Techniques and Protocols for Improving Network Availability

Don Troshynski

dtroshynski@avici.com

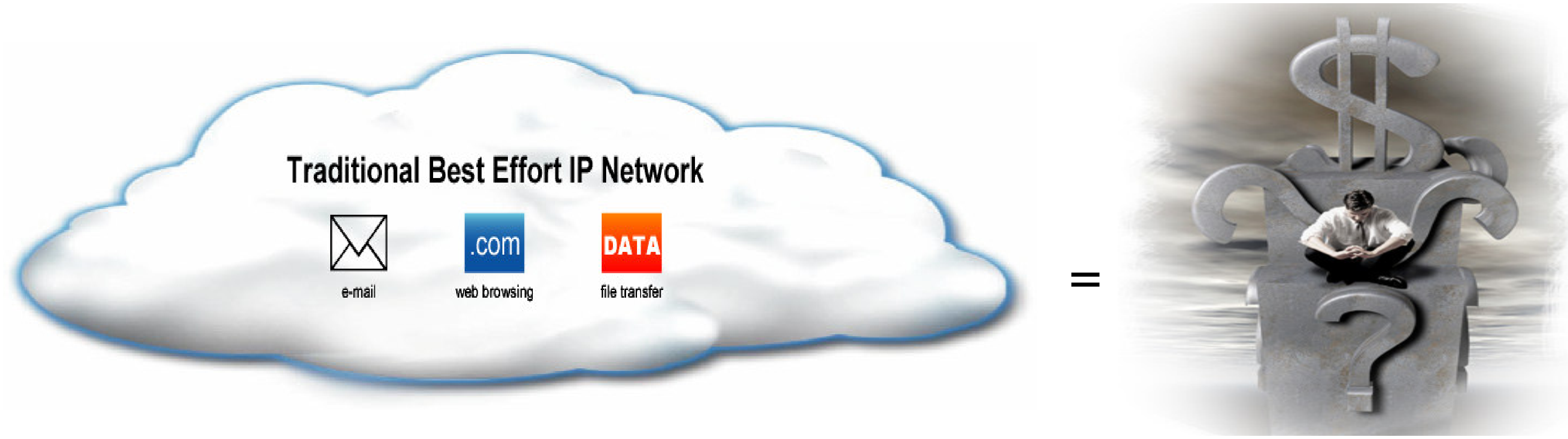
February 26th, 2004

Reliable Routing for the Internet

Outline of Talk

- The Problem
- Common Convergence Solutions
- An Advanced Solution: RAPID
- Increasing RAPID's coverage: U-turn Neighbors
- Applications of RAPID
- Summary

The Profitability Problem: Best Effort IP Network



- **Limited Services**
  - Supports only best effort services due to reliability and stability limitations
- **Low Margins**
  - Commodity pricing of undifferentiated best-effort services
- **High Costs**
  - High CapEx and OpEx outlays
  - Frequent outages and high customer service costs

Who will Capture IP's True Value?

???

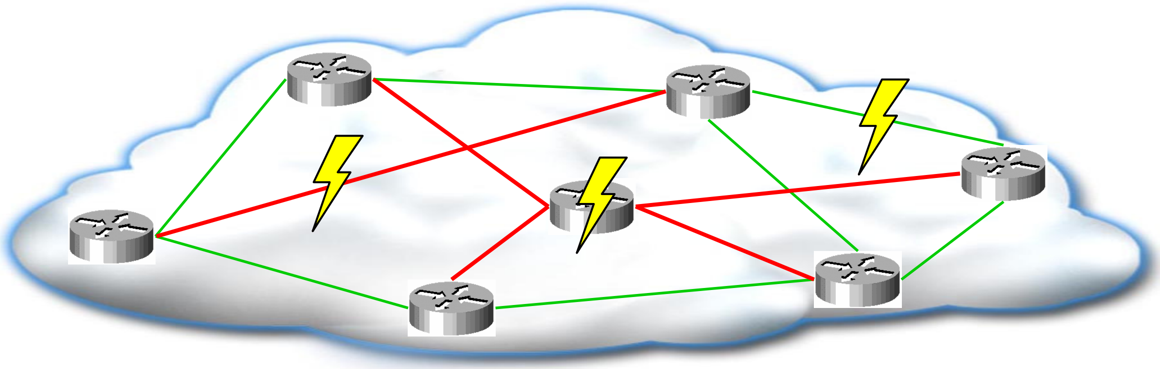
- Unified Messaging
- VPNs/Teleworking
- Delivery of audio, video, software
- Voice over IP
- Interactive Gaming

Service provider IP network today

- E-mail
- Web Browsing

The IP Challenge: To increase market share and gross margins carriers need to deliver more than just pipe

Network Disruptions are Daily Events



Causes
- Router Failure
- Disruptive Operations (sw upgrades, configuration changes, …)
- Link Failure

Service Impact
- Loss of traffic for 10s of seconds
- Disruption of Real-Time Services (voice calls, gaming sessions, video, ATM)

Business Impact
- SLA Penalties
- Customer Service/Maintenance Issues
- Customer Churn
- Inability to support High-Margin Real-Time Services

Traffic Convergence Goal: < 50 ms

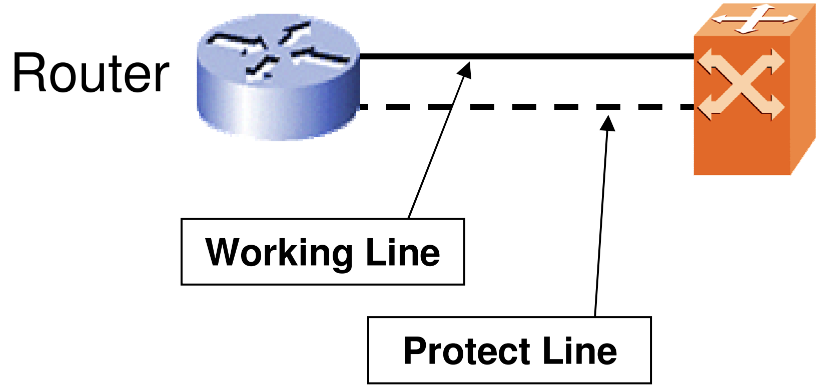
- To support a multi-service network, need to minimize service interruption.
- Network Failures cause service interruption.
  - Node Failure: Avoid disruption with Non-Stop Routing
  - Link Failure: Minimize traffic loss during convergence.
- Traffic Convergence
  - IGP Convergence: SPF provides the basis for all other protocols so must be very fast.
  - BGP Convergence: Using forwarding-plane indirection to IGP next-hop allows traffic restoration for BGP learned destination ***before*** BGP re-computation occurs for many failure scenarios.
  - LDP Convergence: Requires IGP SPF results to install new forwarding plane state.

Outline of Talk

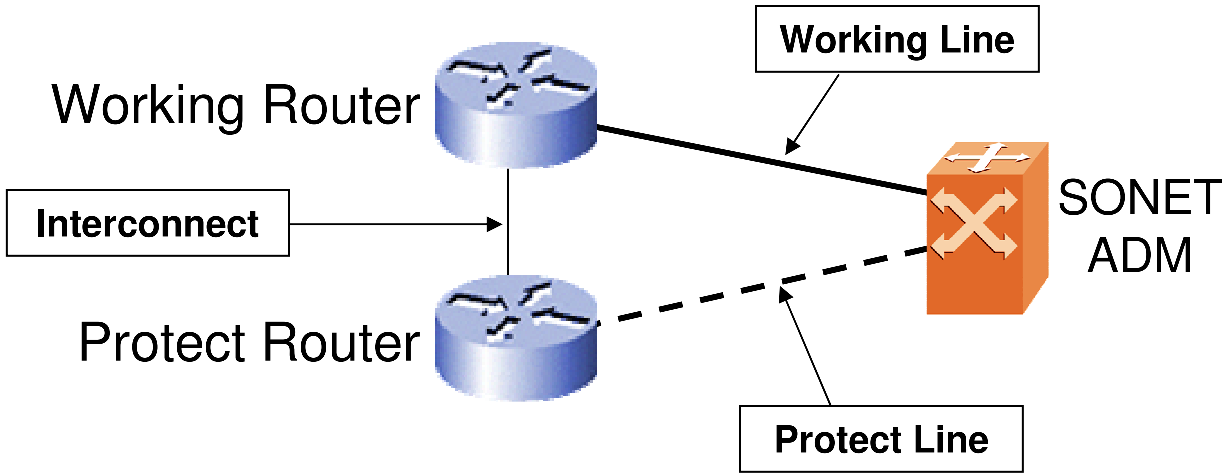
- The Problem
- Common Convergence Solutions
- An Advanced Solution: RAPID
- Increasing RAPID's coverage: U-turn Neighbors
- Applications of RAPID
- Summary

Layer 1: Linear and Distributed APS

- Linear APS
  - Comes in 1+1 and 1:N flavors
  - Works at Line Layer
  - Signaled in K1/K2 bytes

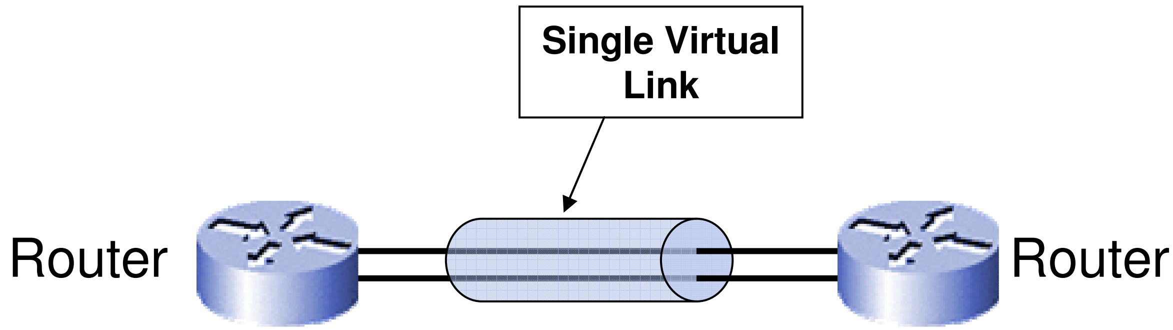


- Distributed APS
  - Like Linear APS, but two routers terminate the working and protect lines
  - Failure of line, or even router, is protected



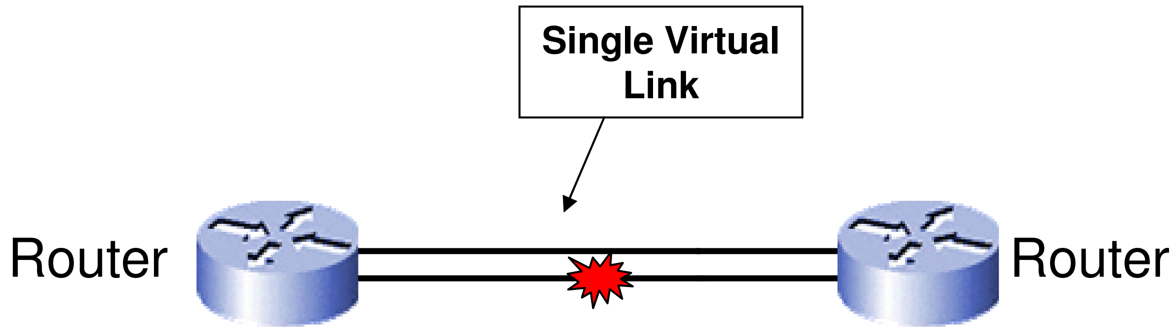
- APS is costly to implement and therefore a targeted solution

Layer 2: POS Bundling Description



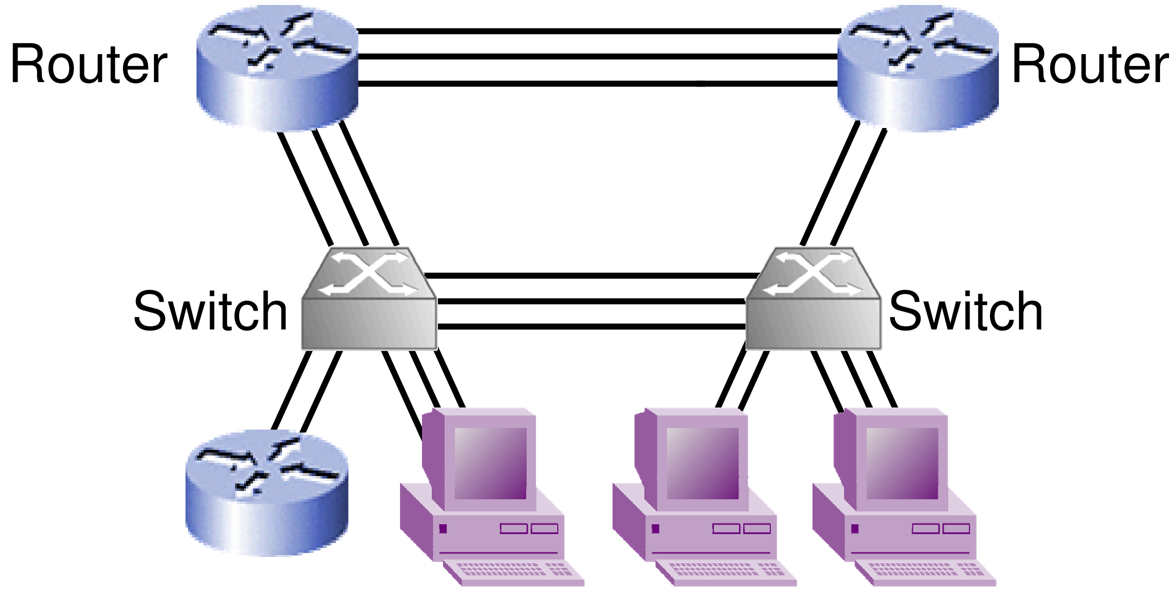
- POS equivalent of GigE Link Aggregation
- A method to aggregate 2 or more physical POS links into a single logical link as observed from Layers 2 and 3
- Network sees a single IP Address/Interface
- “Flows” comprised of IP src/dest or MPLS LSPs routed to a single bundle member
- Source/Dest IP Address and MPLS label based hashing algorithm for traffic flow (same as ECMP)

Layer 2: POS Bundling Protection



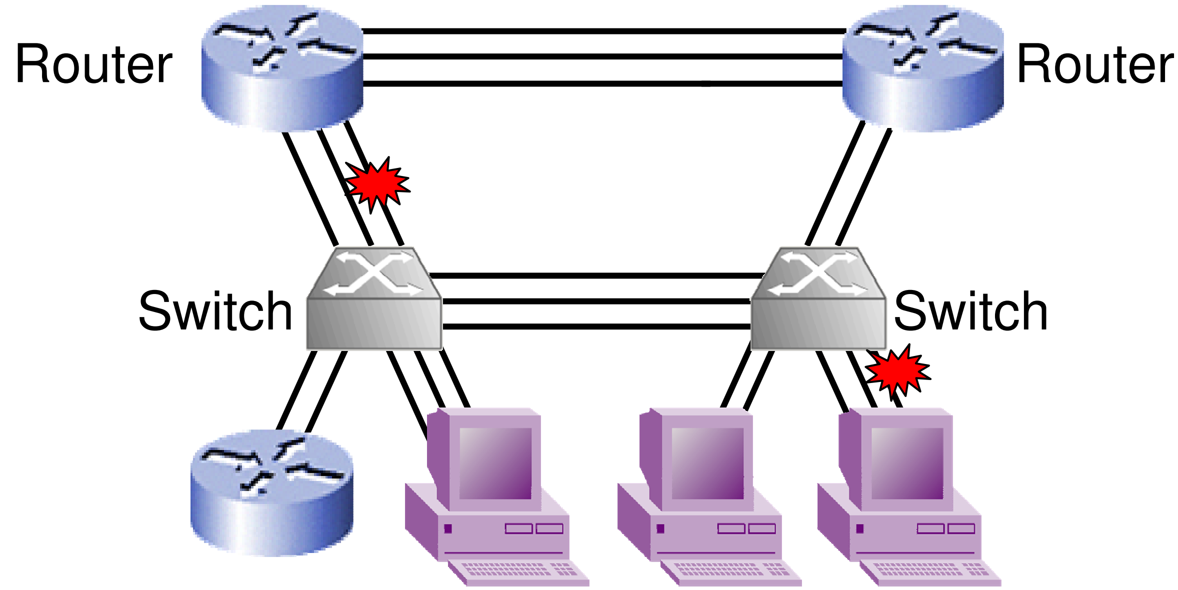
- When one or more fiber fails, traffic shifted to remaining members
- Failure transparent to IP routing layer – bandwidth of “link” just decreased
- Switchover can be performed in <50ms
- Network does not need to re-converge at Layer-3
- Some products even support mixed member link speeds
- Better link utilization than APS but applicable only to parallel links

Layer 2: Link Aggregation Description



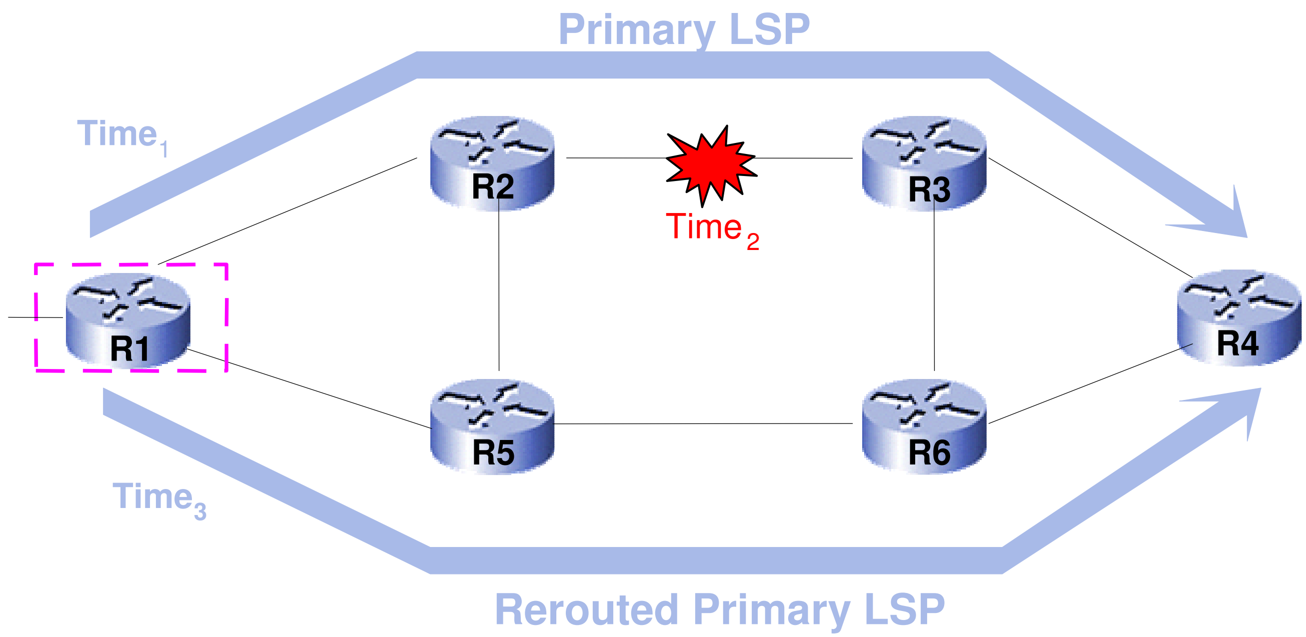
- Aggregate-links are a number of individual Ethernet links that collectively form a single Layer 2 link using the IEEE 802.3ad standard
- Upper layer protocols (Spanning Tree, IS-IS, OSPF, BGP, etc.) and applications see the link aggregation as a single interface
- Conversation “flows”, which could be defined by MAC src/dest or IP src/dest, are kept on the same Link Agg member link

Layer 2: Link Aggregation Protection



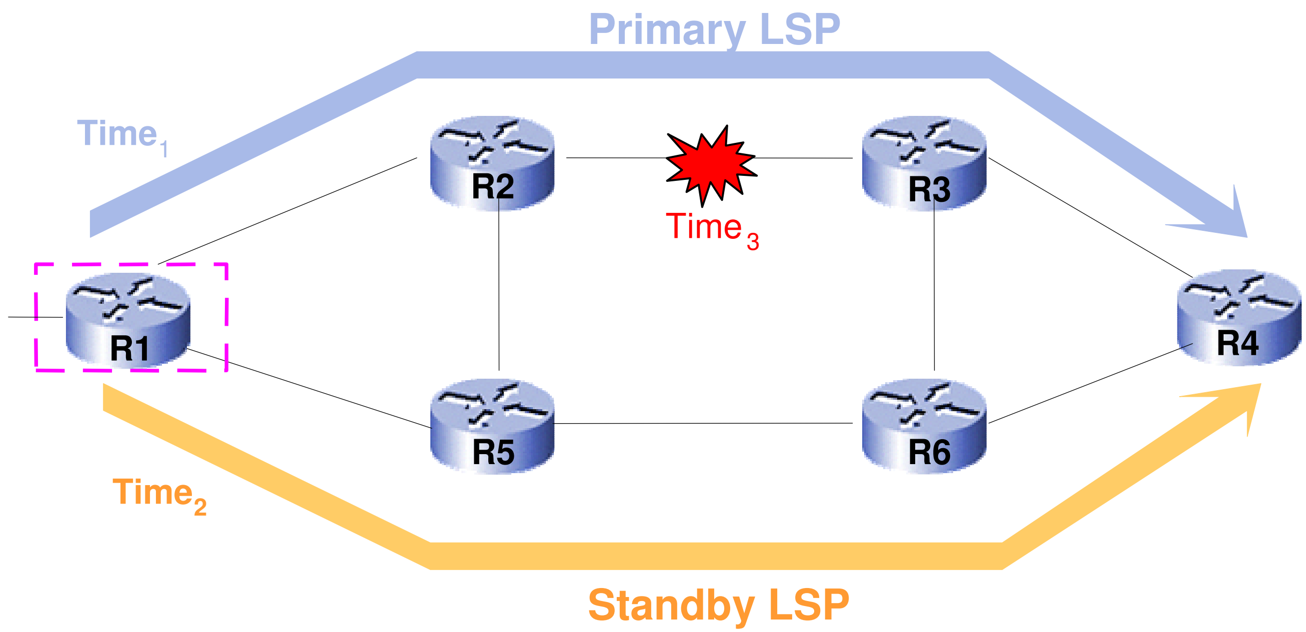
- There is an automatic configuration ability, through the use of Link Aggregation Control Protocol (LACP)
- LACP also provides a keepalive mechanism
- If a failure occurs on a link, traffic shifted over to remaining member links
- Switchover may happen quickly (<1sec) - upper layers don't see the failure

Layer 2.5: Head-end Rerouted LSPs



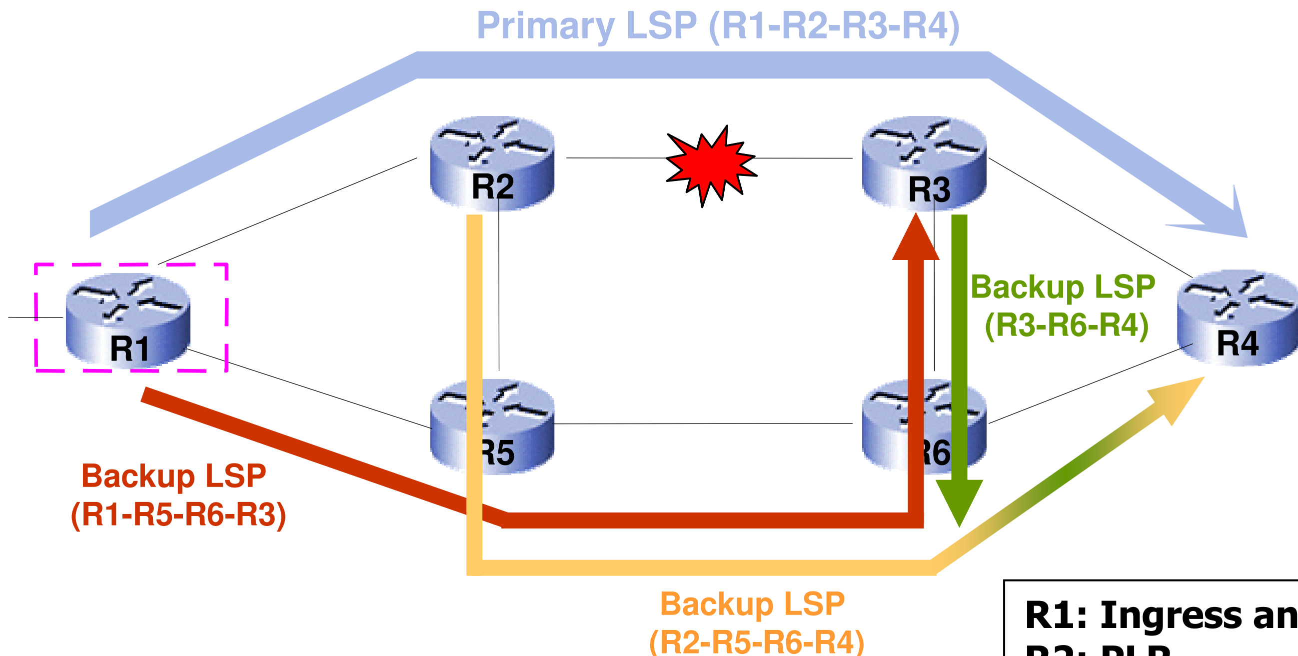
- Planning Occurs After Failure
  - Tunnel Ingress Detects Failure
  - Perform CSPF to Reroute LSP
- Recovery is Order(seconds)
- Packet Loss INCREASES as failure moves away from Ingress
- CSPF and flooding is very sensitive to size of TE topology

Layer 2.5: Pre-Signaled Standby LSPs



- Planning Occurs Before Failure
  - Tunnel Ingress Detects Failure
  - Move Traffic to use standby LSP
- Recovery can be in 100s of milliseconds
- Packet Loss INCREASES as failure moves away from Ingress

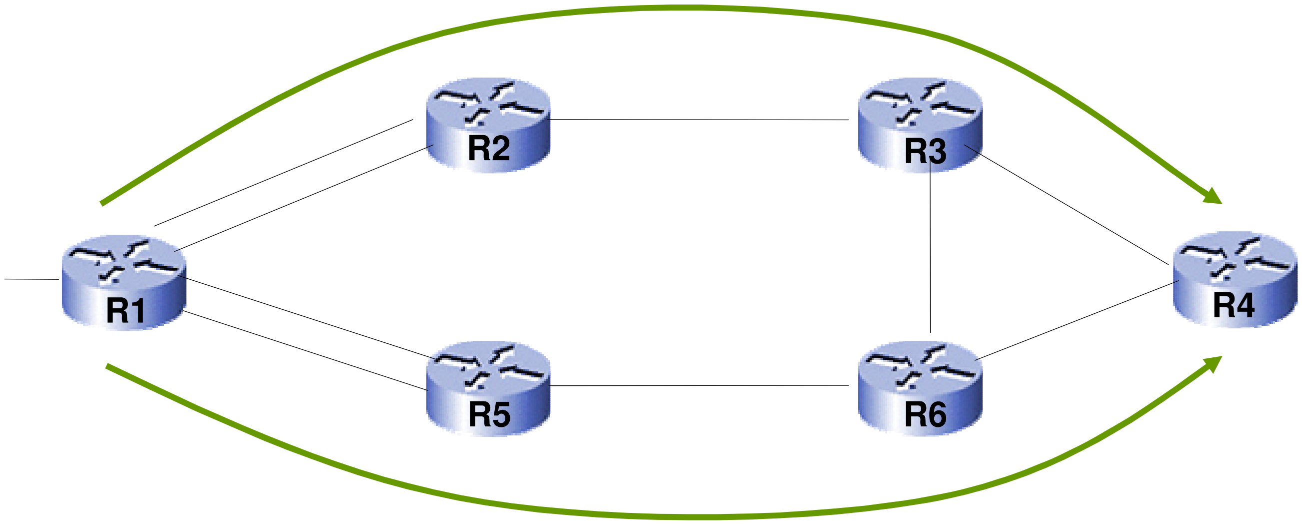
Layer 2.5: Fast Reroute



R1: Ingress and PLR
R2: PLR
R3: Merge Node and PLR
R4: Egress and Merge Node
R5, R6: Backup Midpoints

- Planning Occurs Before Failure
  - PLR Detects Failure
  - Move Traffic to use Backup LSP
- Recovery in 10s of milliseconds
- Increased state in network and potential for unused bandwidth
- Quickly becomes complex to manage and troubleshoot

Layer 3: Equal Cost Multi-Path (ECMP)

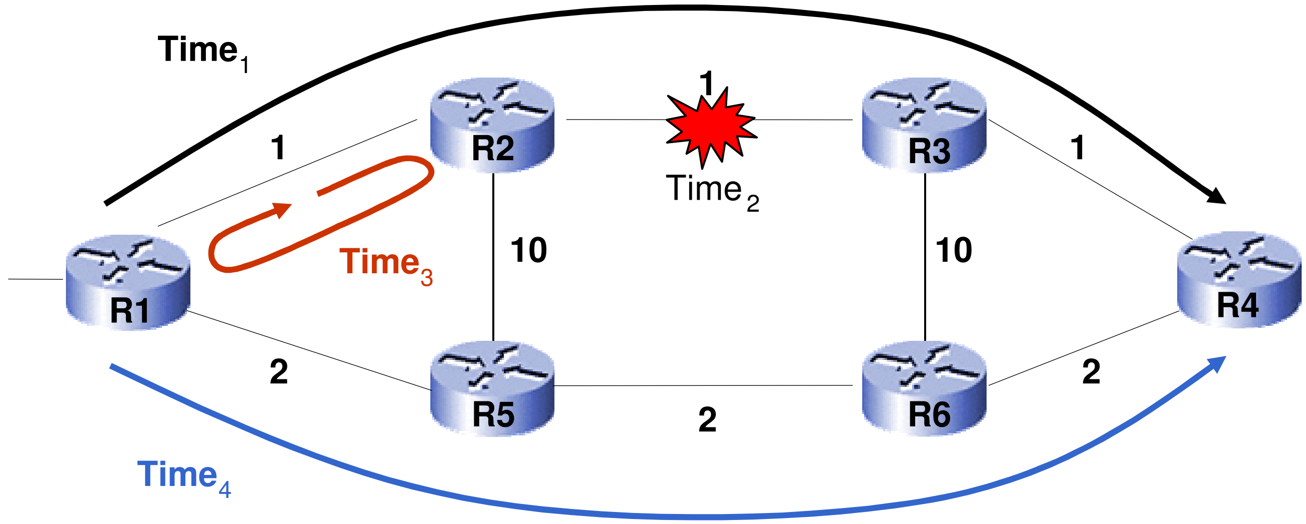


- ECMP can be used for either node or link protection
- Problems with ECMP:
  - Can have long failover times due to IGP flooding and SPF (same issue as IGP convergence, if failure occurred remotely)
  - IGP costs have to be the same (potentially complex traffic engineering)
- Really it's the same issues as normal IGP convergence, except half (or more) of the traffic won't be affected by the failure

Outline of Talk

- The Problem
- Common Convergence Solutions
- An Advanced Solution: RAPID
- Increasing RAPID's coverage: U-turn Neighbors
- Applications of RAPID
- Summary

IGP Convergence Time



- Failure detection triggers R2 to re-converge and signal failure
  - Link loss happens in ms, but keepalive timeout can be seconds
  - R2's convergence time doesn't matter – only its failure detection and signalling time does
- Packets now loop until R1 receives signal, processes it, and re-converges (and perhaps R5 needs to as well)

Why It Takes So Long

Detection of SONET layer failure

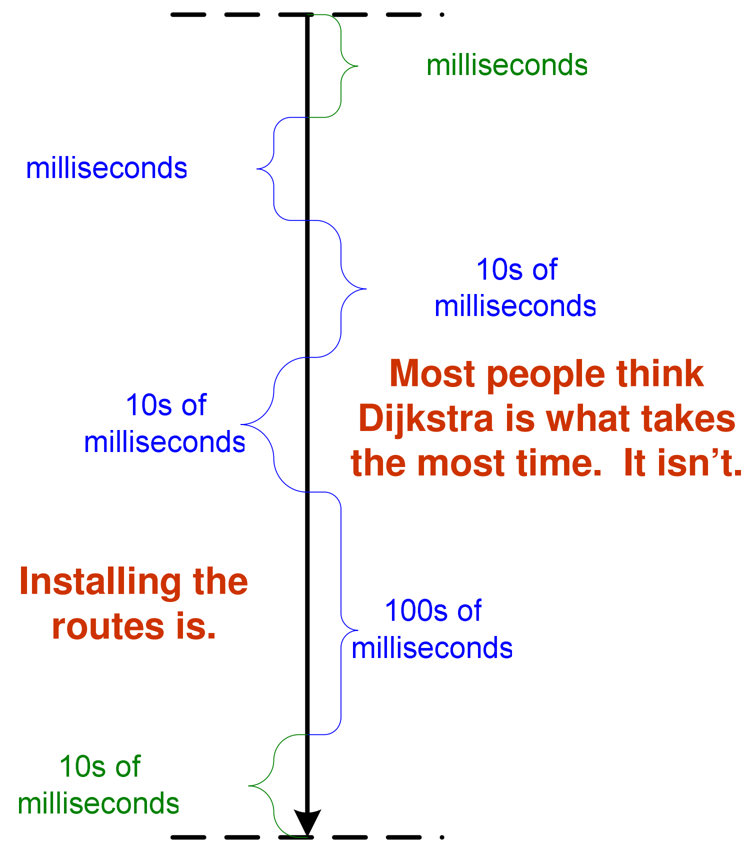
Report failure to Route Controller

Generate and flood an LSP

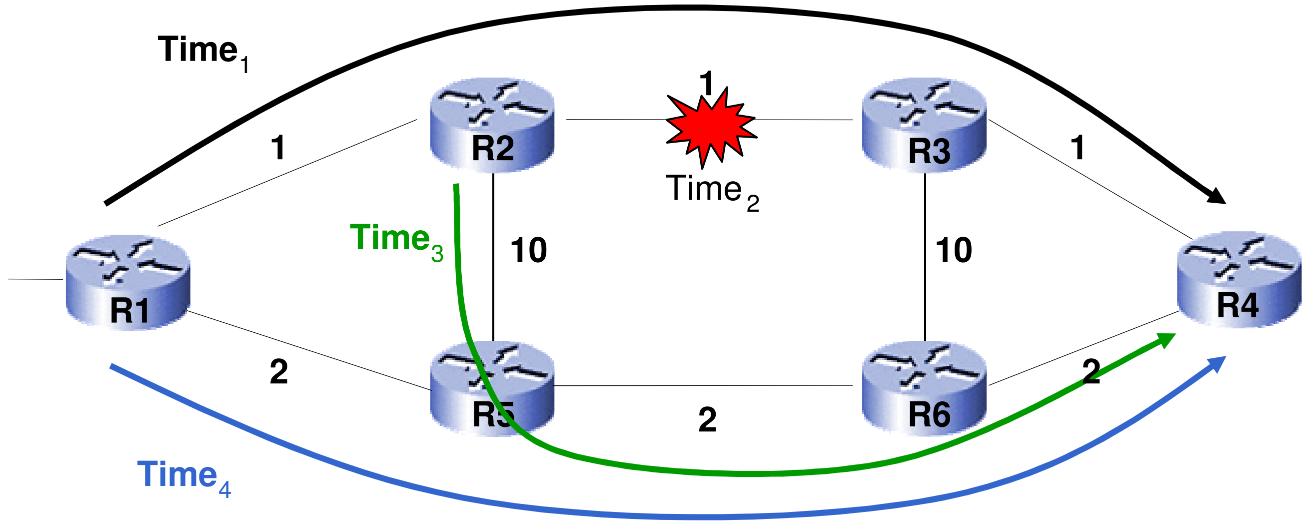
Trigger and Compute an SPF

Communicate new Next-Hops to linecards.

Install new Next-Hops into hardware path on each linecard.



Reliable Alternate Paths for Internet Destinations (RAPID) -- Basic Concept

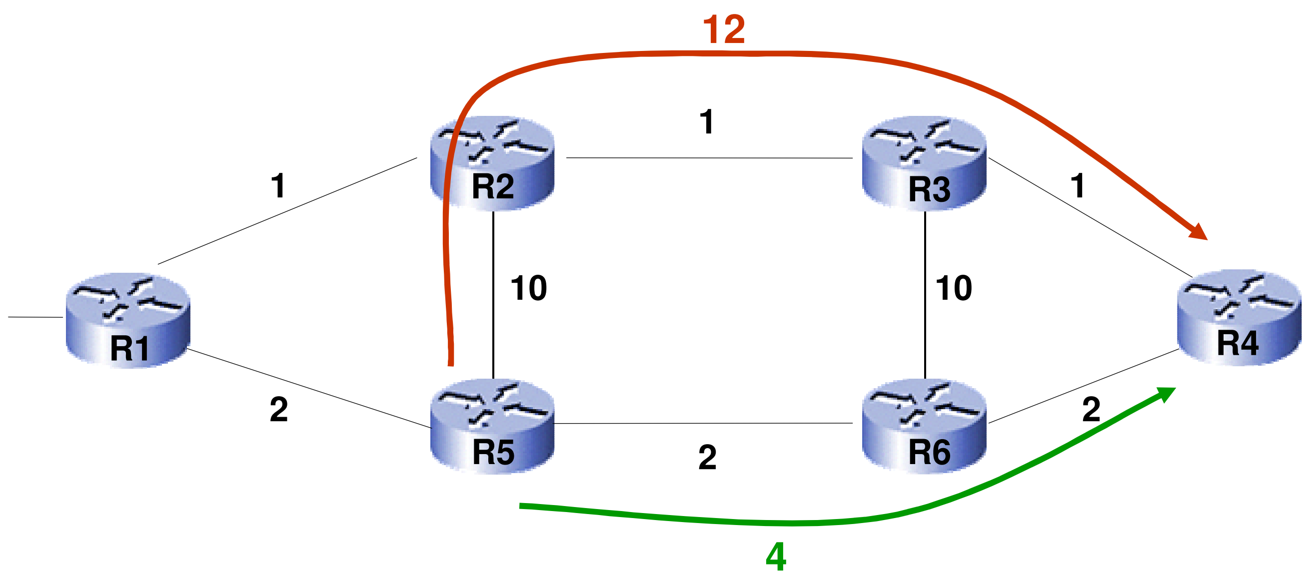


- R2 pre-computes alternate IGP path for R4 traffic in case link fails
- Failure detection triggers R2 to failover to alternate path
  - Failover occurs in **milliseconds** for both IP and LDP
  - R2 also signals failure and runs SPF, but that time does not impact traffic
- Some time later R1 will have run a new SPF and send traffic to R5

Alternate Next-Hops

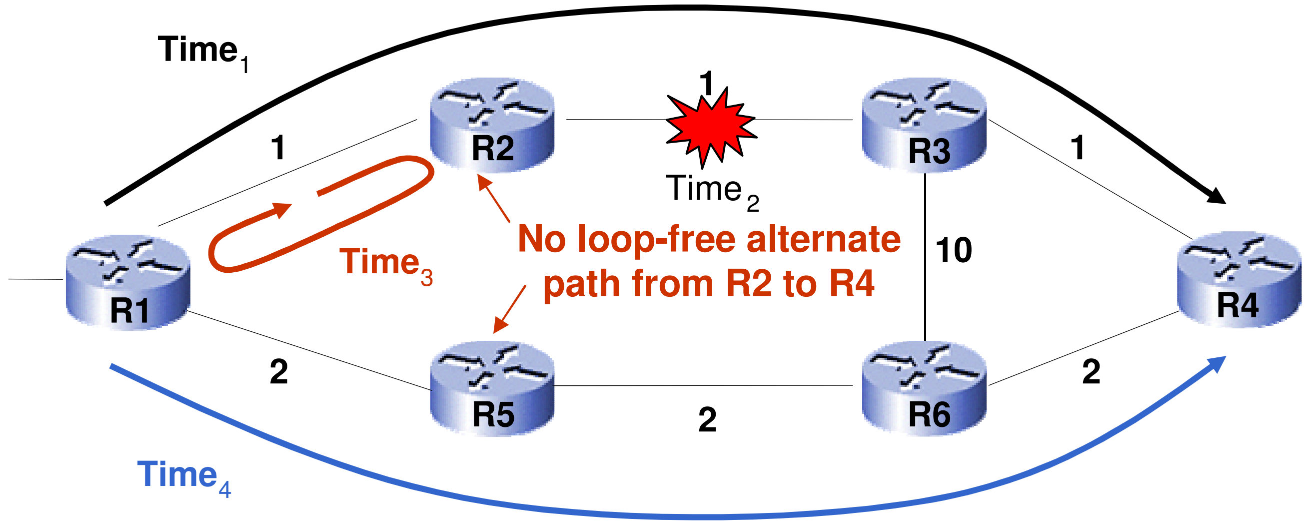
- Pre-computed with previous Dijkstra SPF calculation
- Used during a local link failure while Router is computing a new SPF based on the revised topology and is installing it into the forwarding plane
- Must not cause forwarding loop during failure
- Feasible Alternate Next-Hop can be used for LDP as well as IGP/BGP to provide sub-second traffic re-direction
- Once new SPF runs, it overrides the RAPID alternate path

Finding Loop-Free Neighbors



- R2 can find a loop free neighbor: R5
- R5 is loop-free, because the distance from R5 to R4 is less than the distance from R2 to R4 plus the distance from R5 to R2.
- R2 can know all this because it has the full LSDB
- Only R2 needs to support RAPID to provide protection for its links
- This allows a slow migration to RAPID protection

Loop-Free Coverage

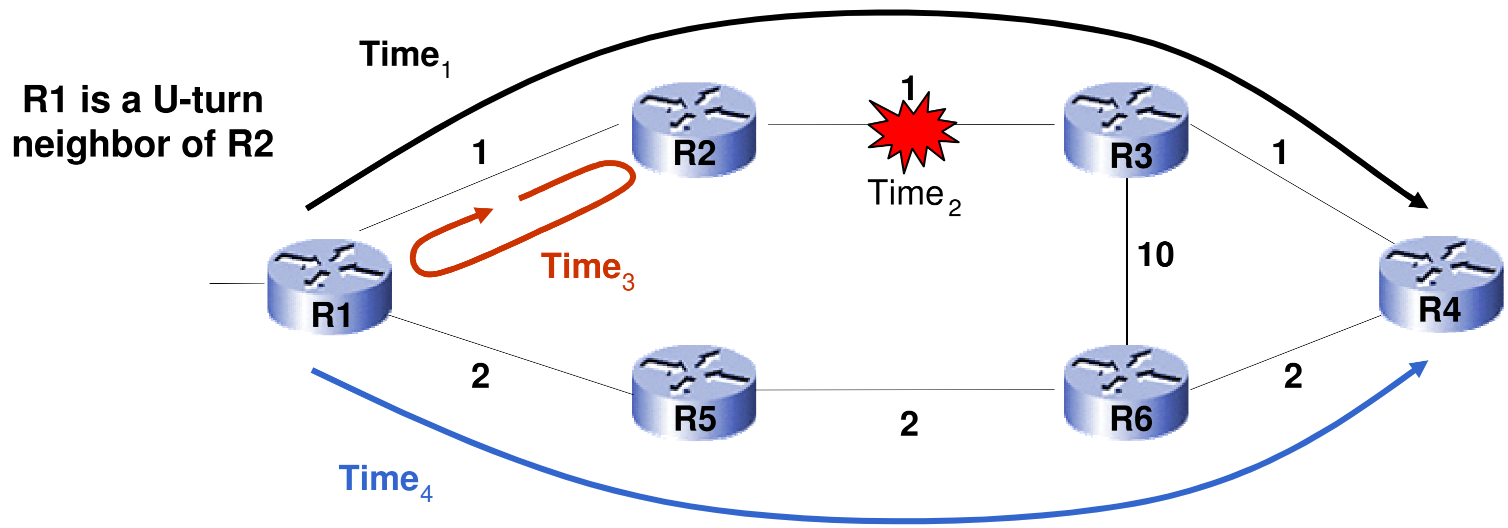


- Many networks don't have alternate links at all points
  - Simple loop-free RAPID provides an average 75% failure coverage
  - But 75% of the links does not equal 75% of the traffic – could be a lot less if the 25% unprotected are important links
- If R2 could use R1 as an alternate, the coverage would increase dramatically

Outline of Talk

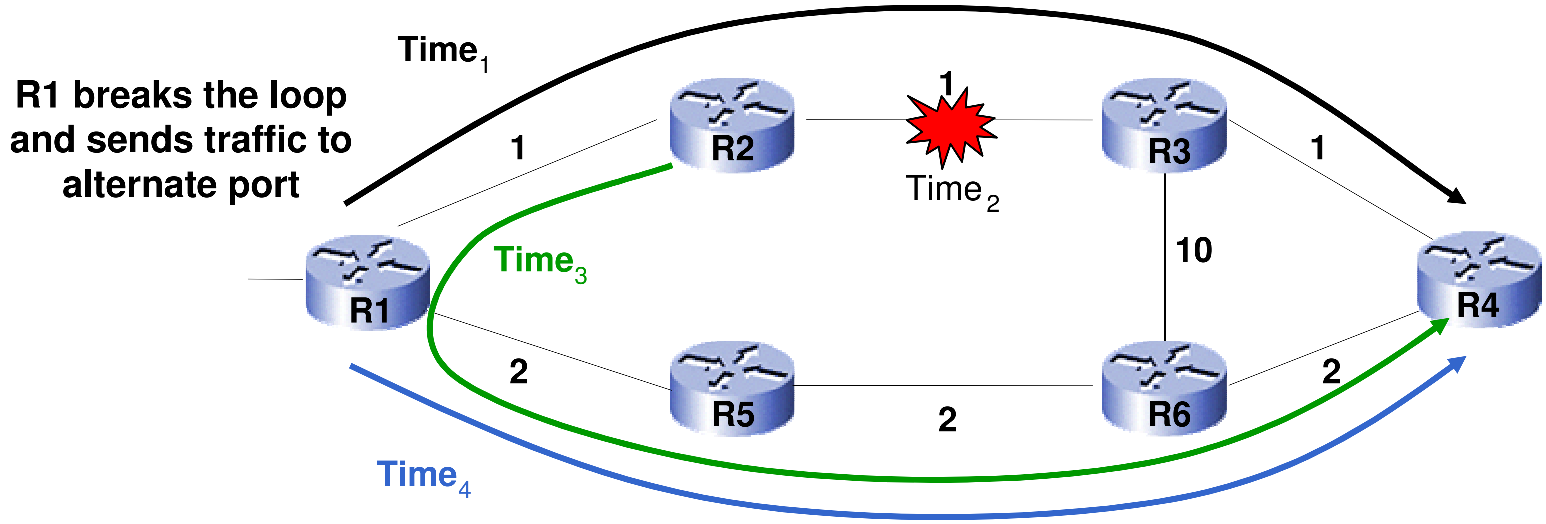
- The Problem
- Common Convergence Solutions
- An Advanced Solution: RAPID
- Increasing RAPID’s coverage: U-turn Neighbors
- Applications of RAPID
- Summary

Breaking the loop: U-Turn Alternates



- R1 is a U-turn neighbor of R2 because:
  - R1 itself has a loop-free alternate path to reach R4
  - R1 can break the loop
- So R2 could use R1 as an alternate, if R1 were capable of breaking the loop when a failure happens

U-Turn Alternates



- R1 can break the loop, if its hardware can forward traffic received on the "wrong" port to the alternate path port
  - The receiving port is "wrong", if it's the port that the traffic should be transmitting back out on
- This means R1 has to be doing RAPID, and have the hardware to be a U-turn alternate
- Thus new IETF drafts to signal capabilities: OSPF, ISIS, LDP

RAPID Details

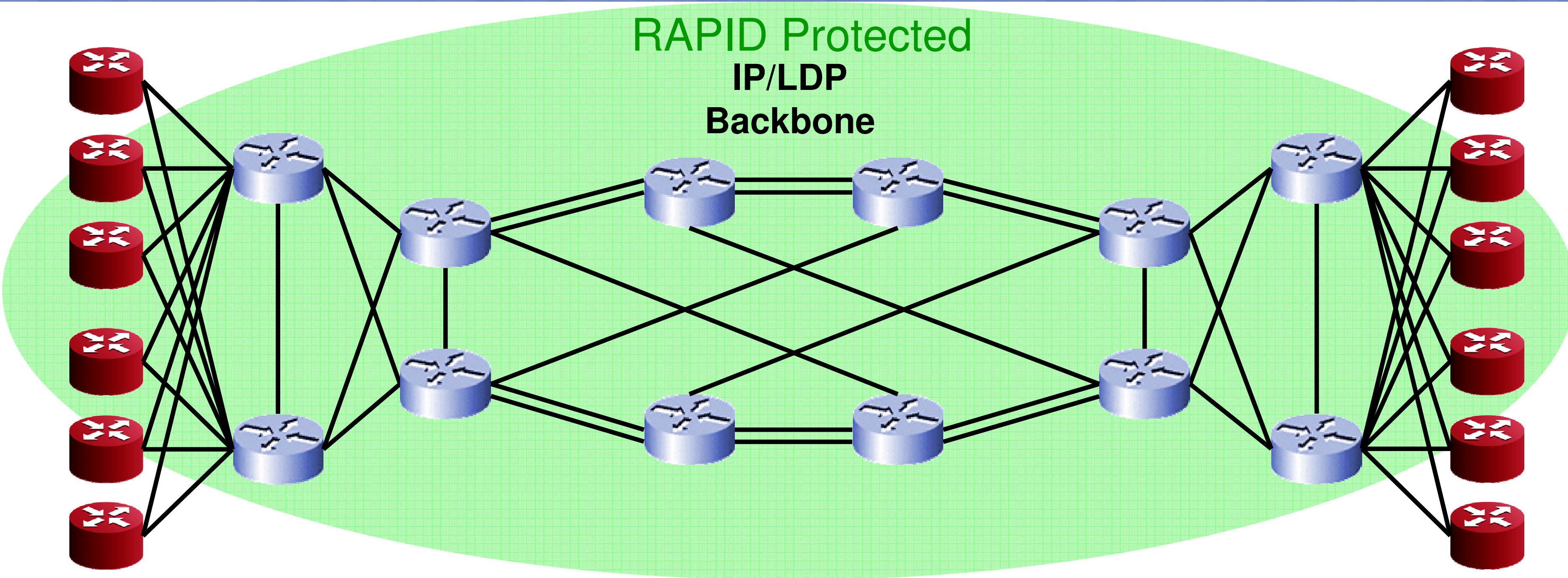
- New IETF drafts define signaling of Router's RAPID capability, and per link capability for IGPs and LDP
- Drafts also define common rules for selecting loop-free and U-turn alternates
- Using U-turn alternates increases protection coverage from 75% average, to 95% average

- User-configurable for simple (non-U-turn) RAPID, or for full RAPID
- Asymmetric costs taken into account
- Currently multicast not covered by RAPID – uses old convergence method (under investigation)

Outline of Talk

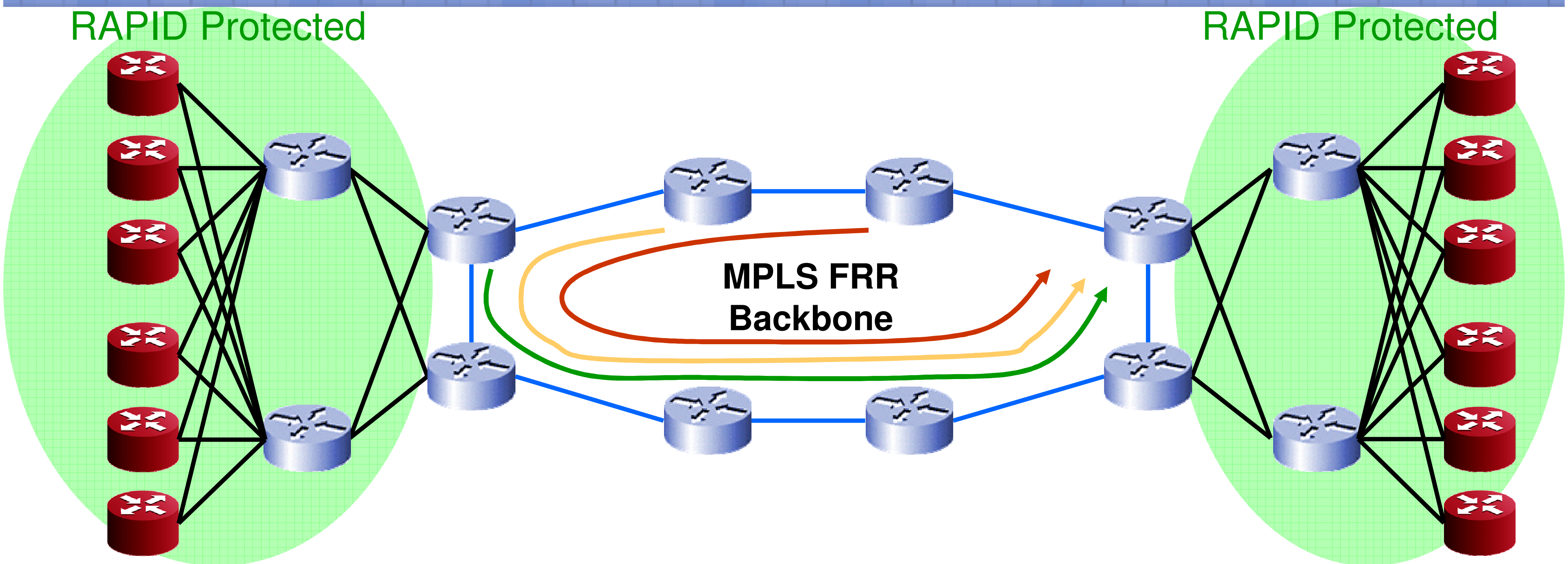
- The Problem
- Common Convergence Solutions
- An Advanced Solution: RAPID
- Increasing RAPID's coverage: U-turn Neighbors
- Applications of RAPID
- Summary

Network 1



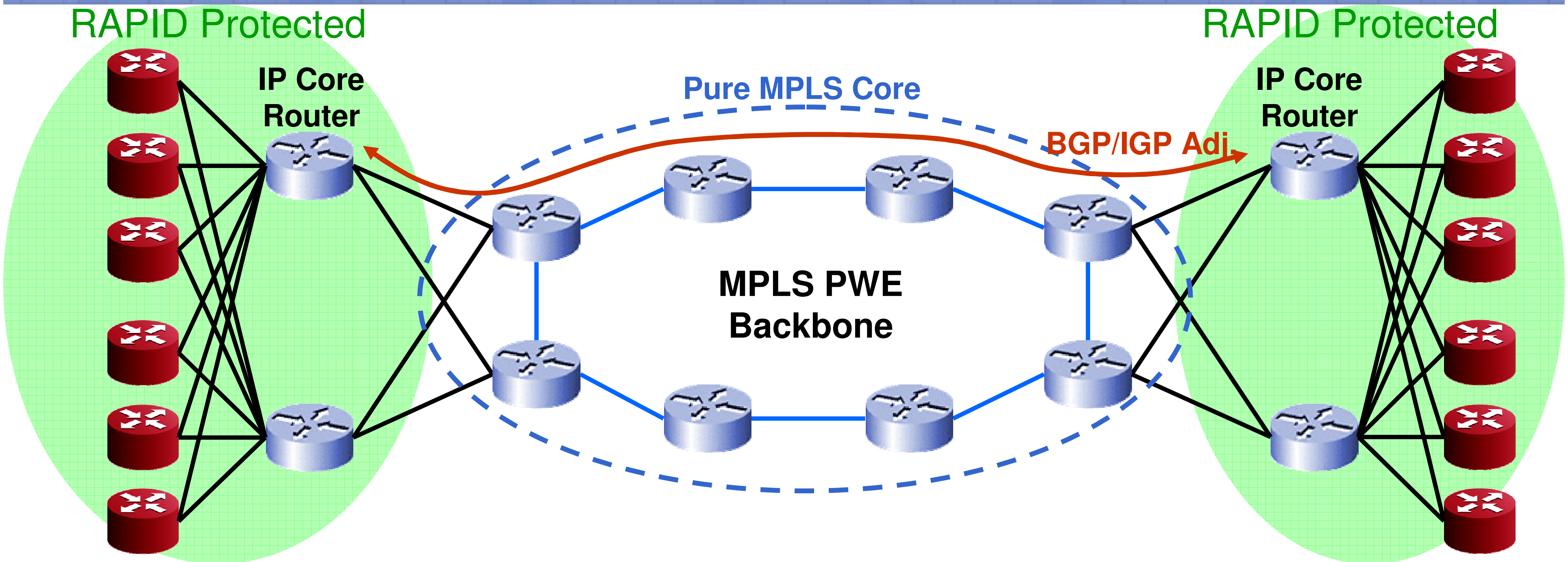
- A highly redundant IP/LDP Backbone – no MPLS-TE
- RAPID provides protection both in the Core, Aggregation, and Edge
- Coverage is very good, if the link redundancy is sufficient

Network 2



- Less redundant design using MPLS-TE in core
- FRR provides loop-free method to backup logical-ring core
- RAPID protects Aggregation/Hub/Edge routers

Network 3



- Separate Transit MPLS PWE/L2-VPN Core design
- IP Core routers do not “see” this MPLS core – they think they have direct connections to the other IP Core routers
- MPLS Backbone can be protected by FRR or Pre-Signaled standby tunnels (more common)
- RAPID protects IP Core Routers (not Transit Backbone Routers) and Edge

SYSTEMS

Outline of Talk

- The Problem
- Common Convergence Solutions
- An Advanced Solution: RAPID
- Increasing RAPID's coverage: U-turn Neighbors
- Applications of RAPID
- Summary

Well-Known Convergence Solutions

- Layer-1/2:
  - APS: standard SONET Line-layer protection mechanism – fast failover, but wastes protection link
  - Composite Links/Link Aggregation: multiple parallel links to neighbor – fast failover, but no node protection/interoperability
- Layer-2.5:
  - Head-end Reroute: re-signal LSP path from ingress to egress – slow failover (<10 seconds)
  - Head-end Standby Tunnels: pre-signaled from ingress to egress – slow failover (100s of milliseconds)
  - MPLS Fast Reroute: MPLS-TE based local protection (1:1 or 1:N) – fast failover, somewhat complicated and doesn't scale well
- Layer-3:
  - ECMP for IP/LDP: multi-path load-sharing based on IGP cost – slow failover and requires careful planning for equal cost paths

RAPID Fast Convergence Summary

- Provide < 50ms traffic convergence in the event of a link failure for IP and LDP traffic.
- Loop-free alternates can be used independent of LDP Fast Convergence on the alternate next-hop.
- U-Turn Alternates expand the potential failure coverage on networks.
- Simple to configure and manage
- Can be incrementally deployed – the benefit of U-Turn Alternates will be seen as more routers are deployed with this feature in the network.

RAPID Benefits vs. Alternatives

	Current Protection Mechanisms (MPLS FRR/ APS/ etc.)	IP/LDP Fast Convergence
Costs	▪ Expensive unused protection capacity	▪ More effective utilization of network assets
Complexity	▪ Requires high skill set ▪ Time intensive and manual ▪ Requires RSVP-TE overlay	▪ Simple configuration requires no specialized training or resources ▪ Automated and adaptive
Flexibility	▪ Provisioning requires constant updates as network changes	▪ Seamless adaptability to network events (topology/ failures/ etc.)
Scalability	▪ Backup tunnels use more resources ▪ Tunnels rarely deployed edge-to-edge	▪ Seamless adaptability to network events (topology/ failures/ etc.) ▪ Protects end-to-end