ROUTING RESILIENCY

DENOG6

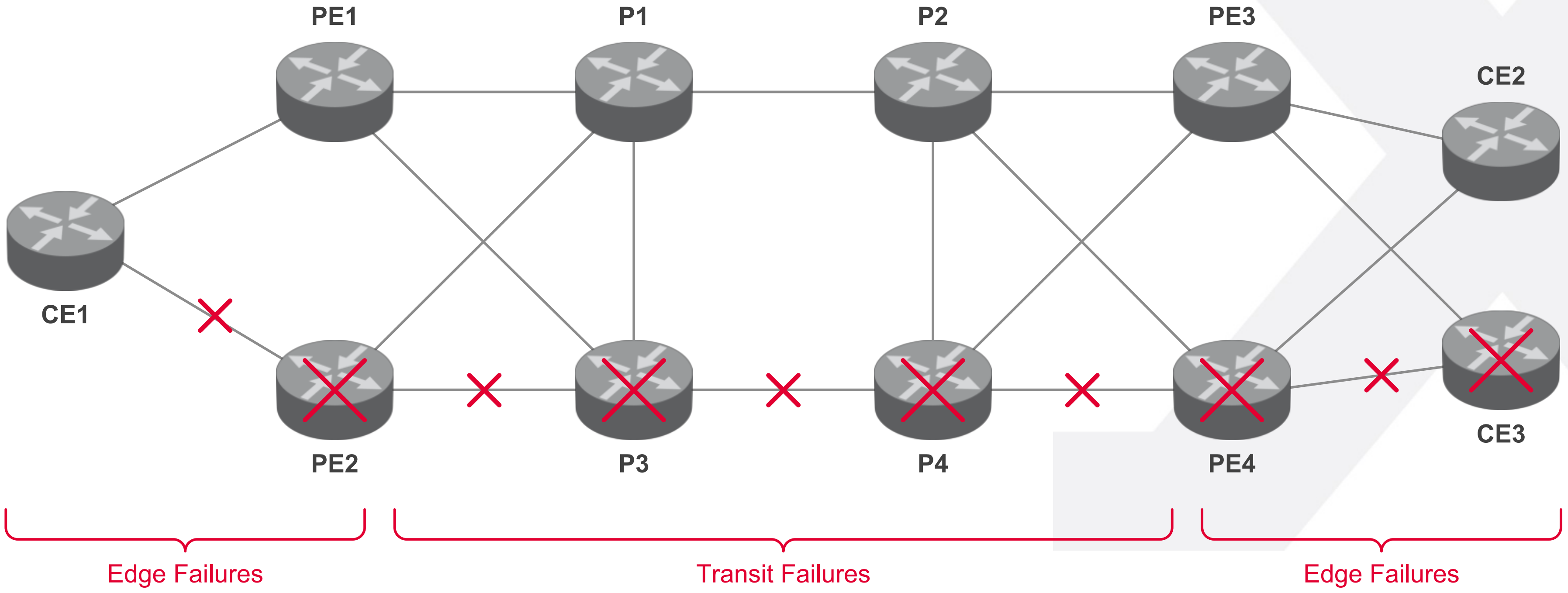
PETER SIEVERS

Agenda

- Problem Definition
- MPLS Fast Reroute
- Loop-Free Alternates
- Segment Routing

What‘s the problem?

- As network infrastructure converges, more and more legacy service with strict requirements for service availability are migrated to the IP/MPLS network
  - Magic number is 50ms (!)
- Consider an IP/MPLS network. How long will it take to reroute traffic?
  - Think about OSPF, IS-IS, LDP, RSVP, or BGP
  - General categories: transit failure (link/node), head /tail end, edge failure



Global vs. Local Protection

Global Protection

- P router adjacent to (egress) PE detects PE failure, and advertises it into IGP
- IGP is used to propagate failure notification to other (ingress) PEs
  - Using OSPF/ISIS flooding procedures, connectivity recovery depends on propagating failure notification
  - Connectivity recovery time can not be less than the time it takes to propagate and process failure notification in ISIS/OSPF
  - Other (ingress) PEs adjust their forwarding tables, once they receive the failure notification via ISIS/OSPF
  - Propagation time involves control plane processing delay on all the intermediate nodes
- Requires signaling to take place, i.e.. restoration time: several 100s of msec

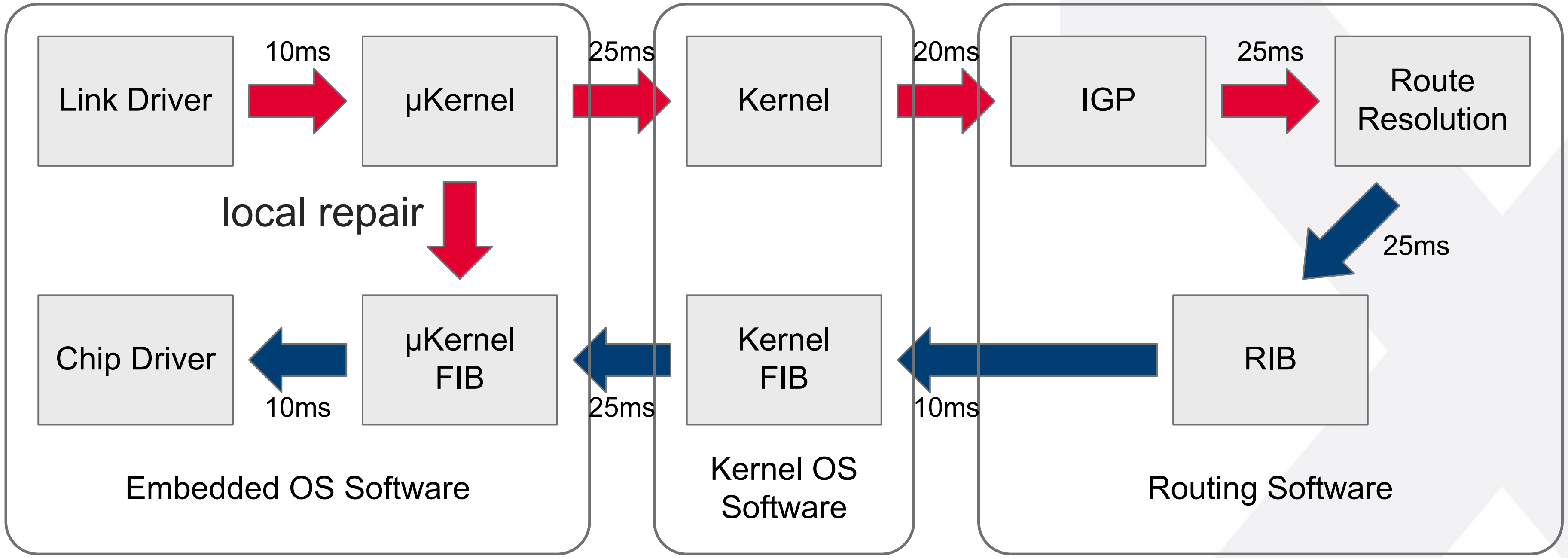
Local Protection

- P router adjacent to (egress) PE detects PE failure
- P router adjacent to PE adjusts its forwarding table
  - P router becomes Point of Local Repair (PLR)
  - Connectivity recovery does not depend on propagating failure notification in ISIS/OSPF
  - Connectivity recovery time does not depend on ISIS/OSPF propagating and processing failure notification all the way to the ingress PEs
  - Connectivity recovery time can be comparable to the time it takes for PLR to detect PE failure
- Based on pre-computed local backup, i.e. restoration in sub-50 msec

Local protection is the fastest and the most scalable way to provide connectivity recovery!

Router event propagation

- Example: Link down event



Agenda

- Problem Definition
- MPLS Fast Reroute
- Loop-Free Alternates
- Segment Routing

MPLS Fast Reroute (FRR) – RFC 4090

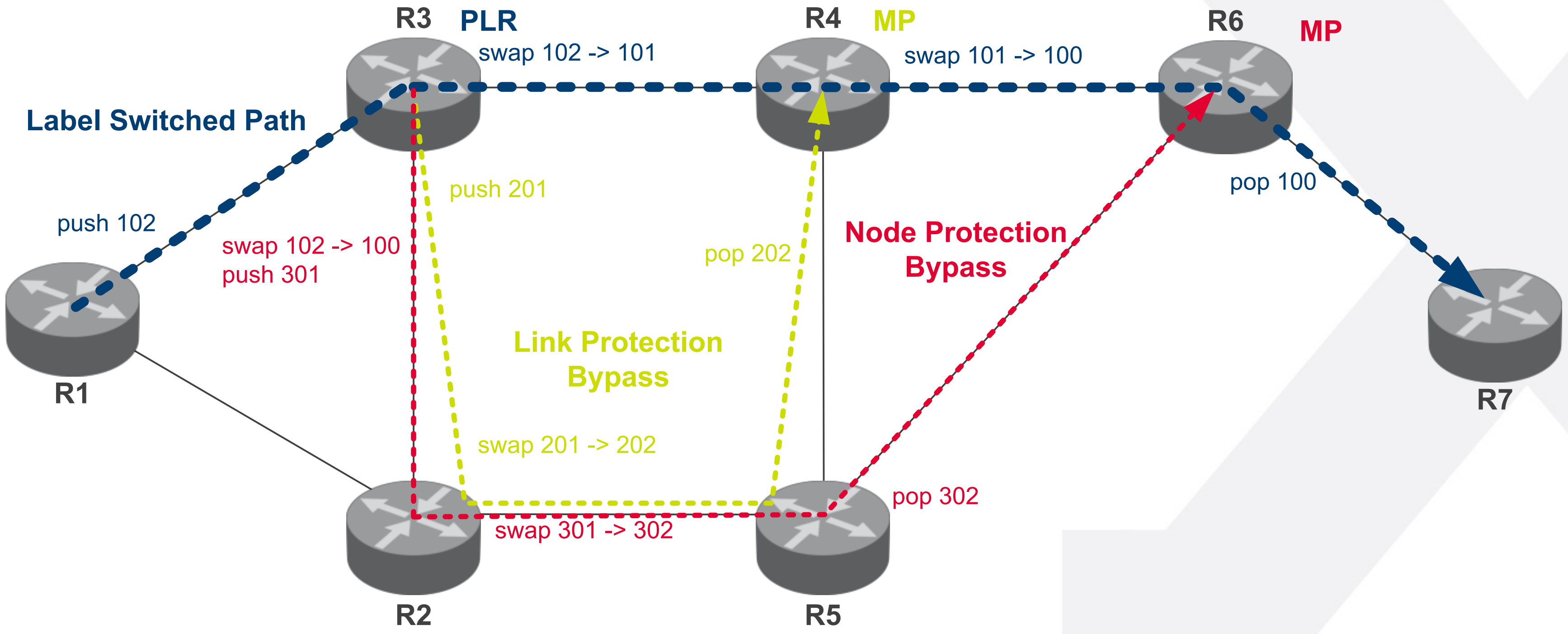
- Let‘s have a look at an RSVP-signaled MPLS LSP. Consider a failure somewhere in the network!
  - Node which discovers failure send ResvTear message towards ingress LSR
  - Ingress LSR re-computes LSP and sends Path messages towards the egress LSR
  - Ingress eventually receives Resv message and maps traffic to new LSP
- That won‘t happen in 50ms! Can we speed up the process?
- MPLS Fast Reroute (FRR) mechanism offers a short-term solution by pre-computing and pre-installing alternate path using detour/bypass LSPs at point of local repair (PLR)
  - Offers link and node protection
  - Option to have one-to-one or facility backup paths
  - Requires RSVP-TE signaling
- Introducing RSVP-signaling to get FRR increases the amount of complexity/states in the network (lot of configuration necessary, huge amount of RSVP states and signaling)

MPLS Fast Reroute (FRR) – after a failure…

- Additional control plane action:
  - Suppression of LSP teardown
    - LSP Head end receives IGP notifications about the failure of the link
    - suppress error generation that would lead to the teardown of the LSP
  - Notification of the LSP head end
    - PLR protect traffic while LSP head end looks for an alternative path
    - PLR takes care of notifying the head end using an RSVP Path Error
      - Including a “Notify” error code “Tunnel locally repaired” subcode
      - Additional flag is turned on in RRO (route record object)
  - New Path computation and signaling
    - Head end recomputes LSP, avoiding failed link
    - Set in make-before-break fashion
    - shared explicit is always used for local protection

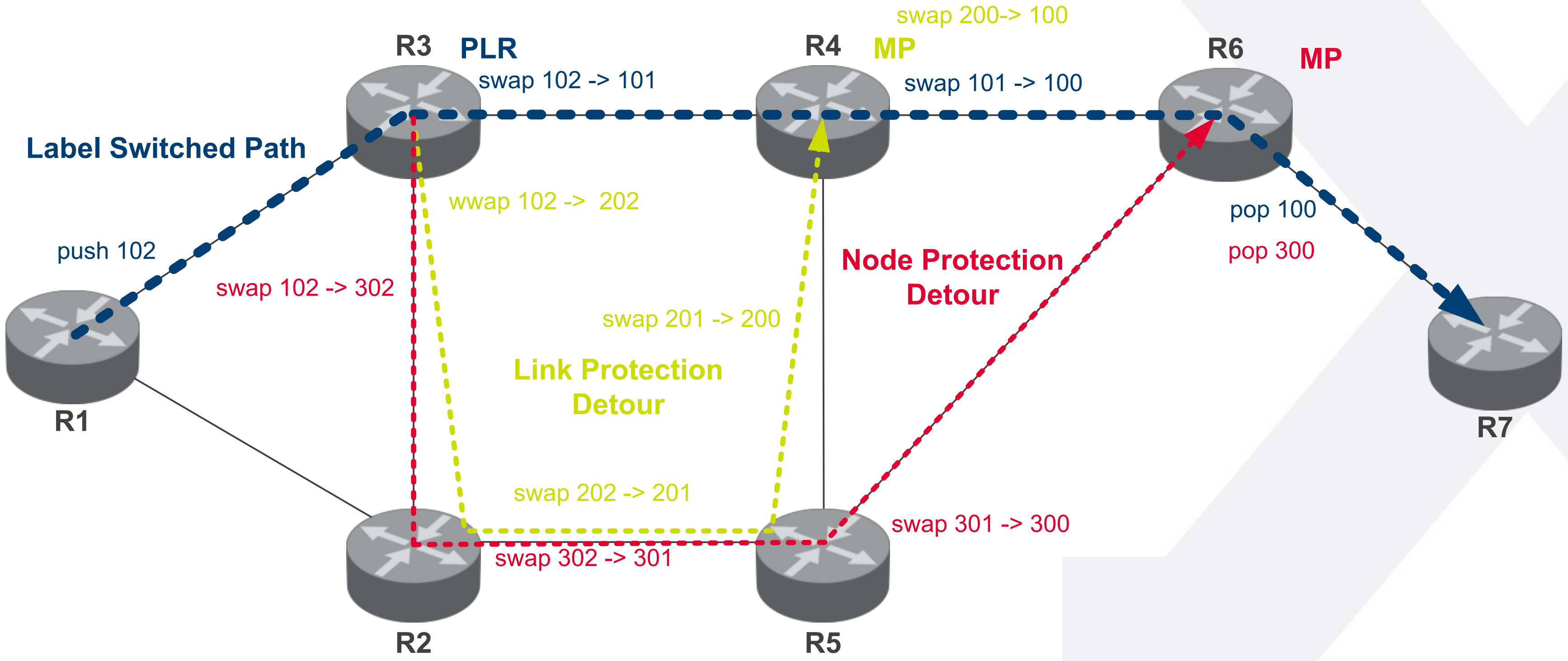
MPLS FRR Explained – Facility backup

- Concept of label stacking is used by PLR (LSP hierarchy)
  - LSR at head end of detour LSPs receives packet identical to the one it would have received on original link (note, that labels have per platform scope)
  - In case of link protection, next-hop bypass LSP will be created
  - In case of node protection, next-next-hop bypass LSP will be created



MPLS FRR Explained – one-to-one backup

- one-to-one backup requires installing new forwarding states at both PLR and MP
- amount of states increases proportionally to the number of LSPs protected
- no need to increase the label stack
- tighter control over backup tunnel and its properties



MPLS FRR One-to-One vs Multiple-to-One Backup

- One-to-one backup
  - One dedicated detour LSP protecting one LSP
  - Best suited if path selection criteria such as bandwidth, priority and link coloring are critical

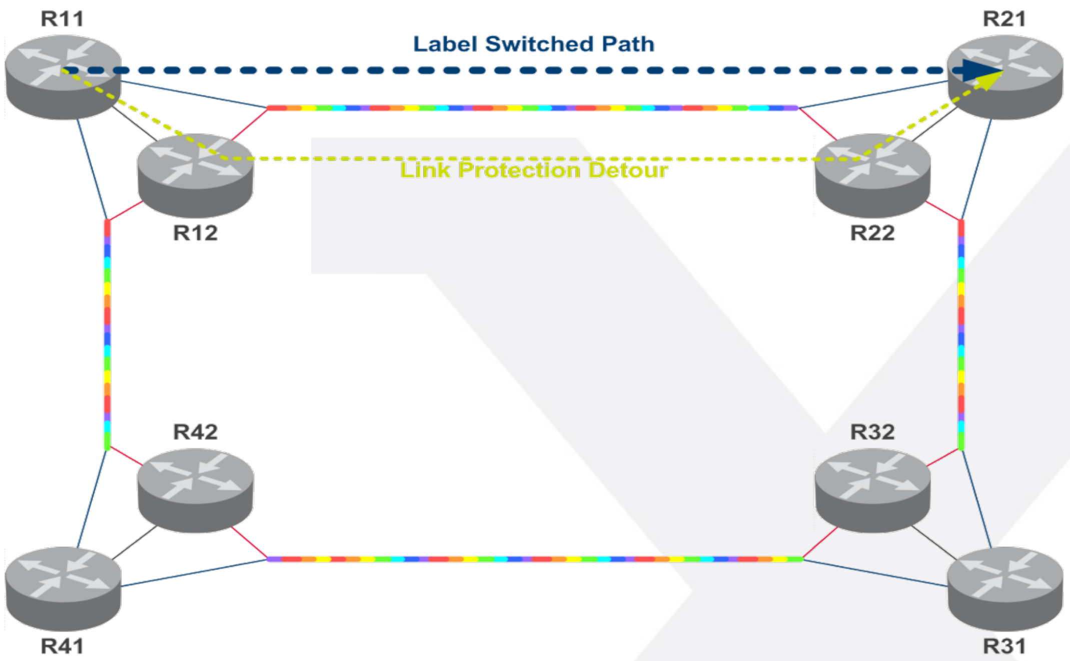
```
[edit protocols]
mpls {
    label-switched-path Example {
        …
        fast-reroute;
    }
}
```

- Multiple-to-one (facility) backup
  - One bypass LSP protecting multiple LSPs at the same time
  - Improves scalability
  - MP is nexthop node or nexthop's nexthop node

```
[edit protocols]
rsvp {
    interface xe-0/0/0 {
        link-protection;
    }
}
mpls {
    label-switched-path Example {
        …
        link-protection|node-link-protection;
    }
}
```

Fate Sharing

- What happens if primary and secondary LSPs are running across common infrastructure, e.g. DWDM equipment, switches, etc.?
- Fate sharing allows grouping of elements with common properties
  - Groups are configured with costs
  - These costs are added to CSPF metric when calculating secondary path
  - Effectively requires standby configuration




```
[edit routing-options fate-sharing]
group PoP1_to_PoP2 {
    cost 1000;
    from R11 to R21;
    from R12 to R22;
}
```

 |

Shared Link Risk Groups (SLRG)

- Fate sharing requires configuration on each individual devices
  - No protocol available exchanging database information
  - Inconsistency due to misconfiguration
- Shared Link Risk Group (SLRG) uses standard-based approach to distribute fate sharing information via IGP TE extensions
  - Requires traffic engineering
  - Support for OSPF (via RFC 4203) and IS-IS (RFC 5307)
  - Introduced in JUNOS 11.4

```
[edit routing-options]
srlg {
    PoP1_to_PoP2 {
        srlg-cost 1000;
        srlg-value 122;
    }
}

[edit protocols mpls]
interface xe-0/0/0 {
    srlg PoP1_to_PoP2;
}
```

Agenda

- Problem Definition
- MPLS Fast Reroute
- Loop-Free Alternates
- Segment Routing

Loop-Free Alternates (LFA)

- How to improve convergence without RSVP-signaling?
- Remember, IGP only calculates best path between source/destination pairs
  - No reason why equal-cost multipath (ECMP) routes cannot be used for local repair
  - No reason why less than equal cost routes canot be used as long as no forwarding loop is created!
- Loop-Free Alternates (sometimes known as IP Fast Reroute) is described in RFC 5286 , RFC 6571 and uses a simple constraint to avoid loop forwarding
  - For a local router R, a neighbor N can provide a LFA for destination X if and only if

    metric(N, X) < metric(R, X) + metric(R, N)
  - LFA is based on IGP information only and provides rerouting capabilities for native IP traffic as well as MPLS traffic (with LDP)
  - LFA is a local decision and does not require any interaction with neighboring routers
- Add a non-best path for backup purpose, but how
  - Shared, common link state database
  - Place the SPF root at your neighbors

Loop-Free Alternates Example

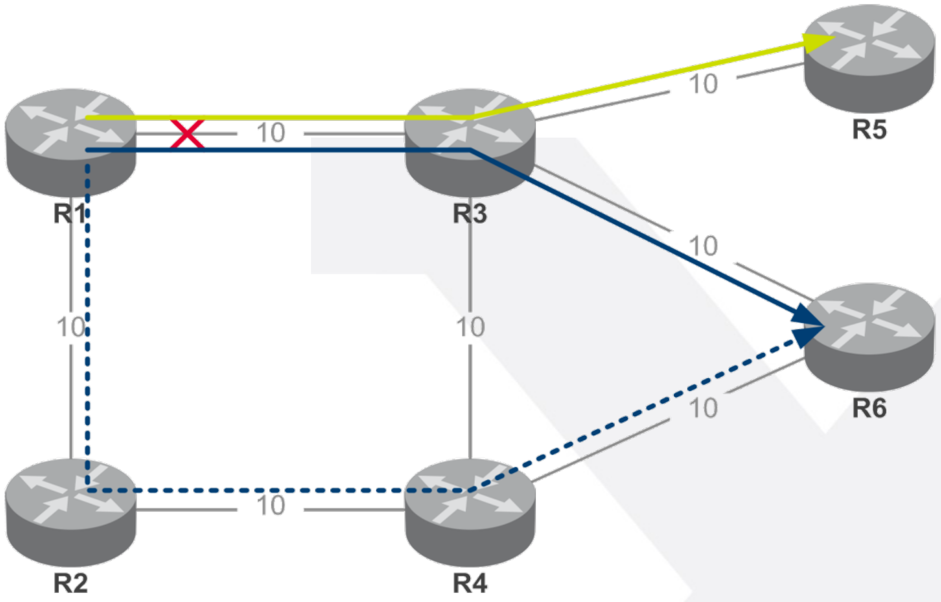
- Look at ingress router R1's routing table
  - R5 reachable via R3 with cost of 20
  - R6 reachable via R3 with cost of 20
- Consider link failure between router R1 and R3!
  - There exists an LFA for destination R6 via R2 because

    [metric(R2->R6)=20] <

    [metric(R1->R2)+metric(R1->R6)=30]
  - There is no LFA for destination R5 because

    [metric(R2->R3)=30] =

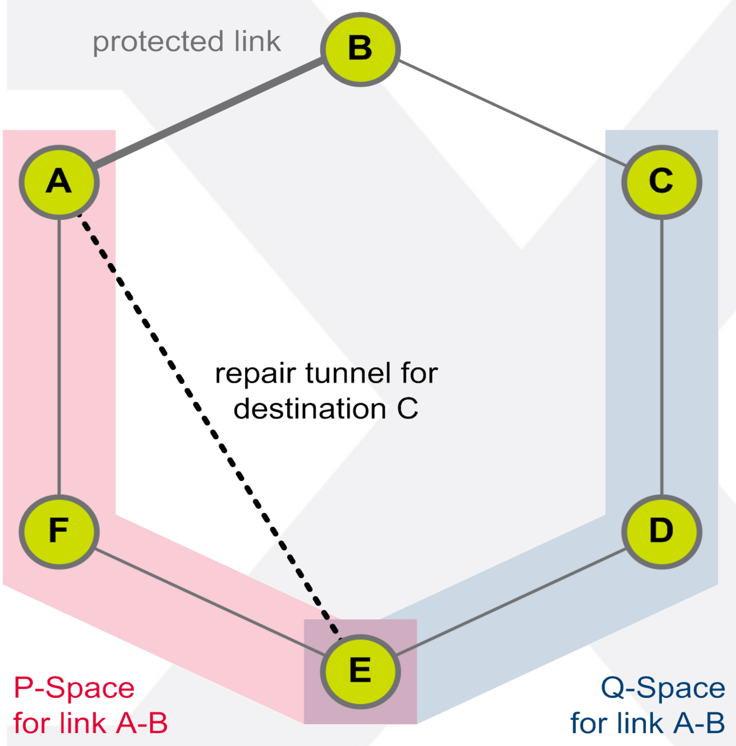
    [metric(R1->R2)+metric(R1->R5)=30]




```
[edit protocols isis]
interface all {
    link-protection|node-link-protection;
}
```

Remote LFA

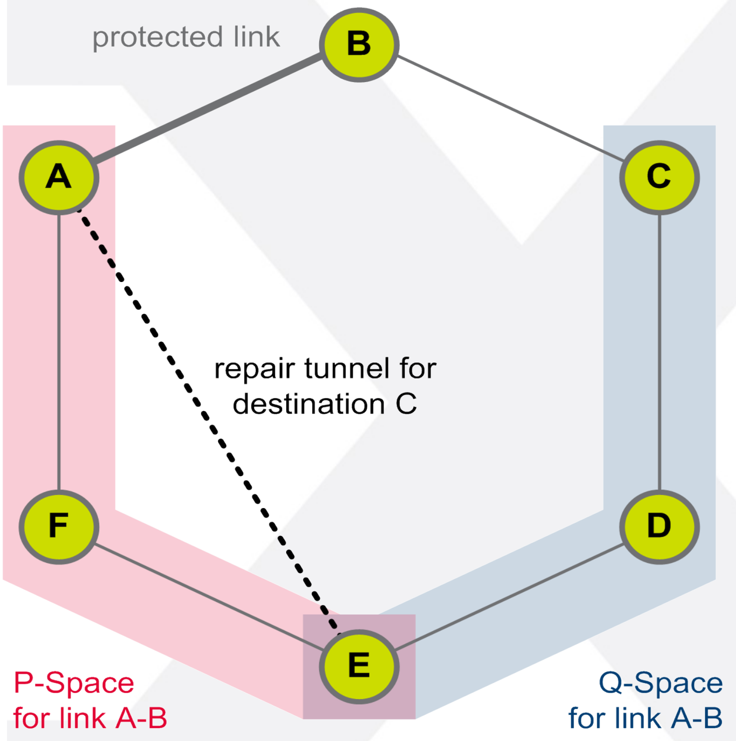
- LFA provides good repair coverage in many topologies, especially if highly meshed.
  - However, some topologies (e.g. rings) are not well protected by LFA alone
- Remote LFA (rLFA) uses tunnels to provide additional logical links used as LFAs where none exist in the original topology
  - P-space: set of all node reachable from source without traversing protected link
  - Q-space: set of all nodes which can reach destination without traversing protected link
  - Tunnel endpoint defined by intersection of P-space and Q-space
- Most be done on a per-prefix basis
- Consider traffic travelling from A to C via B
- Still no guarantee for 100% coverage
- Hard to achieve node protection



 |

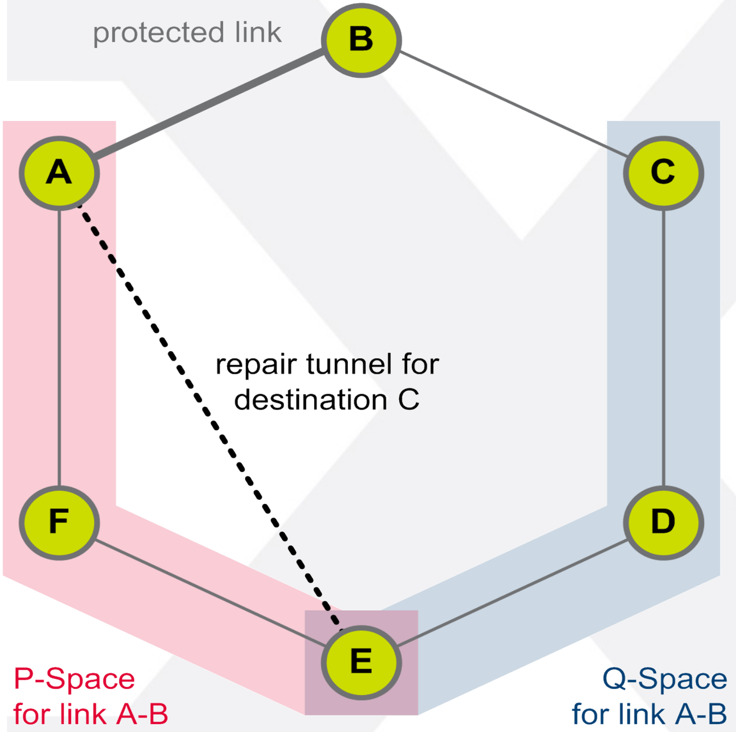
Remote LFA – Case IP Packet

- In the case of IP traffic being protected, A pushes the LDP label required to reach E on top of the IP packet.
- Using Existing LDP LSP to E
- Assuming PHP, packet arrives at E as a plain IP packet.
  E then forwards the packet to D, as this is on the best path towards the destination, C.



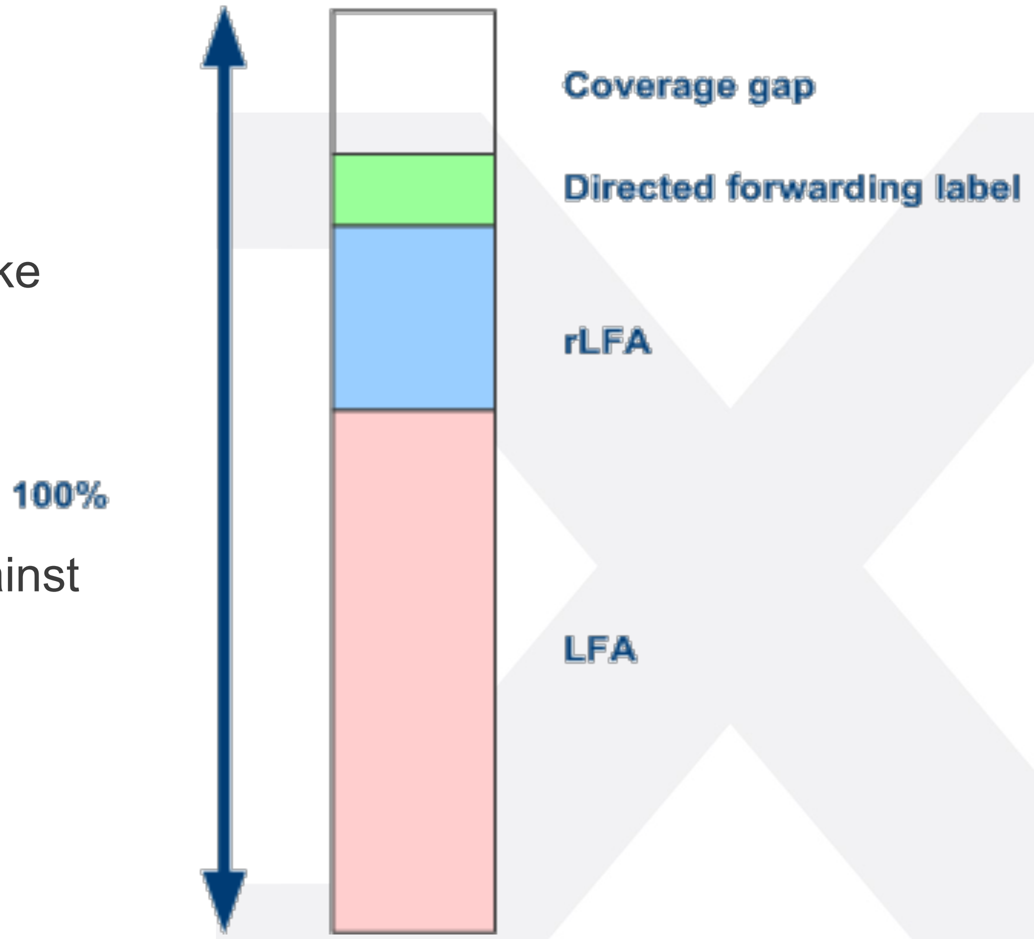
Remote LFA – Case MPLS (LDP) Packet

- In the case of LDP traffic being protected, a stack consisting of two LDP labels is used by A, i.e. “LDP over LDP”.
- The outer LDP label, is the label required to reach RE.
- The inner LDP label, is the label required to reach C from E.
- A targeted LDP session (automatically created) is needed between A and E, so that A can learn the label, advertised by E to reach C.



Difficulty of Attaining full coverage with LFA

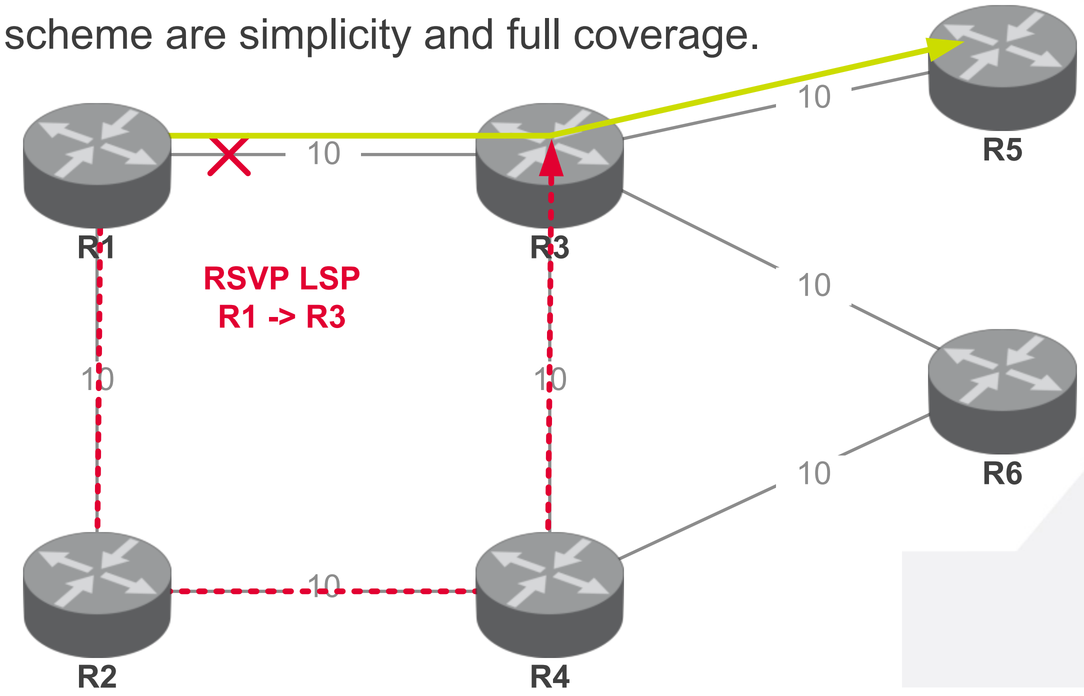
- Difficult to reach 100% coverage without caveats
- The closer we get to 100%, the more difficult is it to make further improvements
- Fundamental problem is that we are trying to „fight“ against the IGP metrics.



Coverage Extension using dynamic RSVP LSP

- An RSVP bypass LSP is automatically created
  - RSVP LSP goes all the way to the node on the far side of the protected link
  - From Egress Node of the LSP the packet then travels to ist original destination

- LFA + RSVP for full coverage
  - The advantages of the scheme are simplicity and full coverage.




```
[edit protocols ldp]
interface all {
    link-protection {
        dynamic-rsvp-lsp;
    }
}
```

Agenda

- Problem Definition
- MPLS Fast Reroute
- Loop-Free Alternates
- Segment Routing

 |

Segment Routing Goals

- Simplicity and Scale
  - Operators want to reduce number of protocols use to simplify network architecture and ease troubleshooting
  - Need to have Fast Reroute capability for any topology without explicit configuration of thousands of RSVP tunnels
  - Leverage all existing services supported over MPLS networks today
    - Source routing, Fast Reroute, VPNv4/6, VPLS, L2VPN
  - Avoid #millions of labels, tunnels and TE LSPs
- Application Centric Networking
  - Allow Applications to influence forwarding decisions in a scalable way
  - Provide programmatic interfaces and orchestration
- Two main concepts
  - put label advertisement into IGP
  - Forwarding based on a label stack

Segment Routing Overview – focus on MPLS dataplane

- New approach standardized in the IETF
  - draft-filsfils-spring-segment-routing-04
  - draft-filsfils-spring-segment-routing-use-cases-01
  - draft-filsfils-spring-segment-routing-mpls-03
- Forwarding state (aka segment) is established by IGP (either OSPF or IS-IS)
  - No need to run LDP or RSVP-TE as a control plane protocol
  - Existing MPLS data plane remains without any modification RFC 3031
    - push, swap and pop
    - segment = label
- A segment identifies respective can represent any instruction
  - Service
  - Context
  - Locator
  - IGP-based forwarding construct
  - BGP-based forwarding construct
  - Local value or Global Index
- the source chooses a path and encodes it in the packet header as an ordered list of segments.
- Per flow state only at ingress SR Domain edge node
  - Ingress edge node pushes the segment list on the packet

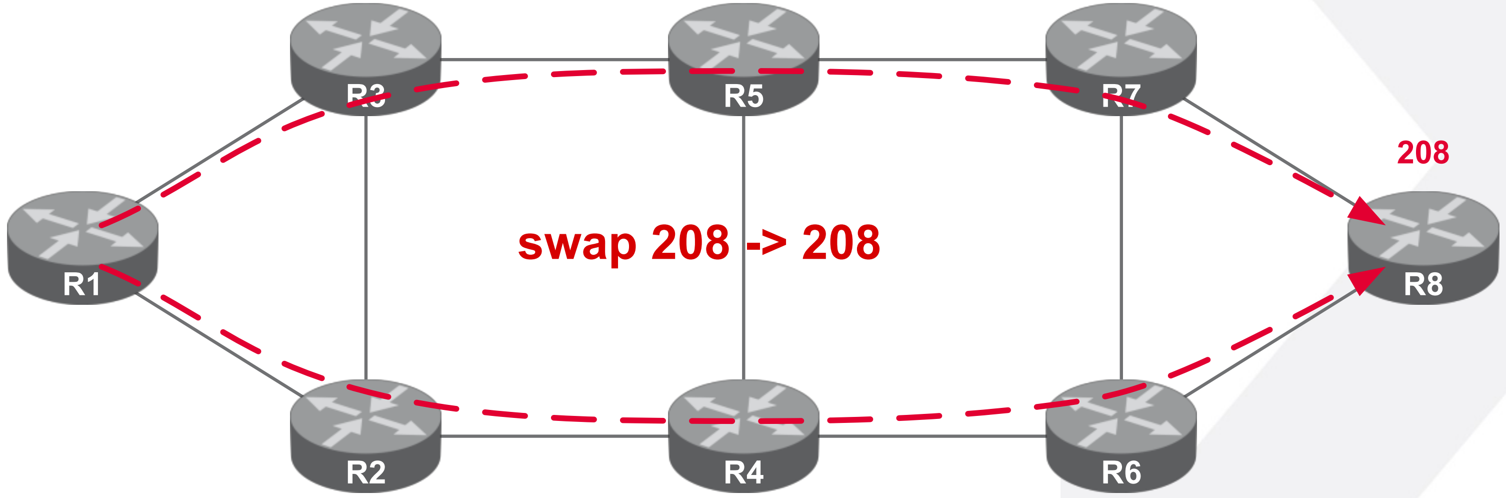
Segment = Instruction
„use shortest path to reach R8“

IGP Segment IDENTIFIERS.

- Node SID
  - prefix that identifies a specific node (e.g. the prefix is its loopback)
  - R1 lo0 1.2.3.4/32 (Node-SID: 201)
  - Global Label Allocation indicating SPT to advertising Node (special Prefix SID)
- Adjacency SID
  - Local Label Allocation indicating a link (or set of links) within the IGP topology
  - Local segment related to a specific SR node
- Prefix SID
  - Local Label Allocation indicating a IGP “leaf” IP prefix (attached node)
  - 2.2.2.2/32 (Prefix---SID: 2222)
- SR Global Block
  - SRGB is the set of local labels reserved for global segments
  - All the global segments must be allocated from SRGB
  - Operator manages SRGB like an IP address block: it ensures unique allocation of a global segment within the SR domain

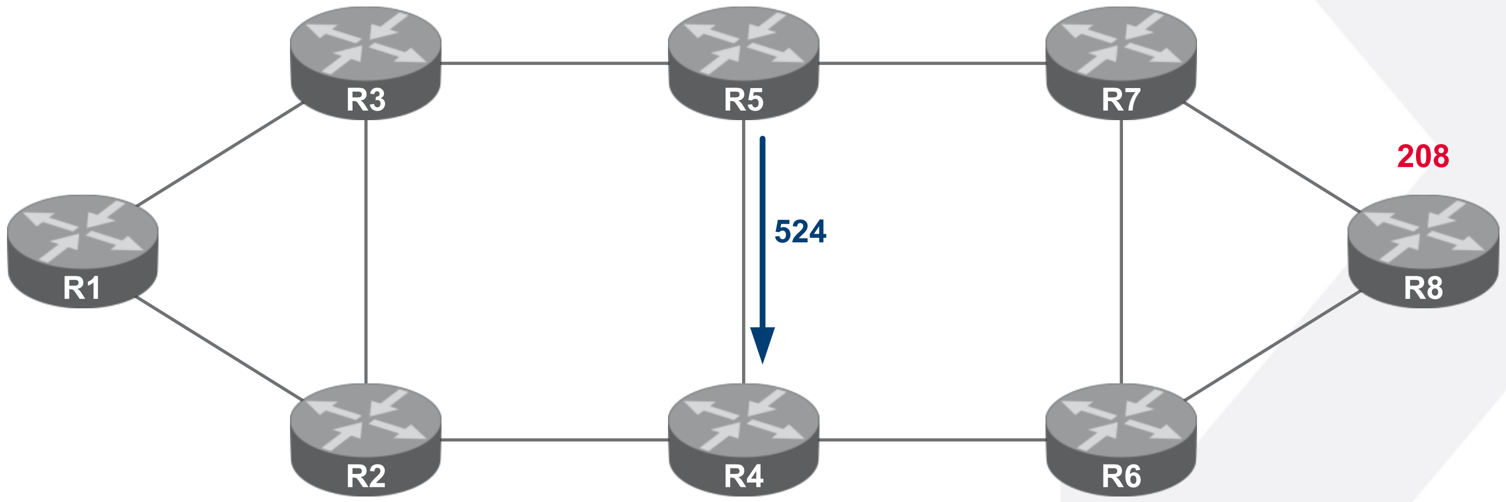
IGP Prefix or Node Segment

- Each router gets one unique label from SR range, router R8 gets **208**
- Router R8 advertises its global prefix segment **208** with his loopback address
  - ISIS sub-TLV extension (draft-previdi-isis-segment-routing-extensions-01)
  - OSPF opaque sub-TLV extension (draft-psenak-ospf-segment-routing-extensions-01)
- All remote routers install the prefix segment to R8 in the MPLS data plane along the shortest path to R8/32
  - Packet injected anywhere with active segment 208 will reach router R8 via ECMP-aware shortest path



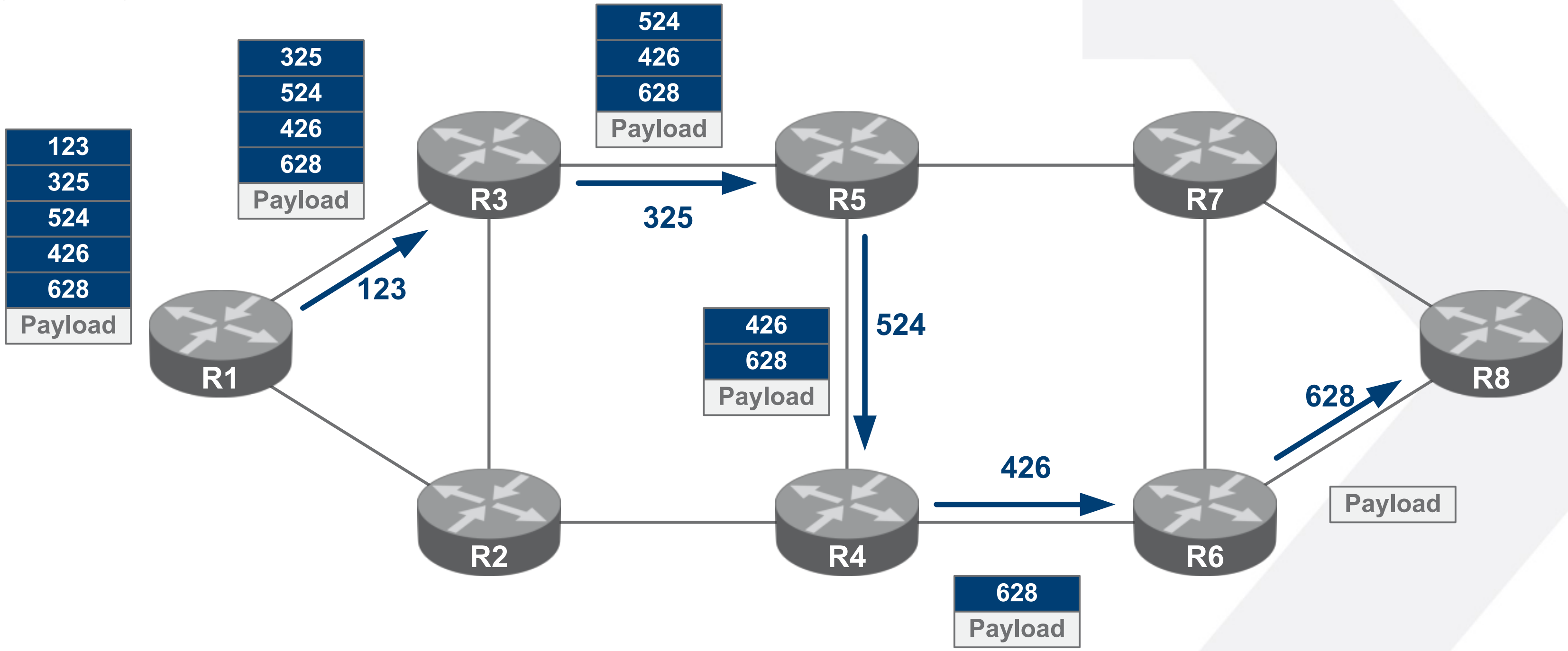
IGP Adjacency Segment

- Router R5 allocates a local segment **524** for its adjacency R5-R4 and advertises the segment in IGP
  - ISIS sub-TLV extension (draft-previdi-isis-segment-routing-extensions-01)
  - OSPF opaque sub-TLV extension (draft-psenak-ospf-segment-routing-extensions-01)
- R5 is the only node to install the adjacency segment in the MPLS dataplane
  - Packet injected at node R5 with active segment 524 is forced through link R5->R4



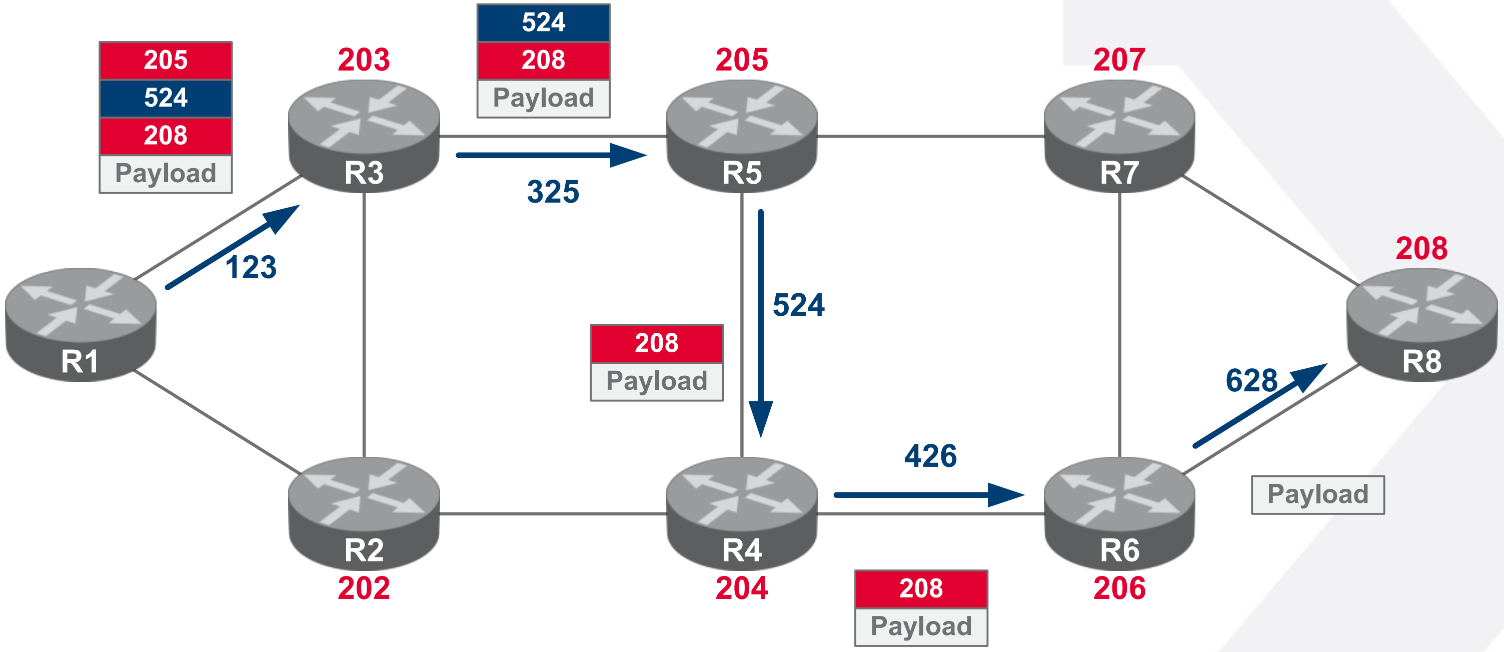
Example: Explicit Path using Adjacency Segments

- Segment Routing provides path control for the entire label switched path
- Source routing along any explicit path
  - stack of adjacency labels



Example: Explicit Path Combining Node & Adj Segments

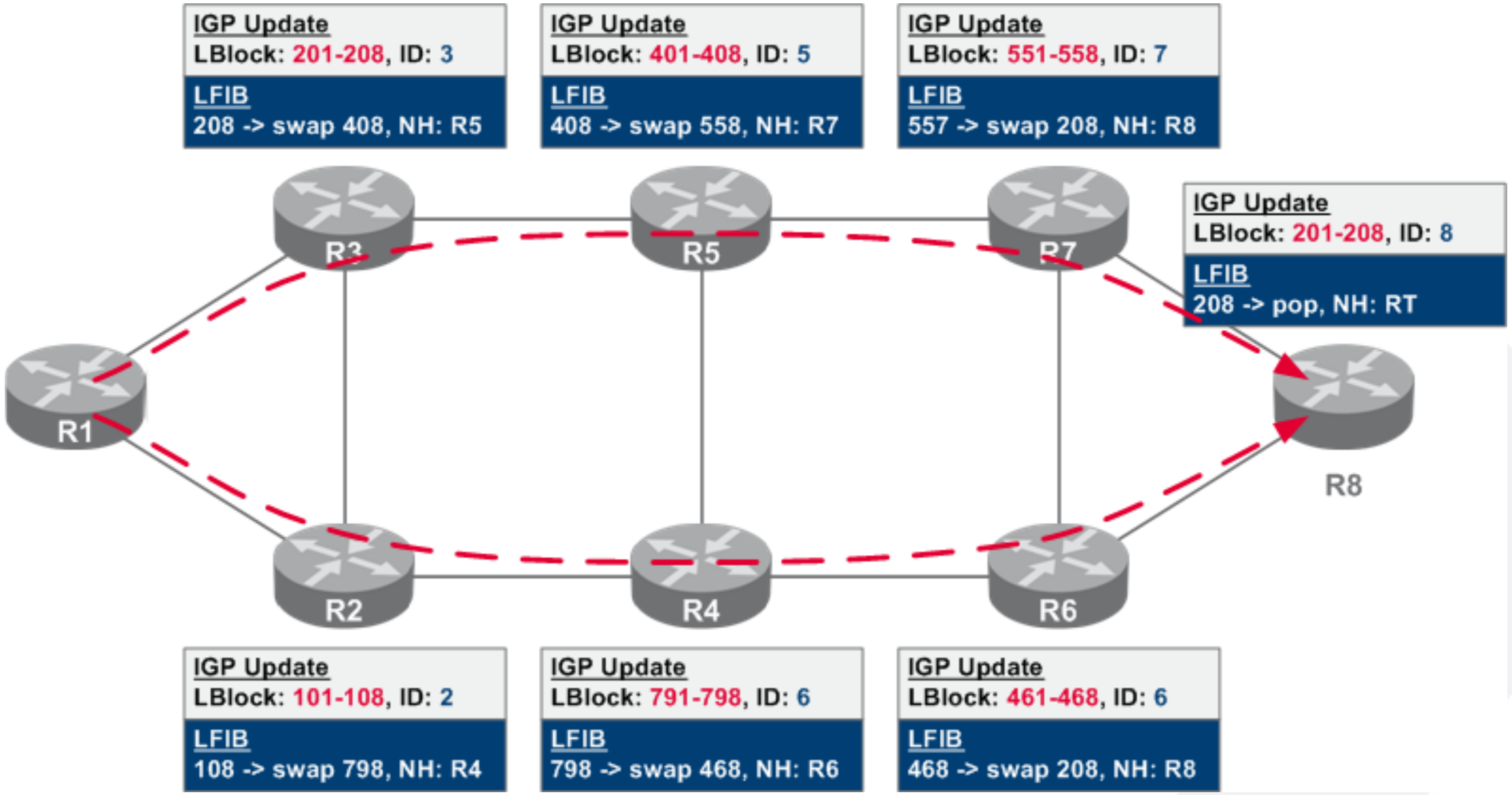
- Any path can be expressed using a combination of IGP prefix (node) segments and adjacency segments
- Excellent Scale: a node installs N+A FIB entries (N node segments and A adjacency segments)



Domain-Wide Labels

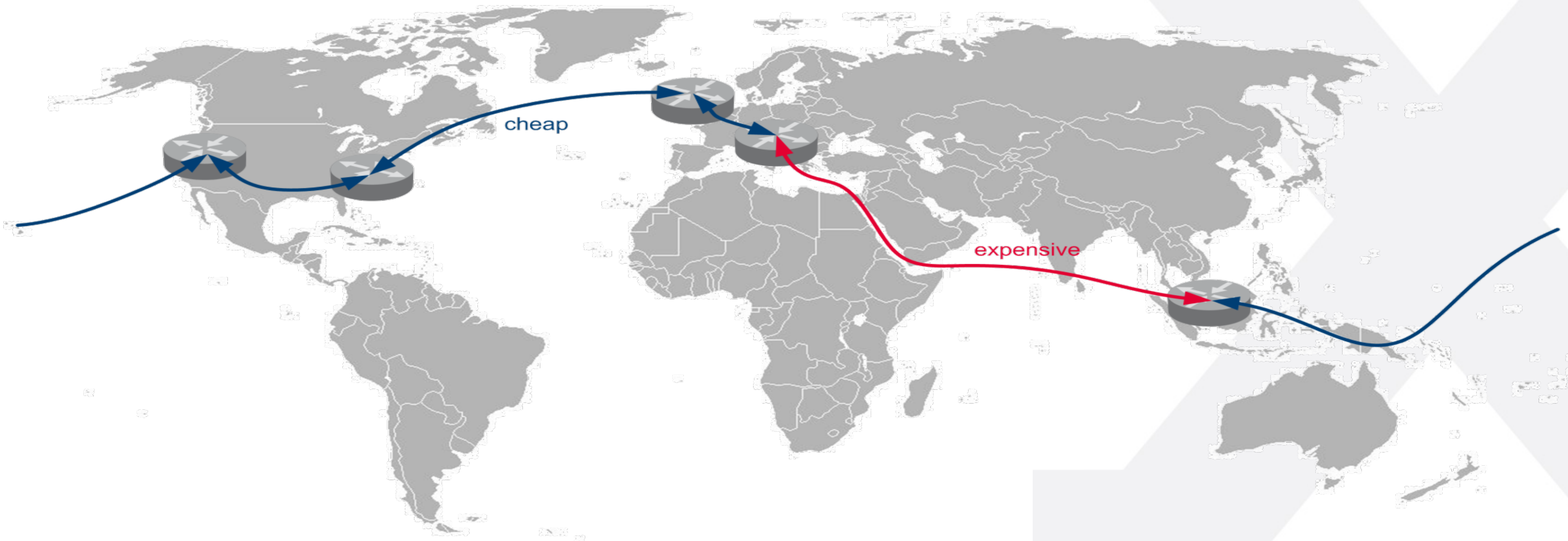
- Segment routing draft requires IGP prefix segments to be globally unique (at least for the node segments)
  - According to RFC5031, MPLS labels only have local significance
  - Introduction of global labels requires significant change of MPLS architecture
    - Today identical labels can co-exist in routing domain
    - Most devices have configurable label-ranges per protocol
    - Interoperability with routers which do not support segment routing
    - Re-use label block semantic used in BGP-based VPLS (RFC 4761)
  - Assign each router a domain-wide unique ID
  - ID is an index to locate the actual label value, inside label block
  - Each LSR allocates and advertises a block of locally significant labels
  - The block should be large enough to accommodate the range of assigned IDs

Label Range Index and LFIB Construction



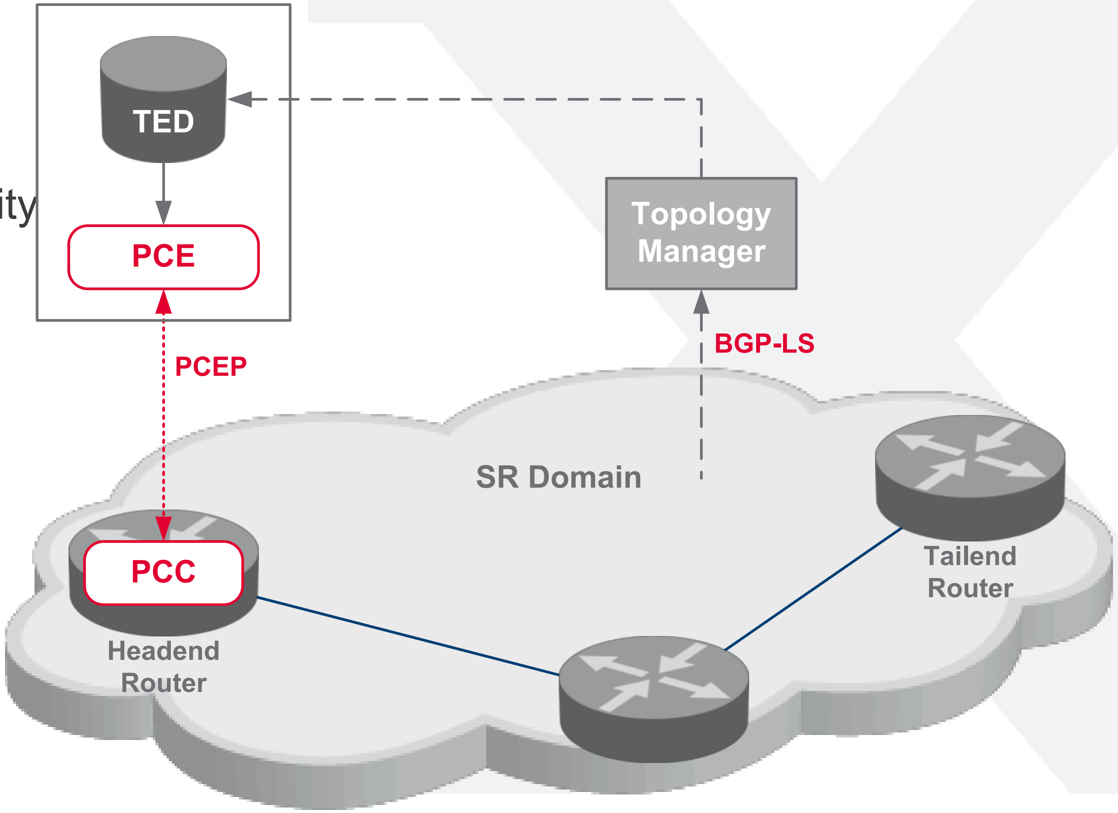
Segment Routing Use Case: CoS-based Routing

- Routing based on service requirements
  - Use direct Asia-Europe path with low latency (expensive)
  - Use path via America with higher latency but reduced costs



Use Case SDN: Self Destructive Networks ;-)

- The heart of the application of SR to the SDN use-case lies in the SDN controller, also called Stateful PCE
- The controller abstracts the network topology and traffic matrix (BGP-LS)
- An SDN Controller (SC) is connected to the network and is able to retrieve the topology and traffic information, as well as set traffic-engineering policies on the network nodes.
- Controller-based Computation
  - Support any other constraint: latency, disjointness
  - SDN-centric: application-based network programmability



Questions?

- Life, the universe, and everything
- Further Reading:
  - http://www.segment-routing.net/
  - http://www.juniper.net/us/en/homepage-campaign.page
  - https://ripe66.ripe.net/presentations/232-SR_RIPE_v2.pdf
  - https://www.ietf.org/