Segment Routing

Clarence Filsfils – cf@cisco.com
Distinguished Engineer
Christian Martin – martincj@cisco.com
Sr. Directior, Engineering

version modified & updated by dlarra for training purposes

Goals and Requirements

- Make things easier for operators
  - –Improve scale, simplify operations
  - –Minimize introduction complexity/disruption
- Enhance service offering potential through programmability
- Leverage the efficient MPLS dataplane that we have today
  - –Push, swap, pop
  - –Maintain existing label structure
- Leverage all the services supported over MPLS
  - –Explicit routing, FRR, VPNv4/6, VPLS, L2VPN, etc
- IPv6 dataplane a must, and should share parity with MPLS

Operators Ask For Drastic LDP/RSVP Improvement

- Simplicity
  - fewer protocols to operate
  - fewer protocol interactions to troubleshoot
  - avoid directed LDP sessions between core routers
  - deliver automated FRR for any topology
- Scale
  - avoid millions of labels in LDP database
  - avoid millions of TE LSP’s inside the network
  - avoid millions of tunnels to configure

Segment Routing

- Source Routing
  - the topological and service (NFV) path is encoded in packet header
  - SDN and virtualization enabled
- Scalability
  - the network fabric does not hold any per-flow state for TE or NFV
- Simplicity
  - automation: TILFA sub-50msec FRR
  - protocol elimination: LDP, RSVP-TE, VxLAN, NSH, GTP, ...
  - straightforward ISIS/OSPF extension to distribute labels
  - leverage MPLS services & hardware
- End-to-End
  - DC, Metro, WAN

Internet Engineering Task Force (IETF) Request for Comments: 8402
Category: Standards Track
July 2018

C. Filsfils, Ed., S. Previdi, Ed. L. Ginsberg ISSN: 2070-1721 Cisco Systems, Inc. B. Decraene S. Litkowski Orange R. Shakir Google, Inc.

Segment Routing Architecture

Abstract

Segment Routing (SR) leverages the source routing paradigm. A node steers a packet through an ordered list of instructions, called "segments". A segment can represent any instruction, topological or service based. A segment can have a semantic local to an SR node or global within an SR domain. SR provides a mechanism that allows a flow to be restricted to a specific topological path, while maintaining per-flow state only at the ingress node(s) to the SR domain. SR can be directly applied to the MPLS architecture with no change to the forwarding plane. A segment is encoded as an MPLS label. An ordered list of segments is encoded as a stack of labels. The segment to process is on the top of the stack. Upon completion of a segment, the related label is popped from the stack. SR can be applied to the IPv6 architecture, with a new type of routing header. A segment is encoded as an IPv6 address. An ordered list of segments is encoded as an ordered list of IPv6 addresses in the routing header. The active segment is indicated by the Destination Address (DA) of the packet. The next active segment is indicated by a pointer in the new routing header.

Concepts

- SR is a form of Source Routing
  - the source chooses a path and encodes it in the packet header as an ordered list of segments
  - the rest of the network executes the encoded instructions
- Segment: an identifier for any type of instruction
  - forwarding or service
  - Here: IGP-based forwarding construct
- Forwarding state (segment) is established by IGP
  - LDP and RSVP-TE are not required in MPLS
  - Agnostic to forwarding dataplane: IPv6 or MPLS

Segment Routing – Forwarding Planes

- MPLS: an ordered list of segments is represented as a stack of labels
  - 1 segment = 1 label
  - sequence of segments = stack of labels
  - leverages MPLS data plane

    No modification required: push, swap and pop

- IPv6: an ordered list of segments encoded in routing extension header
  - 1 segment = 1 IPv6 address
  - sequence of segments = an address list in the SRH
  - leverages RFC8200 provision for source routing extension header

Global and Local Segments

- **Local Segment**
  - Only originating node understands associated instruction
  - MPLS: locally allocated label
- **Global Segment**
  - Any node in SR domain understands associated instruction
  - Each node in SR domain installs the associated instruction in its forwarding table
  - In MPLS:
    - global label value in Segment Routing Global Block (SRGB): 16000-23999 (16000+index)

      Index must be unique in Segment Routing Domain

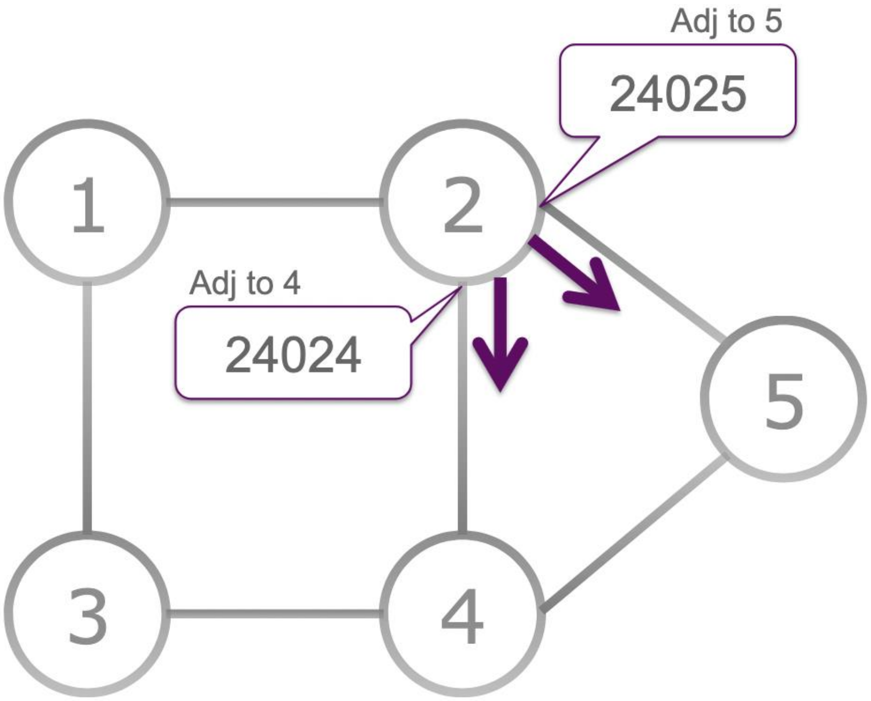
IGP segments

Segments distributed through IGP

- Local: **Adjacency segment**
- Global: **Prefix segments**
  - >Usually prefix = loopback
  - > Global allocation of a segment ID per node
  - > # nodes << SRGB range

IGP Adjacency Segment

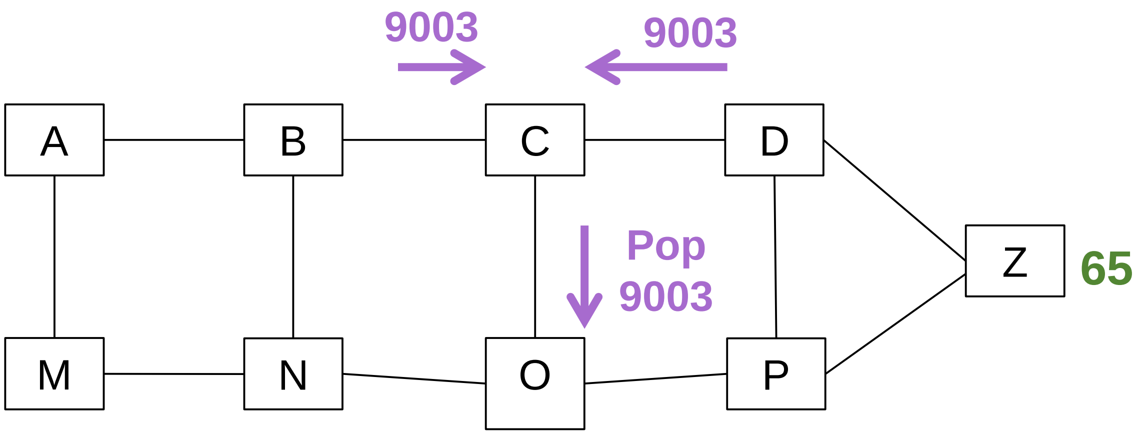
- Forward on the IGP adjacency
- Local Segment
- Advertised as label value
- Distributed by ISIS/OSPF



All nodes use default SRGB
16,000 – 23,999

Adjacency Segment

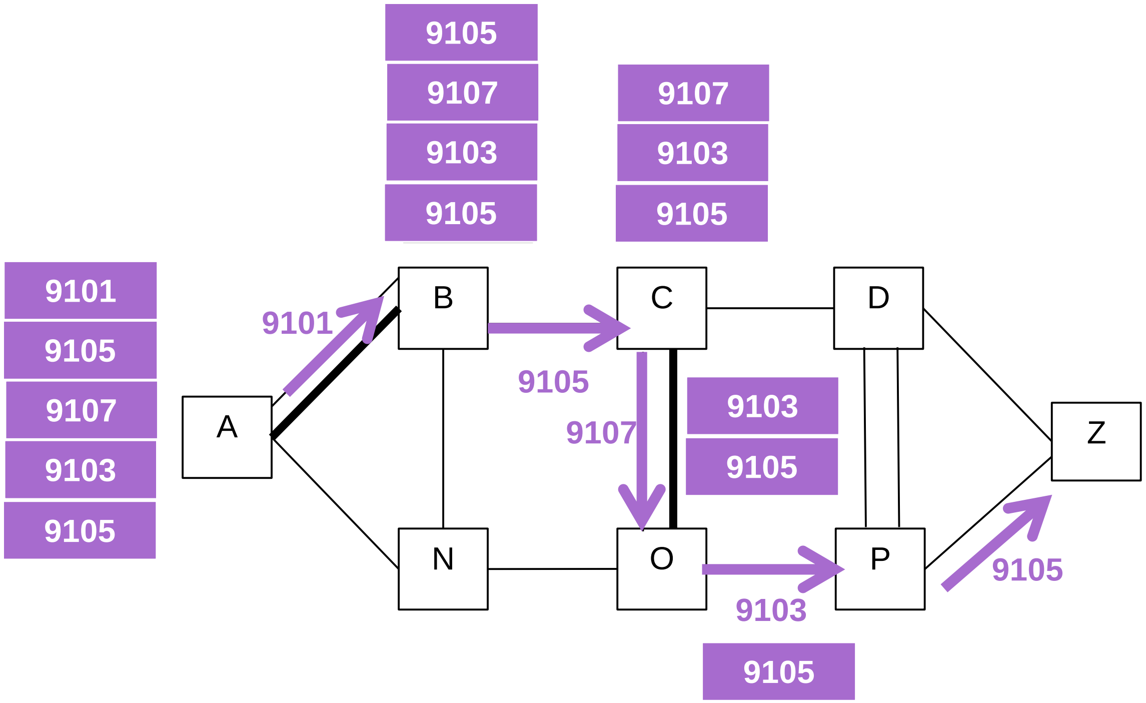
9003 = C’s Adjacency to O



A packet injected at node C with label 9003 is forced through datalink CO

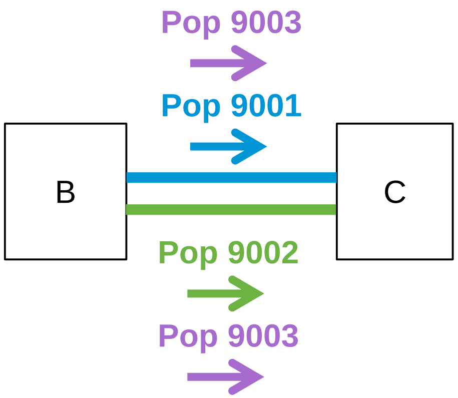
- C allocates a local label
- C advertises the adjacency label in ISIS
  – simple sub-TLV extension
- C is the only node to install the adjacency segment in MPLS dataplane

A path with Adjacency Segments



- Source routing along any explicit path
  - stack of adjacency labels
- SR provides for entire path control

Datalink and Bundle



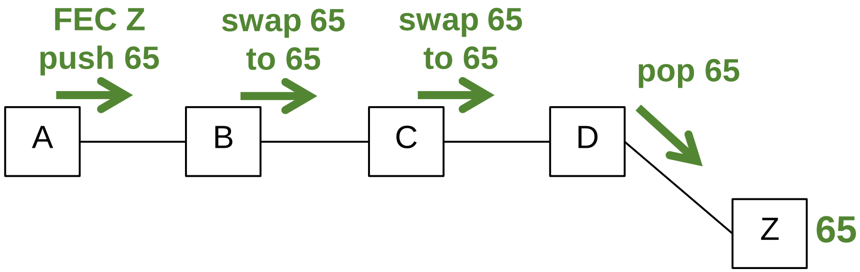
9001 switches on blue member

9002 switches on green member

9003 load-balances on any member of the adj

- Adjacency segment represents a specific datalink to an adjacent node
- Adjacency segment represents a set of datalinks to the adjacent node

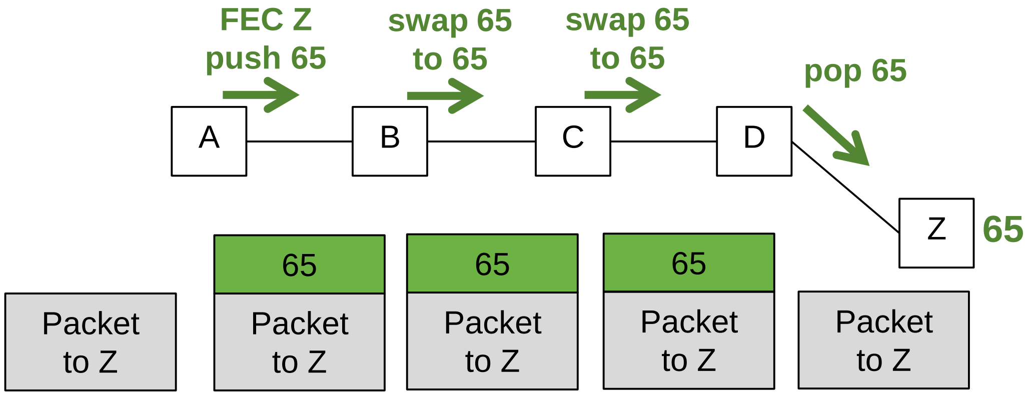
Prefix Segment



A packet injected anywhere with top label 65 will reach Z via shortest-path

- Z advertises its prefix segment: 65 over ISIS or OSPF
  - simple ISIS sub-TLV extension
- All remote nodes install the node segment to Z in the MPLS dataplane
  - # fwding entries proportional to N not to N^2

Prefix Segment

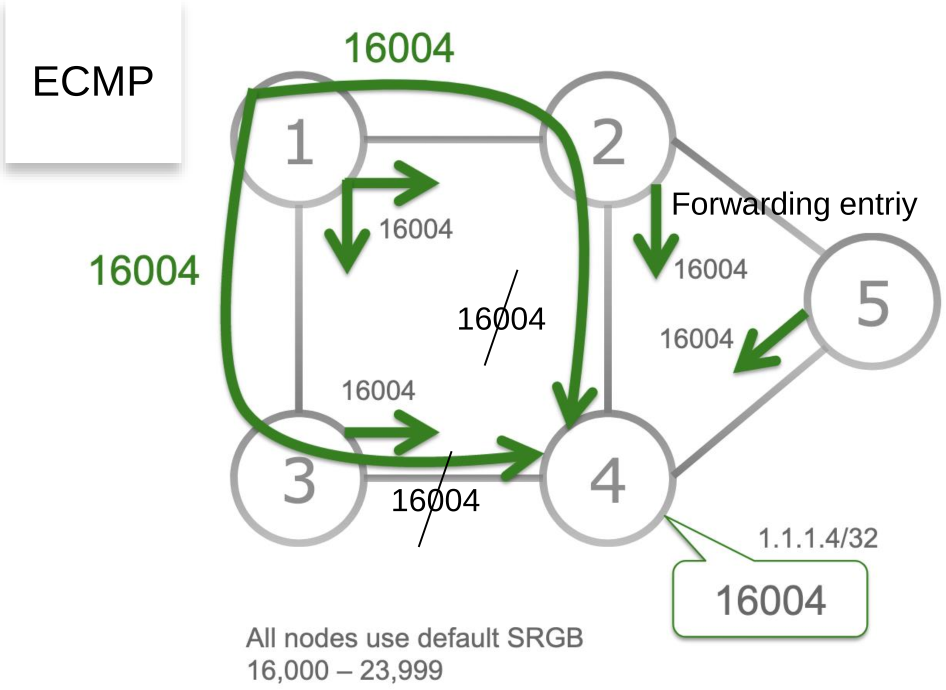


A packet injected anywhere with top label 65 will reach Z via shortest-path

- Z advertises its node segment: 65
  - simple ISIS sub-TLV extension
- All remote nodes install the node segment to Z in the MPLS dataplane
- Shortest path to the IGP prefix. ECMP aware.

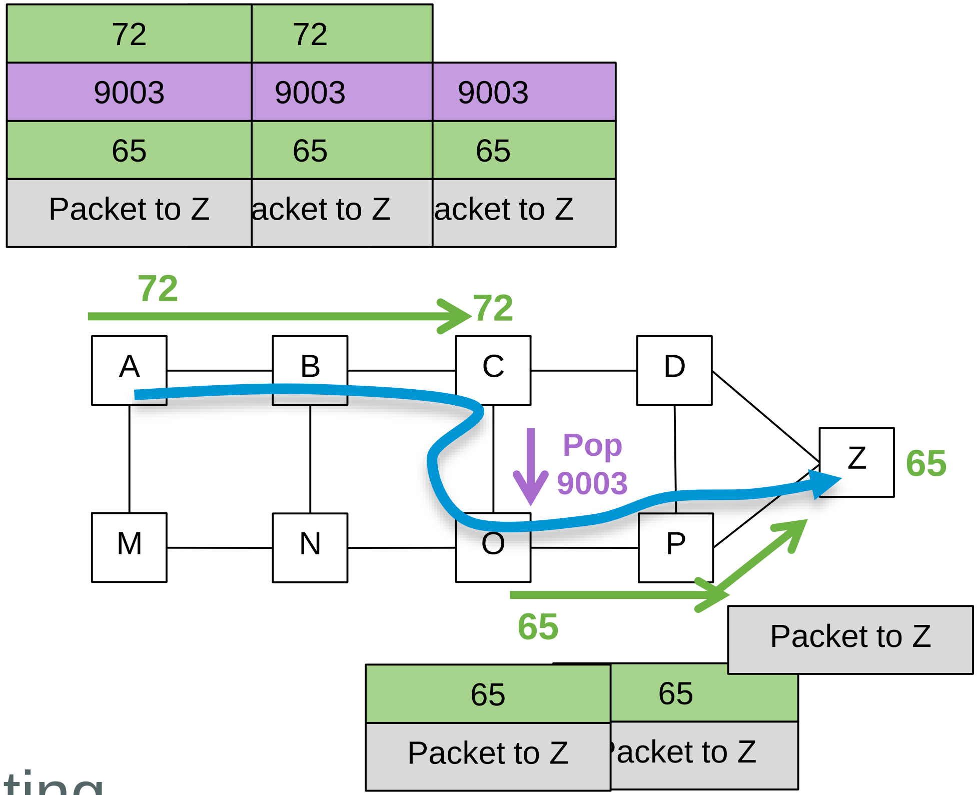
IGP Prefix Segment: ECMP

- Shortest-path to the IGP prefix
  - –Equal Cost MultiPath (ECMP)-aware
- Global Segment
  - –Label = 16000 + Index
    - >Advertised as index
  - –Distributed by ISIS/OSPF



All nodes use default SRGB
16,000 – 23,999

Combining Segments

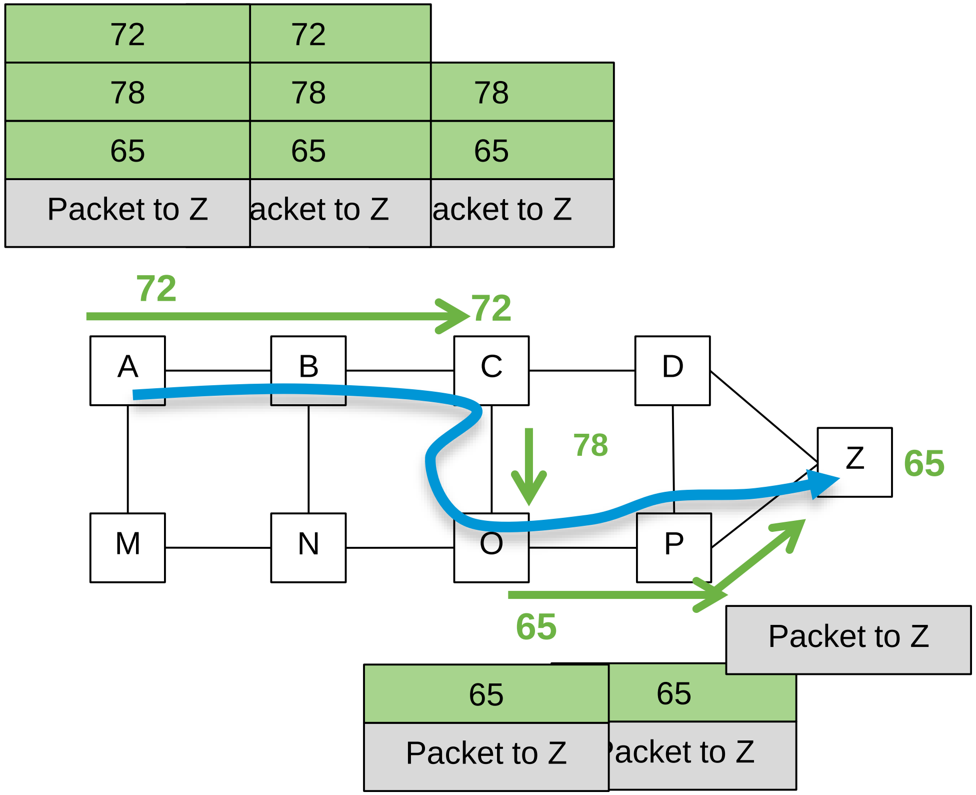


72	72	
9003	9003	9003
65	65	65
Packet to Z	acket to Z	acket to Z

		Packet to Z
65	65	
Packet to Z	'acket to Z	

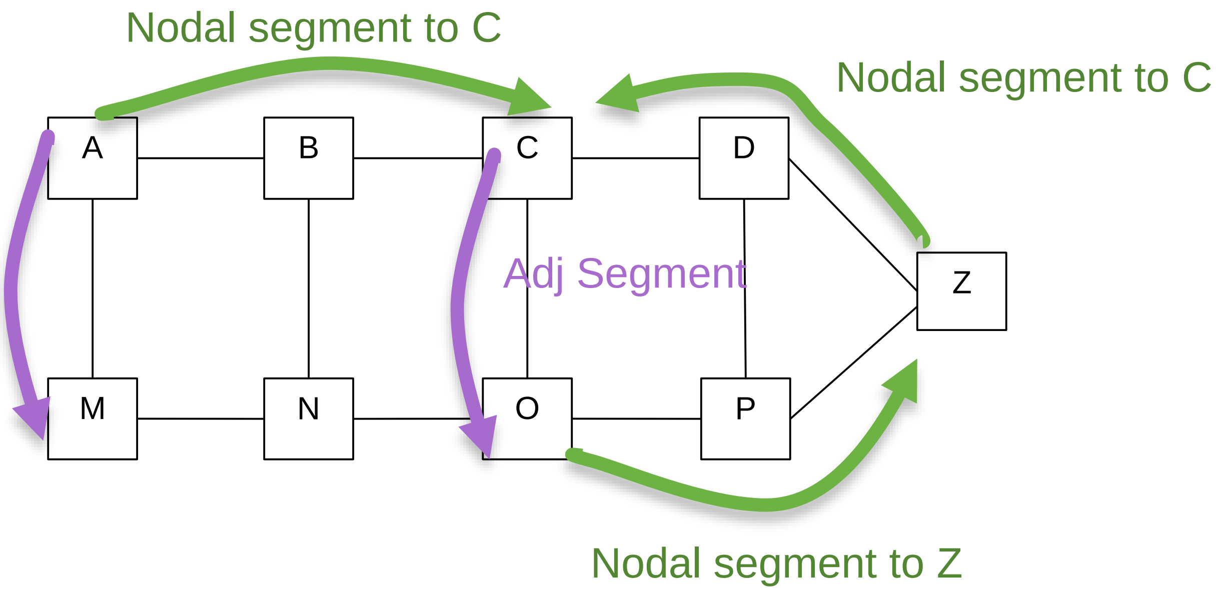
- Source Routing
- Any explicit path can be expressed: ABCOPZ

Combining Segments



- Node Segment is at the heart of the proposal
  - – ecmp multi-hop shortest-path
  - – in most topologies, any path can be expressed as list of node segments

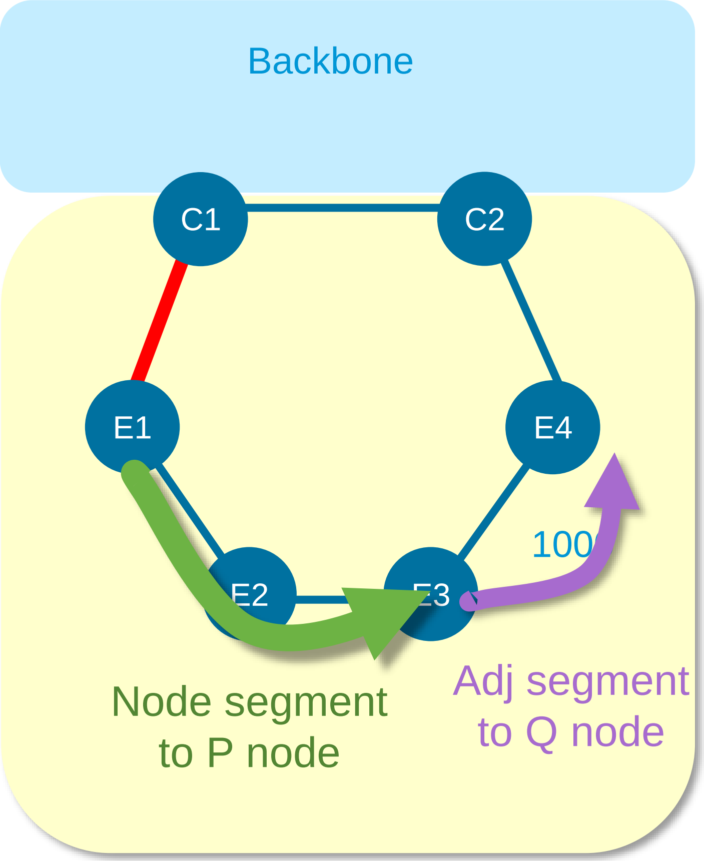
ISIS automatically installs segments



- Simple extension
- Excellent Scale: a node installs N+A FIB entries
  - N node segments and A adjacency segments
- No path is signaled • No per-flow state is created inside

Automated & Guaranteed FRR

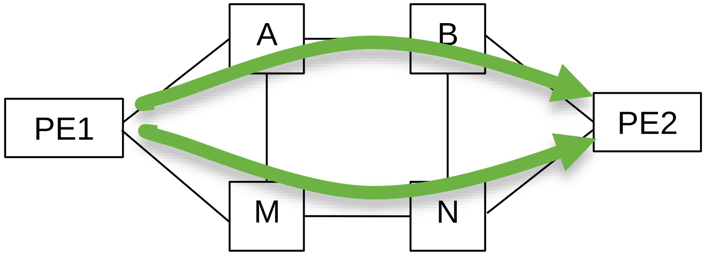
- IP-based FRR is guaranteed in any topology
  - RFC5714 Cisco 2010 informational
- Directed Loop-Free Alternate (DLFA) is guaranteed when metrics are symmetric
- No extra computation
  - No Remote LFA (RLFA) computation
- Simple repair stack
  - node segment to P node
  - adjacency segment from P to Q



Default metric: 10

Use Cases

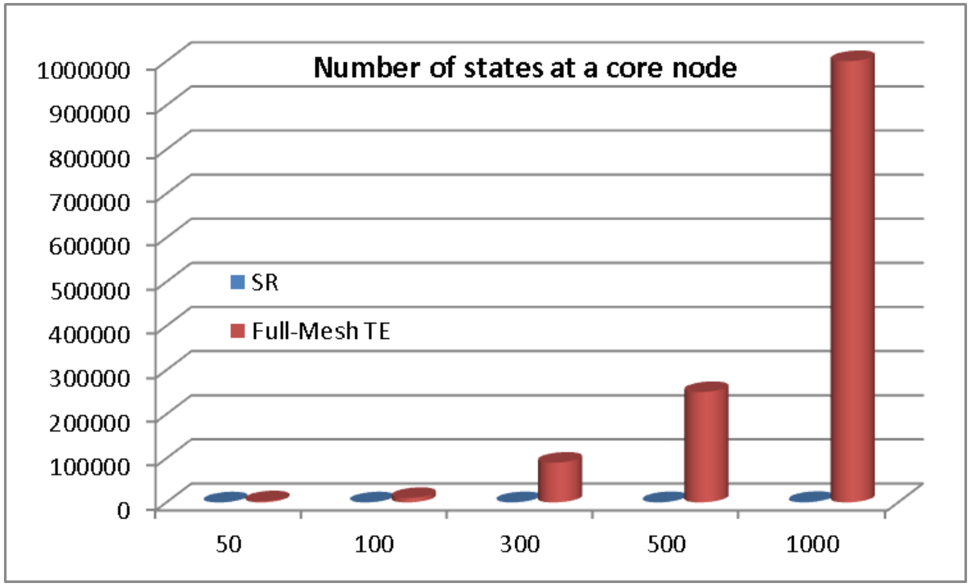
Simple and Efficient Transport of MPLS services



All VPN services ride on the node segment
to PE2

- Full support of ECMP
- Simplicity
  - no complex LDP/ISIS synchronization to troubleshoot
  - one less protocol to operate: LDP (and RSVP-TE)

Scalable TE



- An SR core router scales much better than with RSVP-TE
  - The state is not in the router but in the packet
  - Order N+A vs $(N+A)^2$

N: # of nodes in the network
A: # of adjacencies per node

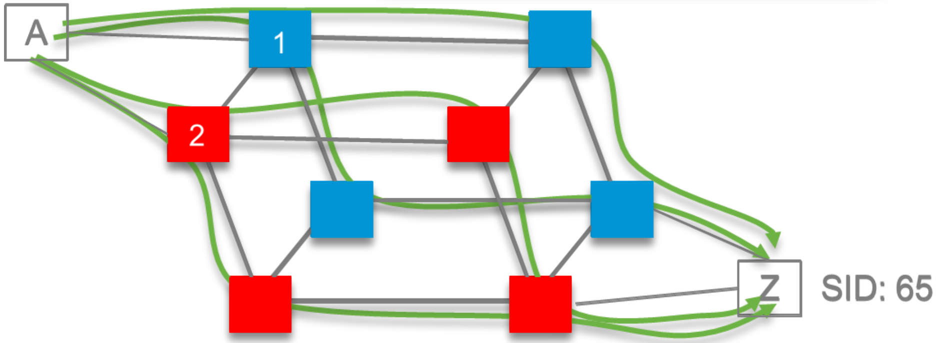
Simple Disjointness

SR avoids state in the core

SR avoids enumerating RSVP-TE tunnels for each ECMP paths

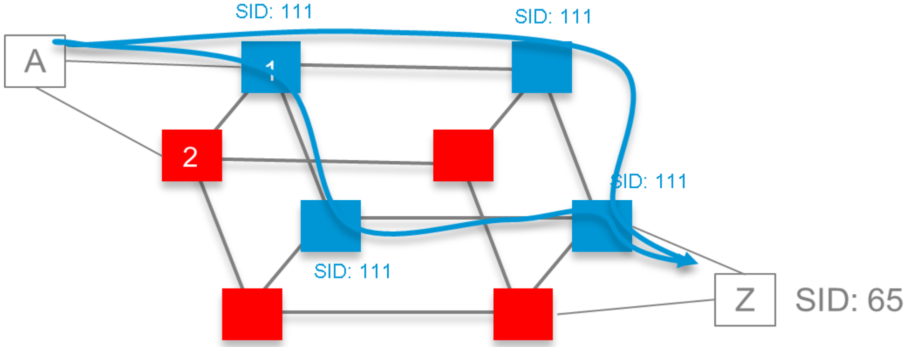
- A sends traffic with [65]

  Classic ECMP “a la IP”



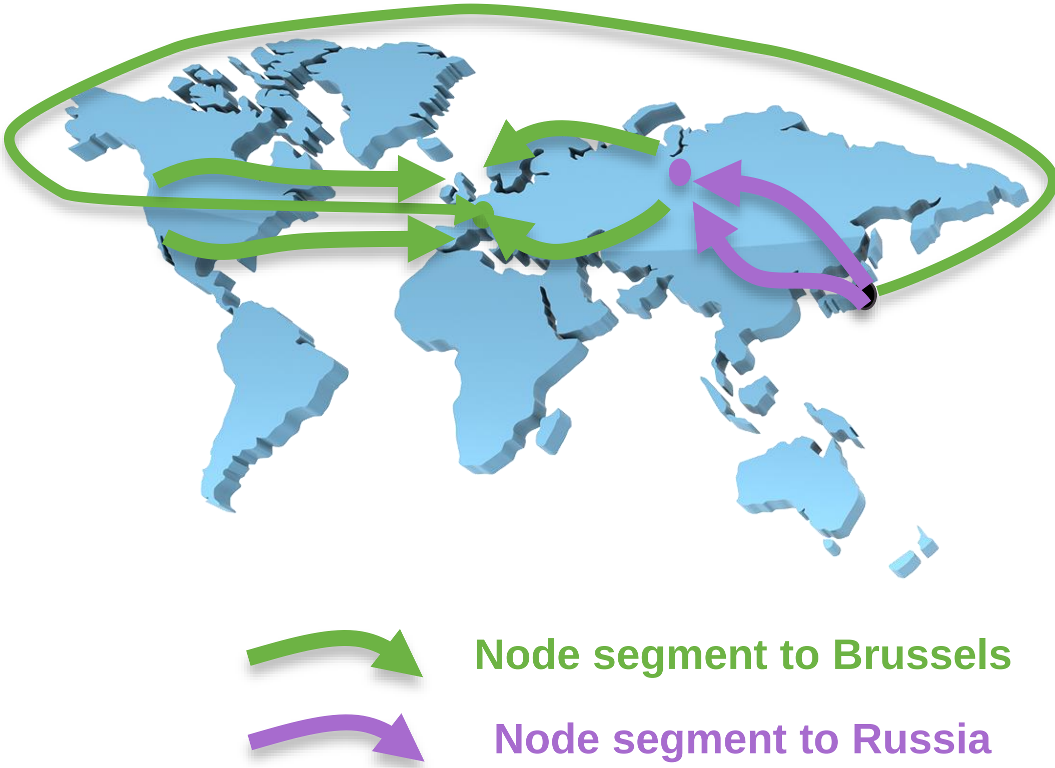
- A sends traffic with [111, 65]

  Packet gets attracted in blue plane and then uses classic ecmp “a la IP”



ECMP-awareness!

CoS-based TE



Node segment to Brussels

Node segment to Russia

- Tokyo to Brussels
  - data: via US: cheap capacity
  - voip: via russia: low latency
- CoS-based TE with SR
  - IGP metric set such as
    - > Tokyo to Russia: via Russia
    - > Tokyo to Brussels: via US
    - > Russia to Brussels: via Europe
  - Anycast segment “Russia” advertised by Russia core routers
- Tokyo CoS-based policy:
  - Data and Brussels: push the node segment to Brussels

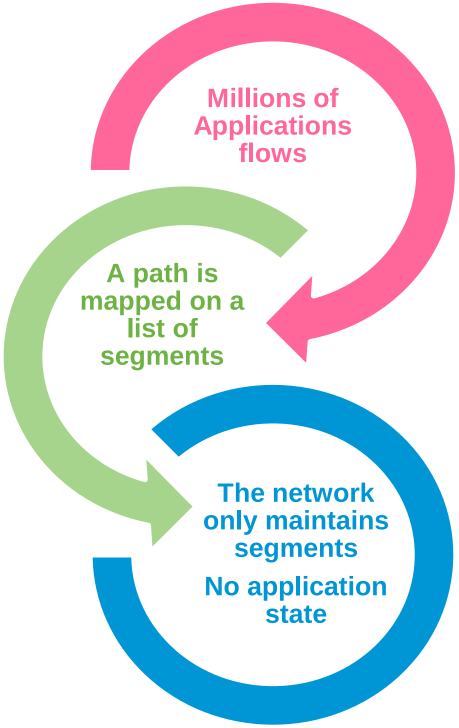
    ➔ ECMP-aware shortest-path to Brussels
  - VoIP and Brussels: push the anycast node to Russia, push Brussels

    ➔ ECMP-aware shortest-path to Russia, followed by ECMP-aware shortest-path to Brussels

No TE tunnel enumeration,
no TE state in the core

Conclusions

- **Simple**: fewer Protocols, less Protocol interaction, less state
  - - No need for RSVP-TE, LDP
- **Scalable**: fewer Label Databases, fewer TE LSPs
  - - Leverage MPLS services & hardware
- **Forwarding** based on MPLS label
  - – minor change to MPLS forwarding plane)
- **Label distributed** by the IGP protocol
  - – simple ISIS/OSPF extensions
- **50msec FRR** service level guarantees via LFA
- Leverages multi-service properties of MPLS


Millions of Applications flows

A path is mapped on a list of segments

The network only maintains segments

No application state

“The state is no longer in the network but in the packet”

Source: Cisco: *https://indico.uknof.org.uk/event/32/contribution/16/.../slides/*