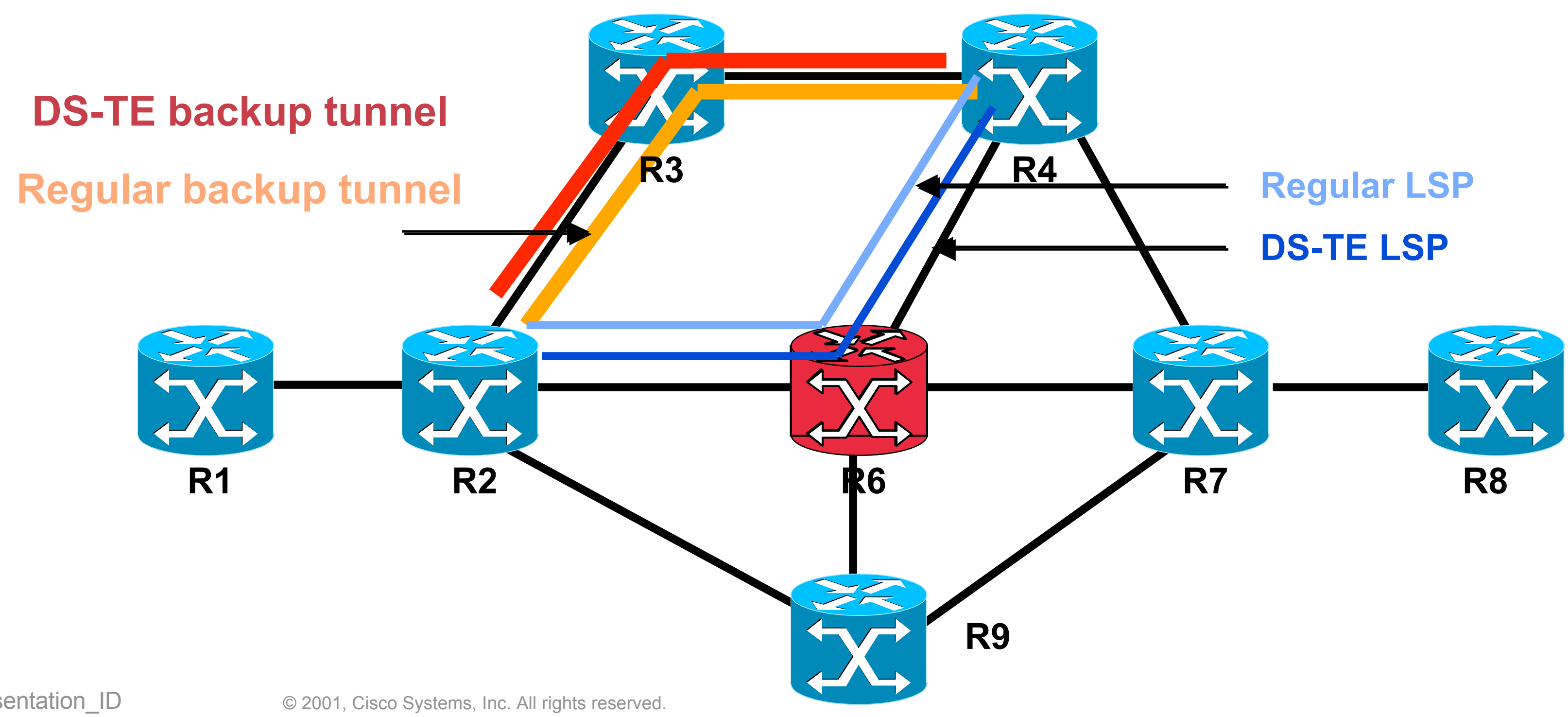


Backup tunnel path computation and provisioning

- Packing algorithm: refers to the method used to select the backup tunnel for each protected LSP.
- For each protected LSP at a given PLR:
  - Select the set of backup tunnel whose merge point crosses the primary path,
  - Find a backup tunnel whose remaining bandwidth is >= of the protected LSP (if bandwidth protection is required)
  - Multiple backup tunnel selection policies are available

Per Class backup tunnel

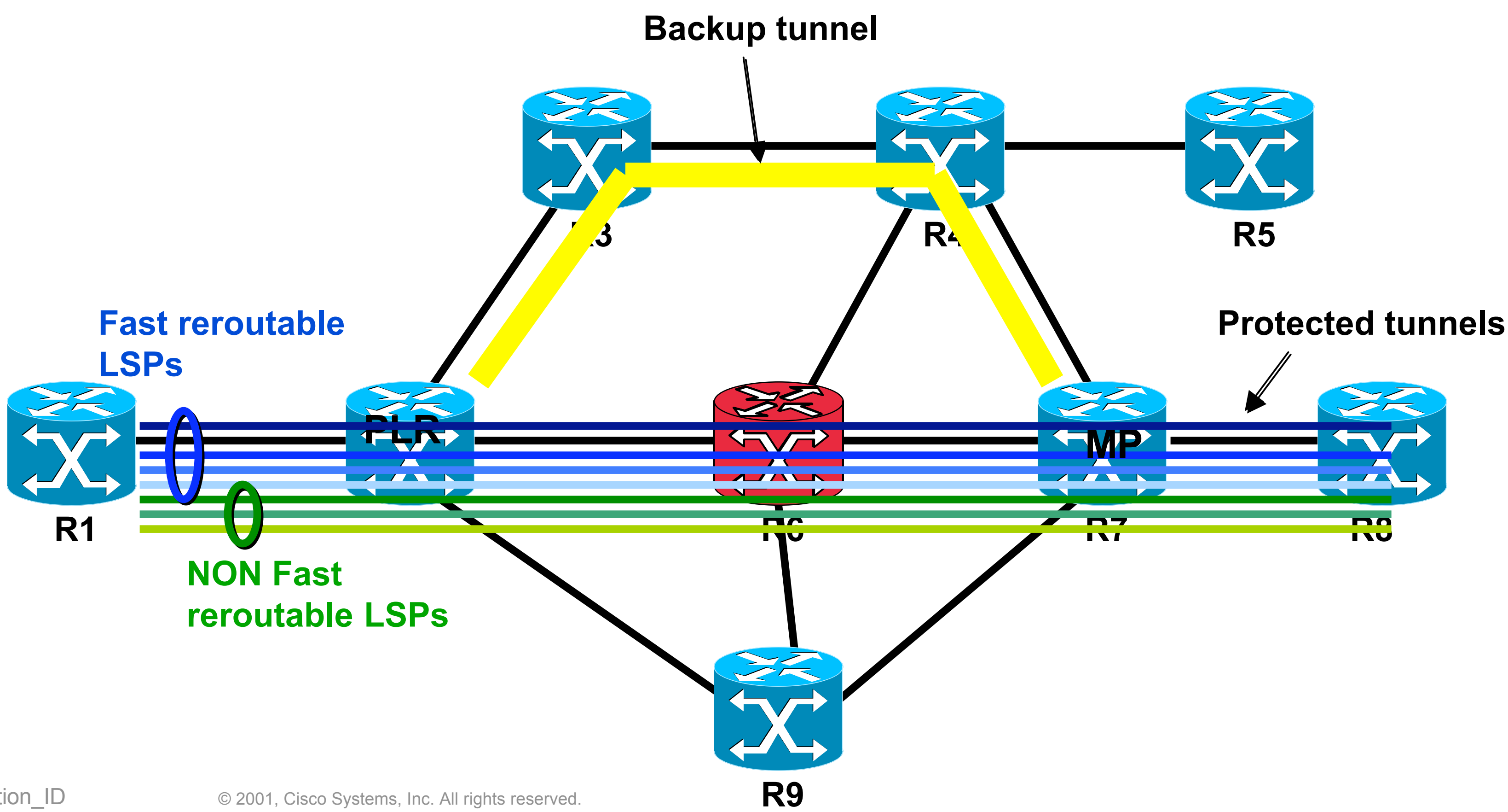
- When using both regular and DS-TE tunnels, it may desirable to configure regular and DS-TE backup tunnels.
- Other combinations are also possible
- Packing algorithm enhancements



MPLS TE FRR Local repair

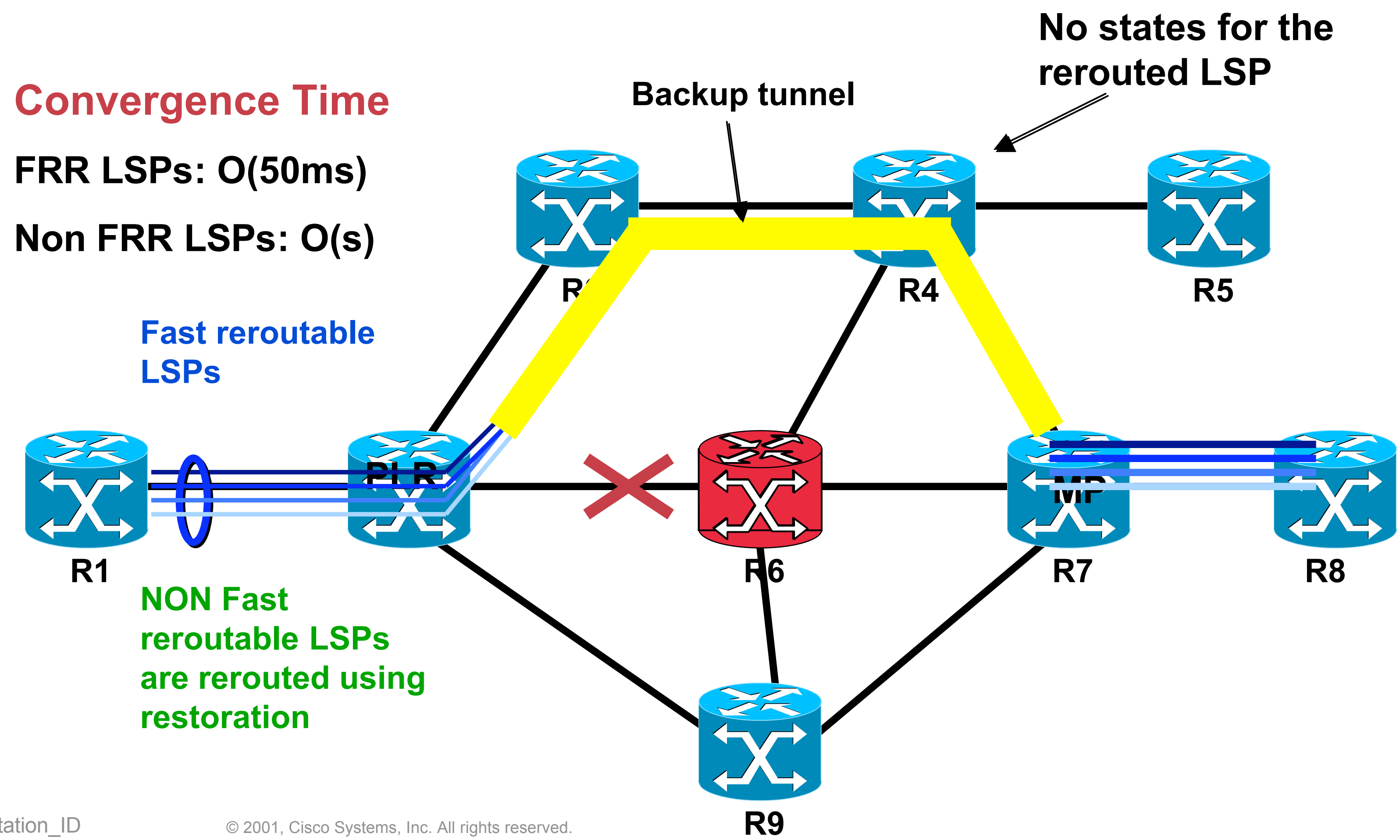
- **Uses nested LSPs (stack of labels)**

    1:N protection is KEY for scalability. N protected LSP will be backed-up onto the SAME backup LSP



MPLS TE FRR Local repair

- **Uses nested LSPs (stack of labels)**



MPLS TE protection/restoration schemes

Cisco.com

Link/Node Failure detection

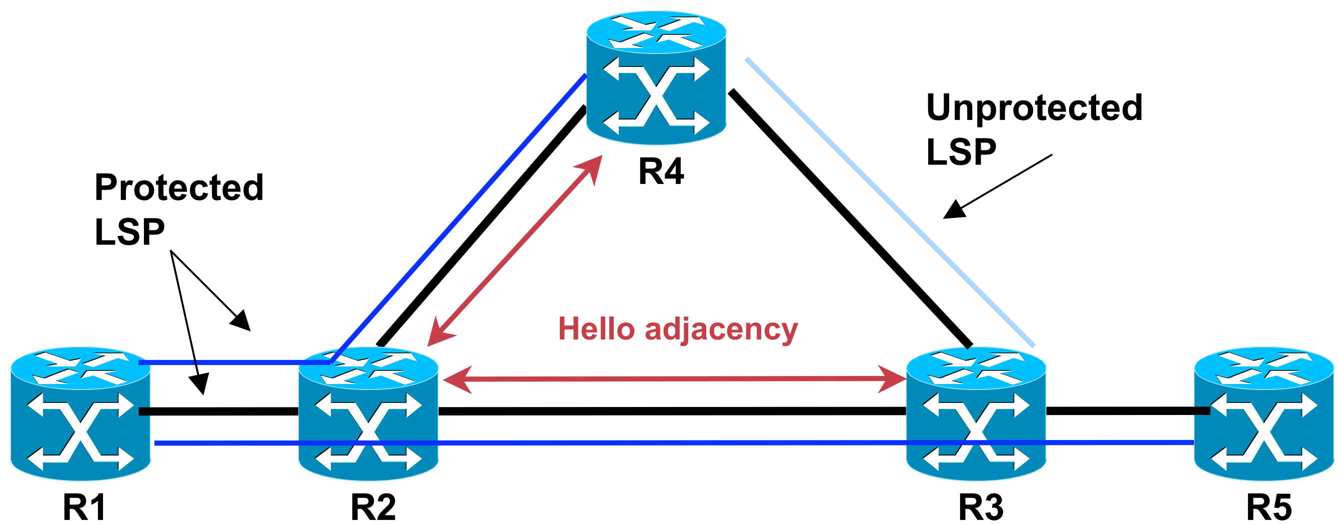
- Link failure detection
  - On POS, link failure detection is handled by Sonet/SDH alarms
    - On Receive side: LOS/LOF/LAIS
    - On Transmit side: LRDI
    - Very fast.
- Node failure detection is a more difficult problem
  - Node hardware failure => Link failure
  - Software failure … Need for a fast keepalive scheme (IGP, RSVP hellos)

RSVP Hellos

- RSVP Hellos extension is defined in RFC3209
- The RSVP hello extension enables an LSR to detect node failure detection
- Allows to detect:
  - Link failure when layer 2 does not provide failure detection mechanism,
  - Node failure when the layer 2 does not fail.

RSVP Hellos

- RSVP hello adjacency are brought up dynamically (if at least one protected LSP in READY state (with one backup tunnel operational))
- One RSVP hello adjacency per link per neighbor (not per protected LSP !!)

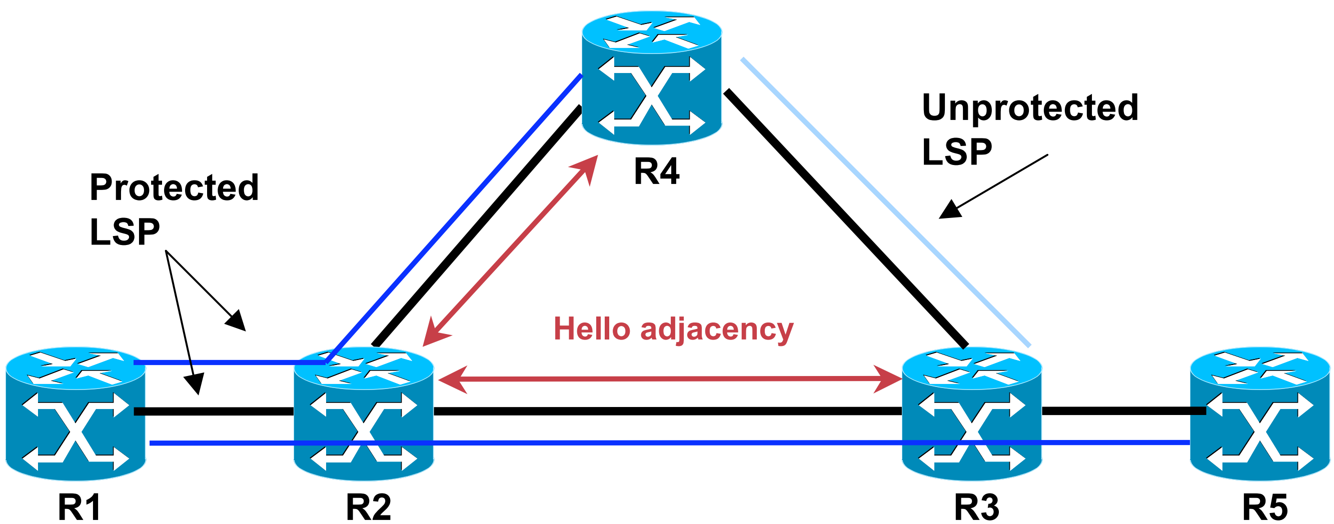


- An hello adjacency is removed when the last protected LSP in READY state is torn down

RSVP Hellos

Cisco.com

- **RSVP hello has been designed for Node failure detection. Fast link failure detection already exist on Sonet/SDH links.**



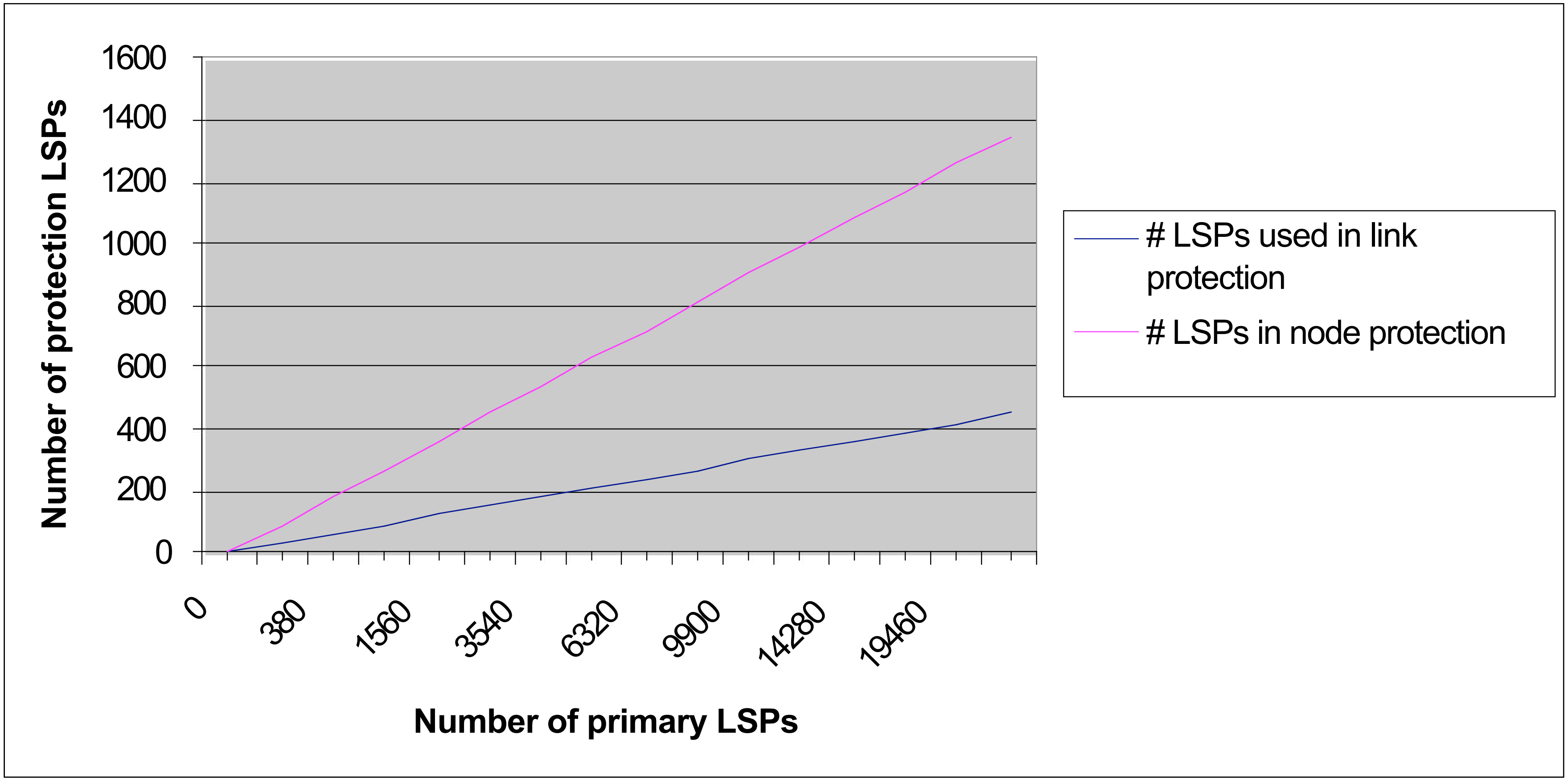
- **But can also be used as a fast link failure detection on GE links (point to point or behind a switch) ➔ FRR over GE links**

MPLS TE protection/restoration schemes

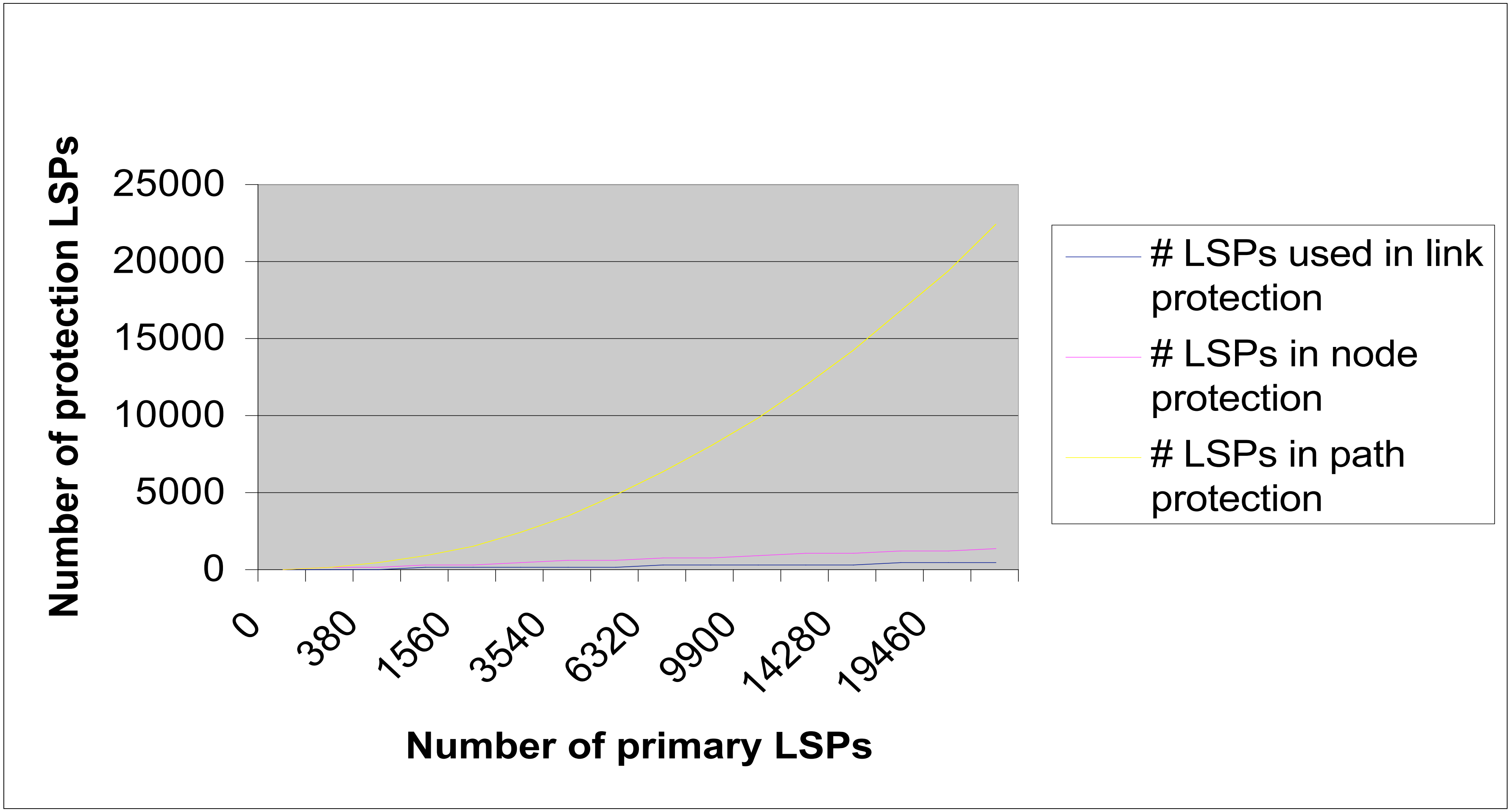
Cisco.com

- **Number of back-up LSPs required (impact on the number of states)**
  - **Primary Tunnels: $O(N^2)$**
  - **FRR Link protection: $O(N \times D)$**
  - **FRR Node protection: $O(N D^2)$**
  - **Path Protection $O(N^2)$**

Link/Node Scalability



Local vs. Path Protection Scalability



Bandwidth Protection

Introduction

Cisco.com

- IETF drafts
  - Local repair technique for fast recovery: draft-ietf-mpls-rsvp-lsp-fastreroute-00.txt FRR
  - Bandwidth protection, and other protection schemes (ie: SONET, Optical 1+1) but with a much more efficient backup bandwidth usage: draft-vasseur-mpls-backup-computation-00.txt
- Bandwidth Protection is required
  - For some, not all, types of traffic
  - In some, not all, networks

Abstract

Cisco.com

- Facility based computation model

    Proposed in draft-vasseur-mpls-backup-computation-00.txt

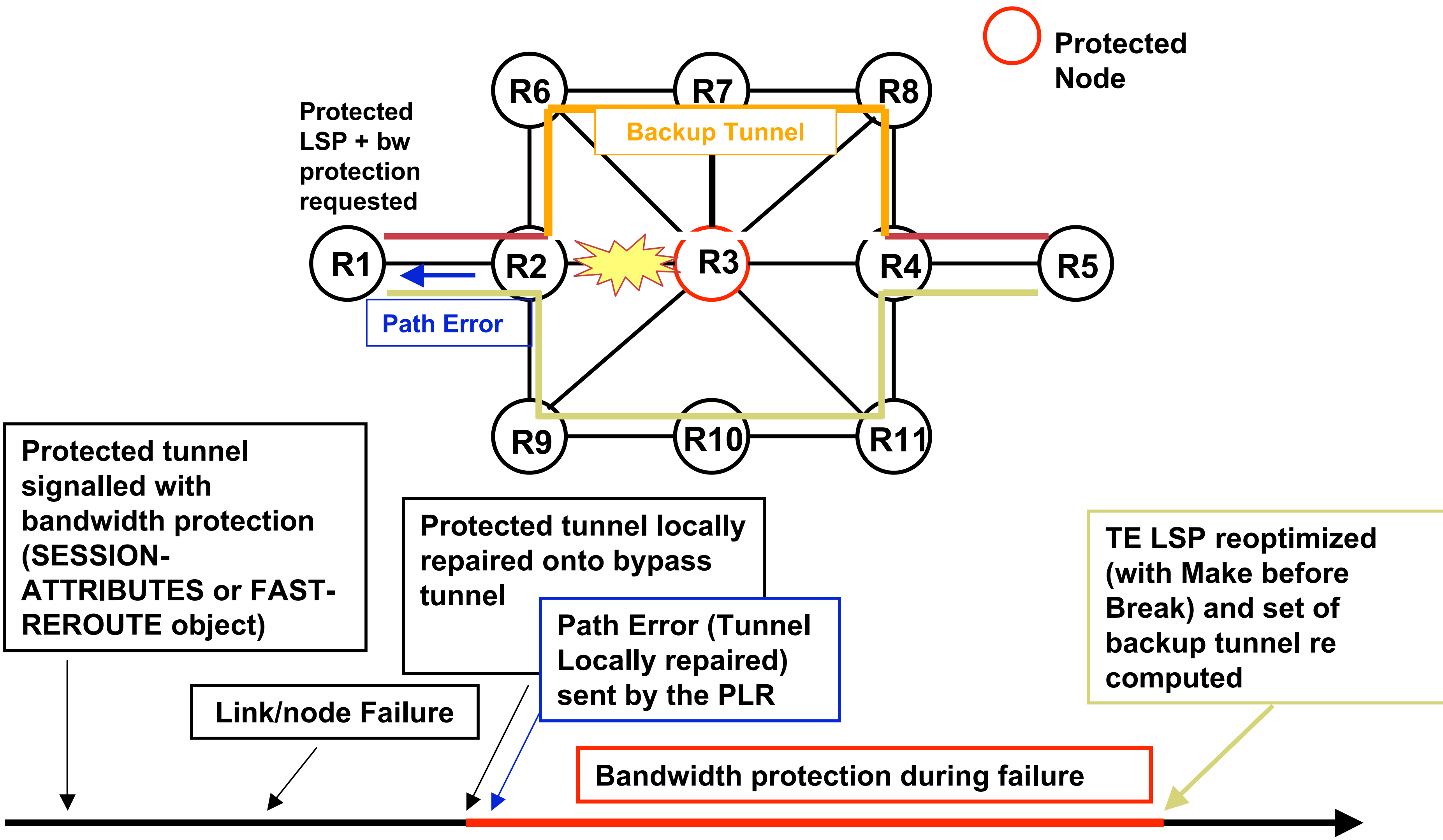
    Model for computing bypass tunnel paths that satisfy capacity constraints in the context of the MPLS TE Fast Reroute

    Guarantees bandwidth protection while allowing bandwidth sharing between backup tunnels

    Protects independent resources while preserving scalability

- Describes centralized and distributed path computation scenarios
- Addresses the required signaling extensions and optional routing extension
- The exact algorithms for the backup tunnel paths computation are beyond the scope of the draft

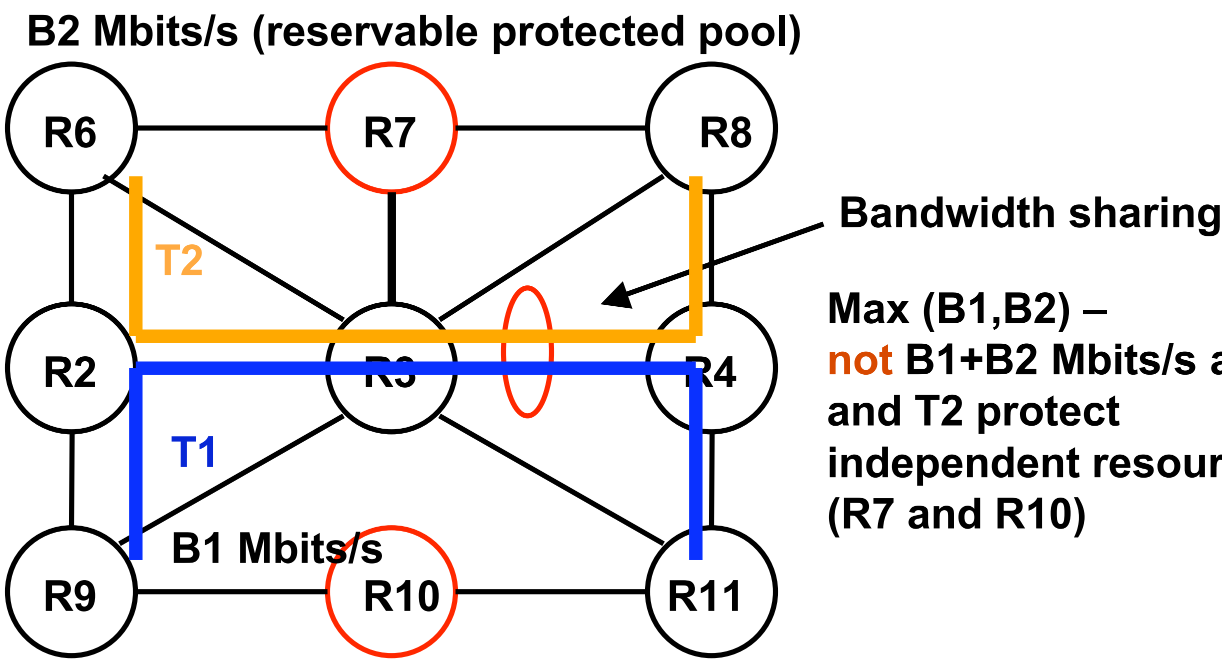
Bandwidth Protection



Bandwidth Sharing

Backup tunnels that protect independent resources (link/node/SRLG) can share bandwidth, resulting in large savings of bandwidth required for protection.

T1 and T2 protect independent resources (R7 and R10)



Max (B1,B2) – not B1+B2 Mbits/s as T1 and T2 protect independent resources (R7 and R10)

The assumption of a single, simultaneous failure is key for bandwidth sharing.

Bandwidth Sharing (Cont.)

- **MPLS TE Fast Reroute is a temporary mechanism**

  Backup tunnels are used until the TE LSPs are rerouted or reoptimized by head-ends and then traverse a protected path

  Only a short period of time

- **In practice, during that period:**

  Pb (multiple failures) <<1

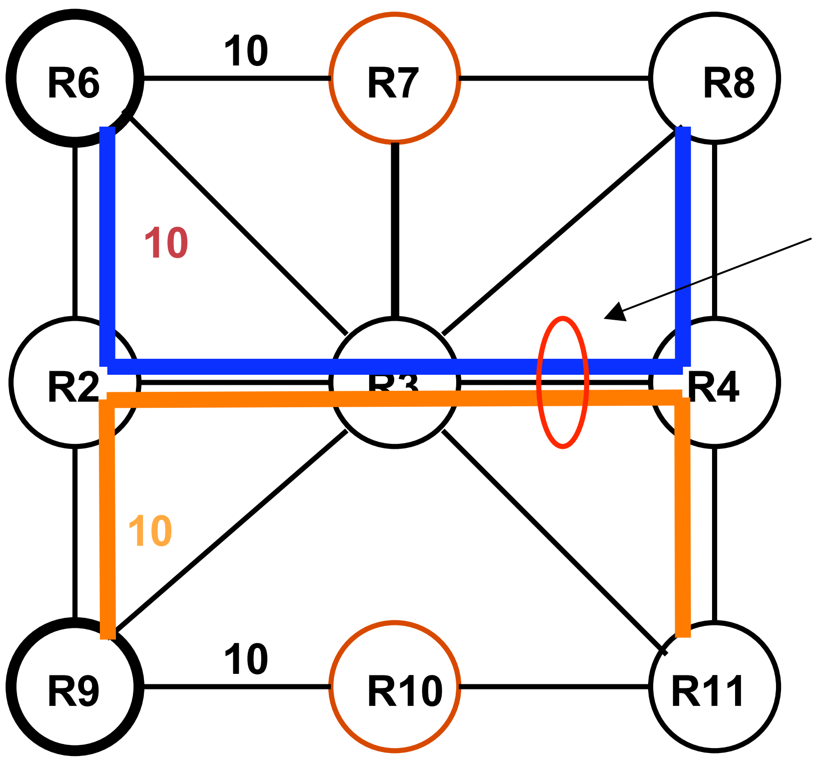
- **Validates the assumption of a single, simultaneous failure**

Naïve Model

- **Description**
  - Each backup tunnel is computed by its head-end (ie: using CSPF)
  - Backup tunnels are signaled with their respective bandwidth
- **+ Simple method**
- **- Limitations**
  - Inability to perform bandwidth sharing
  - Potential inability to find a solution when one does exist
  - Change of placement may help; however Naïve model cannot control that

Independent CSPF-based Computation Model

No bandwidth sharing



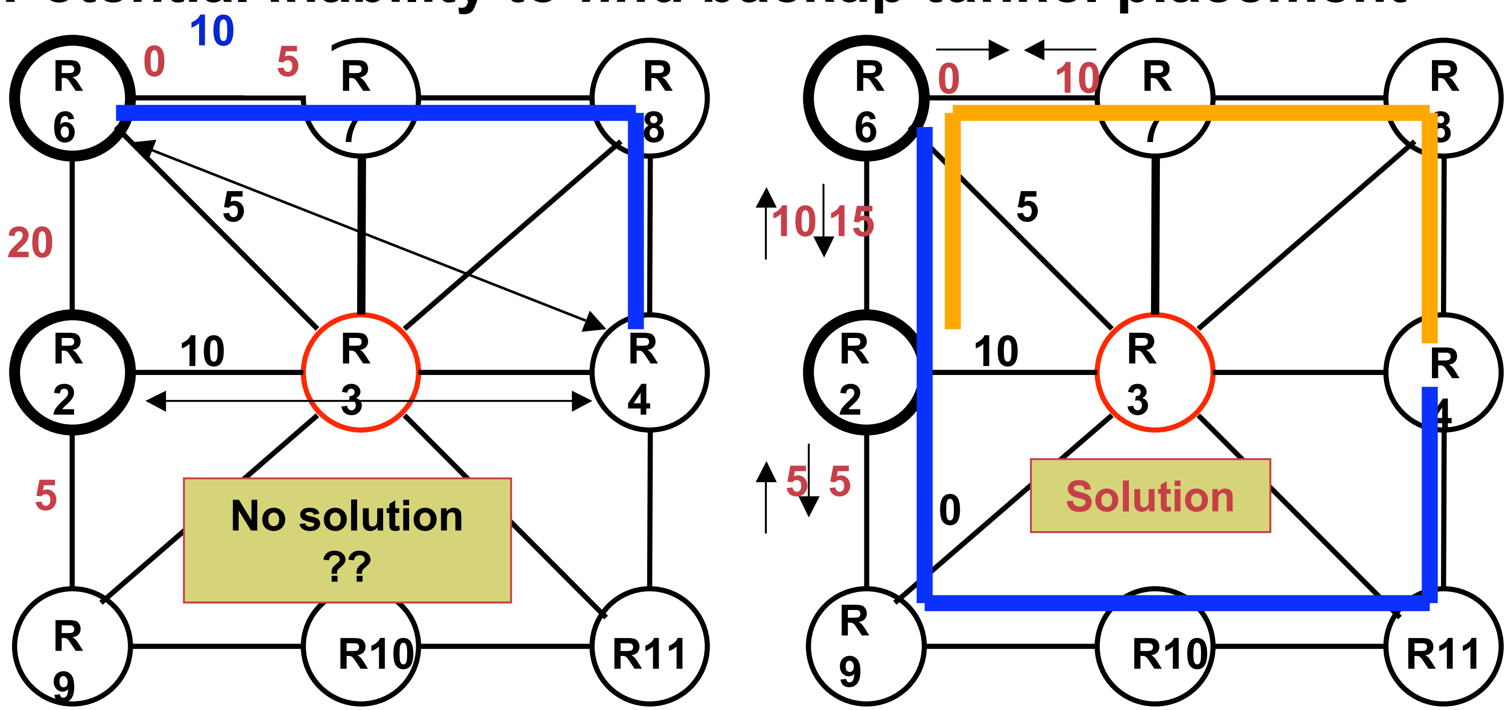
Lack of bandwidth sharing by two backup tunnels protecting independent resources.

10M+10M is reserved

Independent CSPF-based Computation Model (Cont.)

Cisco.com

Potential inability to find backup tunnel placement



R6 first sets up a 5M backup tunnels following the R6-R7-R8-R4 path => R2 can no longer find a 10M backup path

Bandwidth to protect

Available bandwidth

The problem comes from the non collaborative nature of this distributed computation

Facility-based Computation Model – Cisco Approach

Cisco.com

- One PCS computes paths for all backup LSPs that protect a given facility (even when they start on different head-ends)
- Note that a facility is one of the following

  Link (bi-directional)

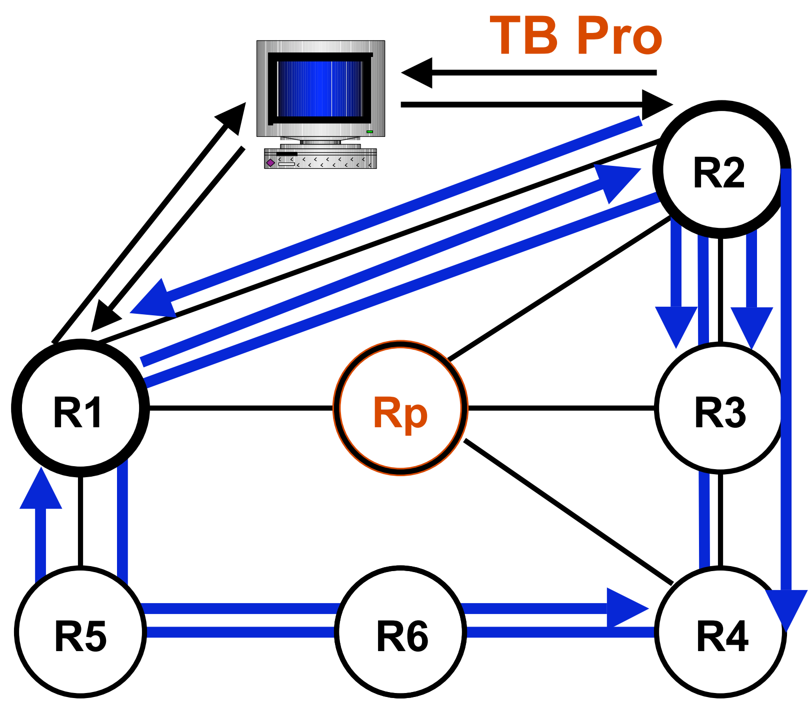
  Node

  Shared Risk Link Group (SRLG)

Complete sharing AND scalable (small amount of signalling and routing extensions)

Computation Scenario – Centralized Backup Tunnel Path

- Example: protection of Rp
- For each protected router, the PCS computes a set of backup tunnel to every NNHOP (or NHOP)
- For Ri I=<1…6>, Rp computes a set of backup tunnels from Ri to Rj with i<>j, whose paths exclude Rp, satisfying the bw constraints.
- So for R1: R1-…-R2, R1-…-R3, R1-…-R4.



Fast ReRoute

- Introduction
- Terminology of Protection/Restoration
- MPLS Traffic Engineering Fast Reroute
- Conclusion

Conclusion

Cisco.com

- **MPLS Traffic Engineering Fast Reroute provides:**

  → fast recovery using local protection

  →A wide scope of recovery: link/node/SRLG

  → in a scalable manner (BYPASS makes use of label stacking limiting the number of backup tunnels)

  → with stability … something crucial in large networks (fast local rerouting followed by the Head-end reoptimization)

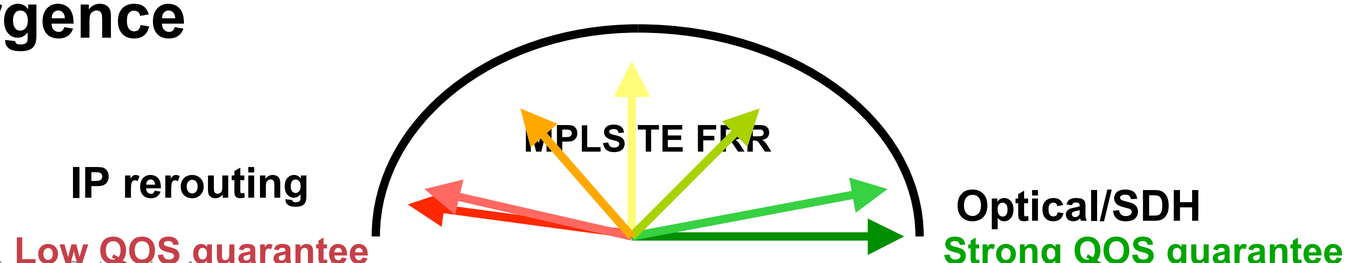
Conclusion

- **MPLS Traffic Engineering Fast Reroute provides:**

  → Bandwidth protection as other very well known and deployed protection schemes (SONET, Optical 1+1, …) but with a much more efficient backup bandwidth usage, in a scalable manner

  → With a high granularity. Different level of protection may be applied to various classes of traffic.

  Ex: an LSP carrying VoIP traffic will require a 50ms protection scheme as Internet traffic may rely on IP convergence



Thank You !