The bridge to possible

Transport Protocols Evolution

Alex Mitev
Cisco

Agenda

- Welcome and Intro
- MPLS, TE, Historical context
- SR MPLS, SRv6. The new way!
- Use Cases
- Appendix

When the present determines the future, but the approximate present does not approximately determine the future.

Edward Lorenz

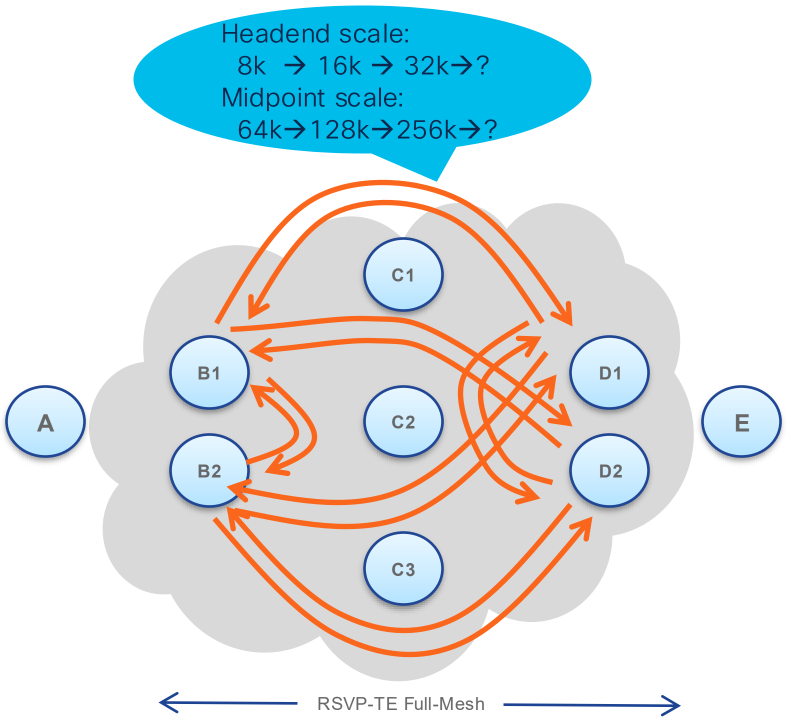
MPLS, RSVP, TE
Historical context

MPLS and TE

- It started with ATM
  - We didn’t like it
  - Expensive HW, cell overhead
  - Another Terrible Mistake?
- TE with IP is hard, really hard and we can’t do it at scale
  - Tuning metrics, redistributing prefixes, leaking specifics
- MPLS TE solved the IP TE scale problem
  - Still hard
  - Distributed

MPLS TE, Challenges

- RSVP gave us more knobs and information at the cost of:
  - Large headend and midpoint scale
  - State maintained at every hop
  - Core devices state k*n^2
- Scaling often requires HW replacements
- No inter-domain; complex steering
- Little deployments
  - Many used it to get FRR
  - Little bit of tactical TE



SR, SRv6
The new way!

SR MPLS

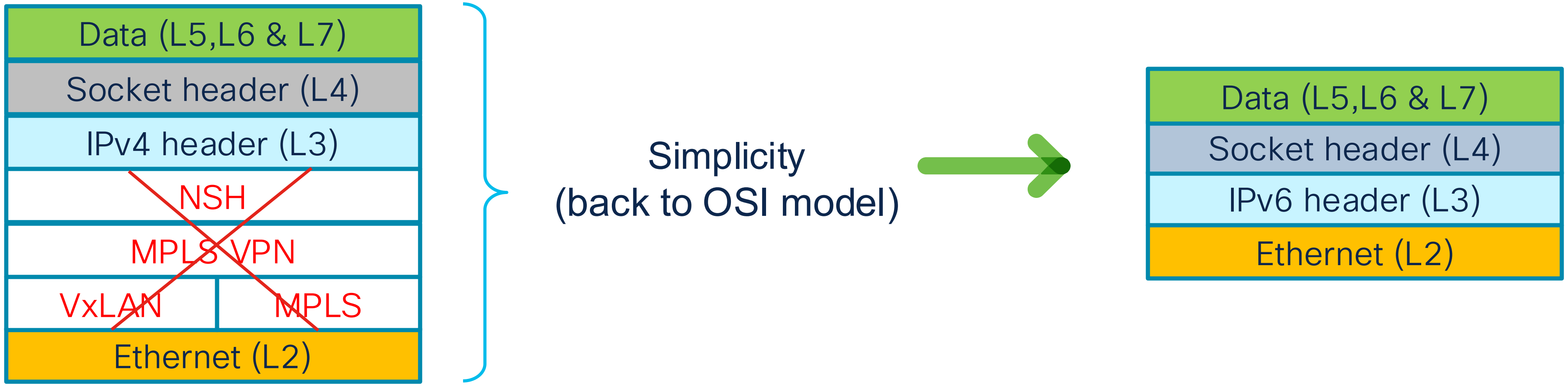
- Simple
  - Eliminate the core scale problem – state is in the packet
  - Utilise existing MPLS data plane
  - No tunnel interfaces
    - Automated steering
  - Multi domain
    - SR PCE
    - BSID
  - SR MPLS eliminates LDP and RSVP protocols

SRv6

Network Functions	IPv6
Reachability	IPv6 Header
Engineered Load Balancing	IPv6 Header
VPN	IPv6 Header
Traffic Engineering	IPv6 Header
Source Routing	IPv6 Header
Service Chaining	IPv6 Header

IPv6 Address 128bits
IPv6 Flow Header
- Engineered Flow optimization

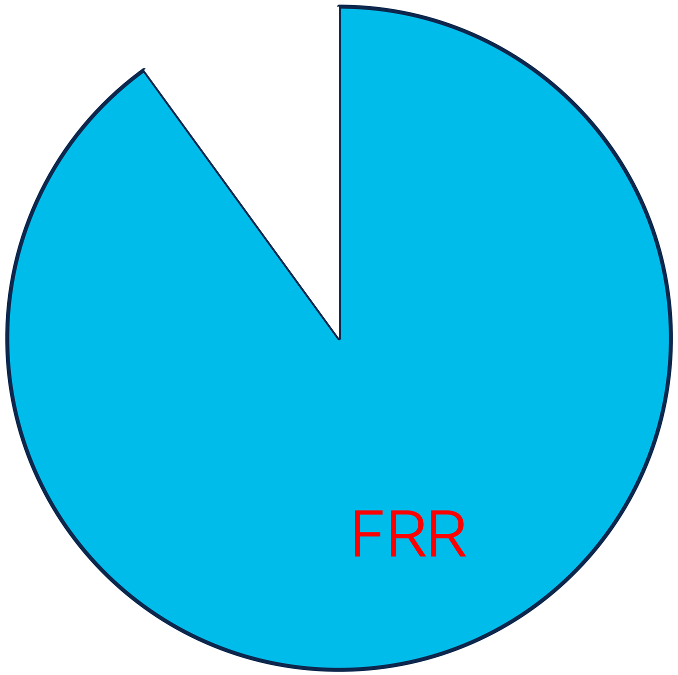
SRv6 Header
- Source-Routing
- Traffic Engineering
- VPN
- Service Chaining



Use Cases

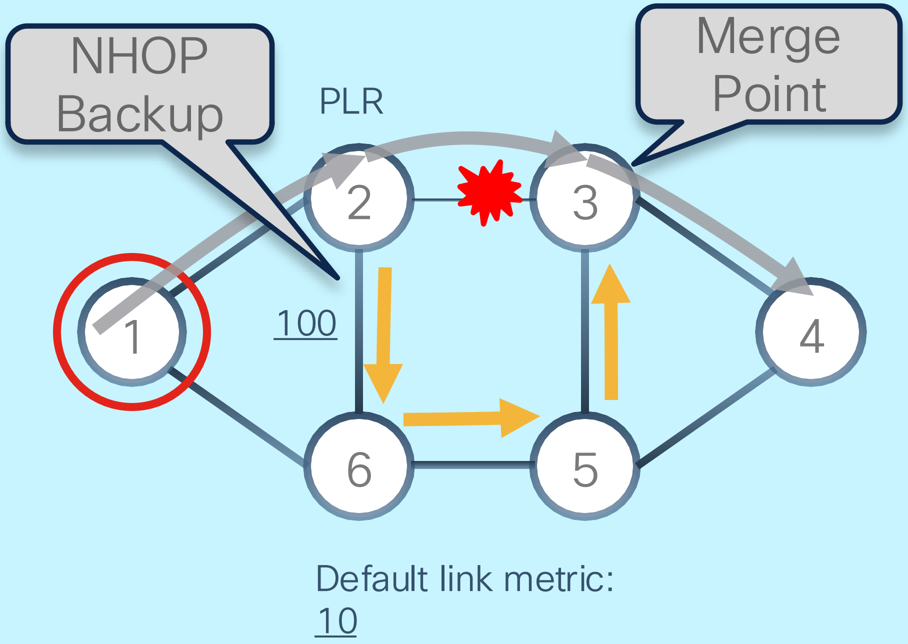
FRR

- 80 to 90%-sh used RSVP TE for FRR
  - Not easy to achieve 100% protection
  - Hard work to set up all paths pre-emptively
    - Housekeeping – making sure TE requirements are always met
  - Slow in some complex topologies
  - Different configuration for link, node, path protection



RSVP TE FRR Configurations - NHOP

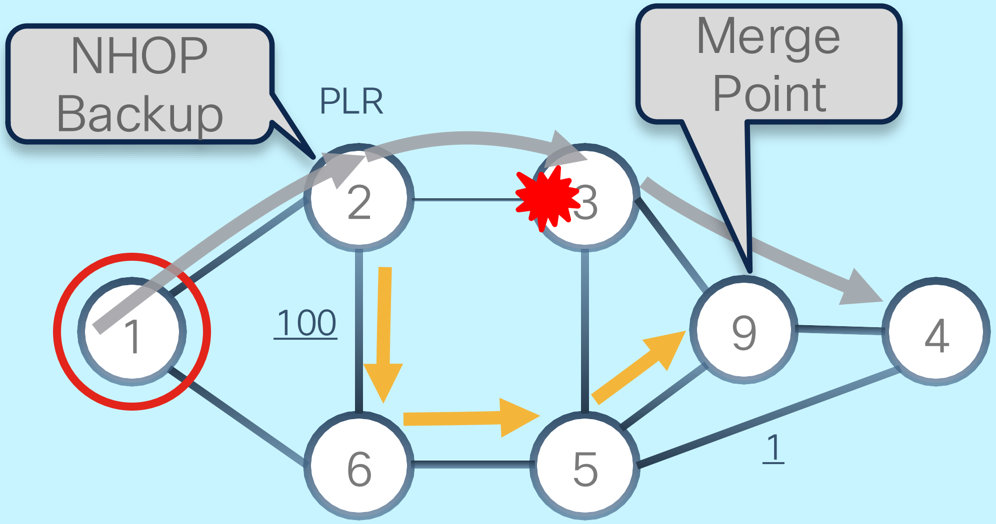
```
PE:
interface tunnel-te0
 ipv4 unnumbered Loopback0
 destination 4.4.4.4
 fast-reroute
 record-route
 autoroute-announce
 path-option 1 explicit name X
```




```
PLR:
interface tunnel-te10
 ipv4 unnumbered Loopback0
 destination 3.3.3.3
 path-option 1 explicit name FRR
!
explicit-path name FRR
 index 1 next-address ipv4 strict unicast 6.6.6.6
 index 2 next-address ipv4 strict unicast 5.5.5.5
!
mpls traffic-eng
 interface TenGigE0/0/0/0
  backup-path tunnel-te 10
```

RSVP TE FRR Configurations - NNHOP

```
PE:
interface tunnel-te0
 ipv4 unnumbered Loopback0
 destination 4.4.4.4
 fast-reroute
 record-route
 autoroute-announce
 path-option 1 explicit name X
```

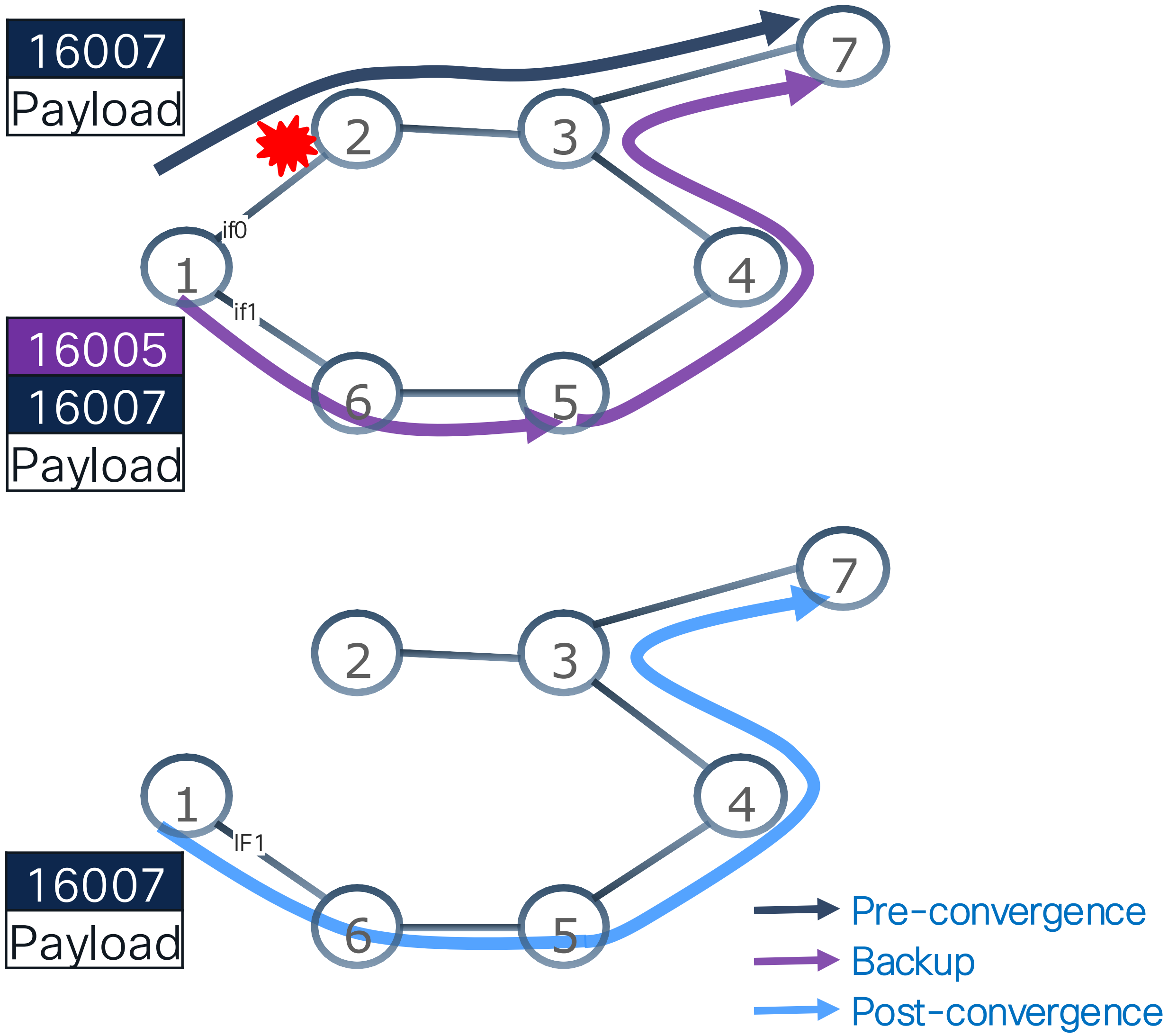


Default link metric: 10

```
PLR:
interface tunnel-te0
 ipv4 unnumbered Loopback0
 destination 9.9.9.9
 autoroute-announce
 path-option 1 explicit name FRR
!
explicit-path name FRR
 index 1 next-address ipv4 strict unicast 26.26.26.6
 index 2 next-address ipv4 strict unicast 5.5.5.5
 index 2 next-address ipv4 strict unicast 59.59.59.9
!
mpls traffic-eng
 interface TenGigE0/0/0/0
  backup-path tunnel-te 0
```

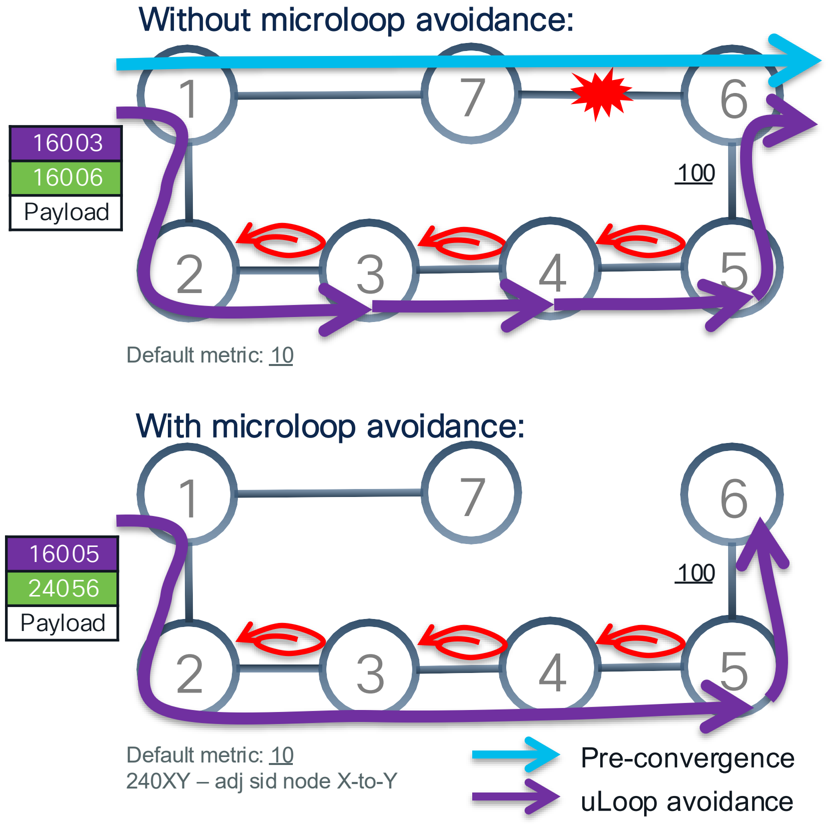
Ti-LFA

- FRR is on the post convergence path
- No sub-optimal backups
- 100% coverage
- ECMP on the backup path



Microloop Avoidance

- Transient packet loop is day 1 problem for the IP networks during convergence
- If there is a possibility of any microloops on the post-convergence path, cmpoute a SID-list to steer the traffic loop free
- Totally stable and predictable network forwarding even after link up/down events



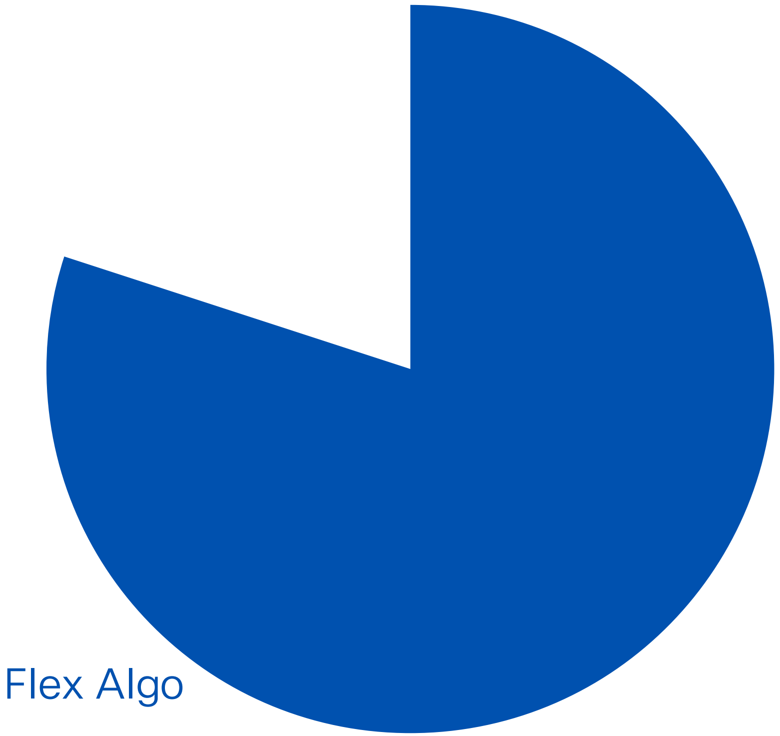
Ti-LFA and Microloop Avoidance

```
router ospf 1
 microloop avoidance segment-routing
 microloop avoidance rib-update-delay 3000
 area 1
  interface GigabitEthernet0/0/2/1
   fast-reroute per-prefix
   fast-reroute per-prefix ti-lfa
   fast-reroute per-prefix tiebreaker node-protecting
index 100
```

```
router isis 1
 address-family ipv4/ipv6 unicast
  microloop avoidance segment-routing
  microloop avoidance rib-update-delay 3000
 interface GigabitEthernet0/0/2/1
  address-family ipv4/ipv6 unicast
   fast-reroute per-prefix
   fast-reroute per-prefix ti-lfa
   fast-reroute per-prefix tiebreaker node-protecting
index 100
```

Flexible Algorithm

- Key building block
  - Majority of the transport intent can be realised with FA metric + constraints
- Optimum Intra/Inter domain paths
- Distributed computation on the HE by IGP
- Native steering of service traffic over FA path
- Min SID list depth - 2



Flexible Algorithm

Flex-Algo	Simplest and most efficient for optimum Intra/Inter-Domain paths for most transport intents (delay, include/exclude affinity, etc.)

A Flex-Algo instance is defined with a **metric** and **constraints**

Minimize path **cumulative metric** based on a given link metric

Metric
IGP
TE
Delay
Bandwidth (max link BW)
Generic

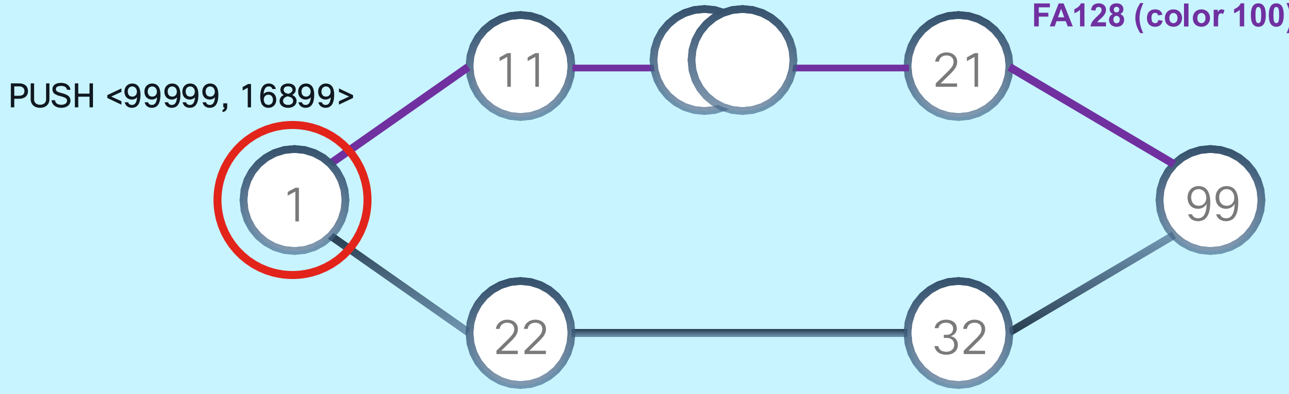
Constrained path computation based on link attributes

Constraints
Affinity (include/exclude)
Reverse Affinity (include/exclude)
Shared Risk Link Groups (exclude)
Min-BW
Max-Delay

FA Use Case Example

```
router isis 1
 is-type level-2-only
 affinity-map PURPLE bit-position 1
 flex-algo 128
  affinity include-any PURPLE
  advertise-definition
 !
 address-family ipv4 unicast
  router-id 1.1.1.1
  segment-routing mpls
 !
 interface Loopback0
  address-family ipv4 unicast
   prefix-sid absolute 16002
   prefix-sid algorithm 128 absolute 16801
 !
 interface GigabitEthernet0/2/0/4
  affinity flex-algo PURPLE
```

```
segment-routing
  traffic-eng
    on-demand color 100
      dynamic
        sid-algorithm 128
```



SR Policies

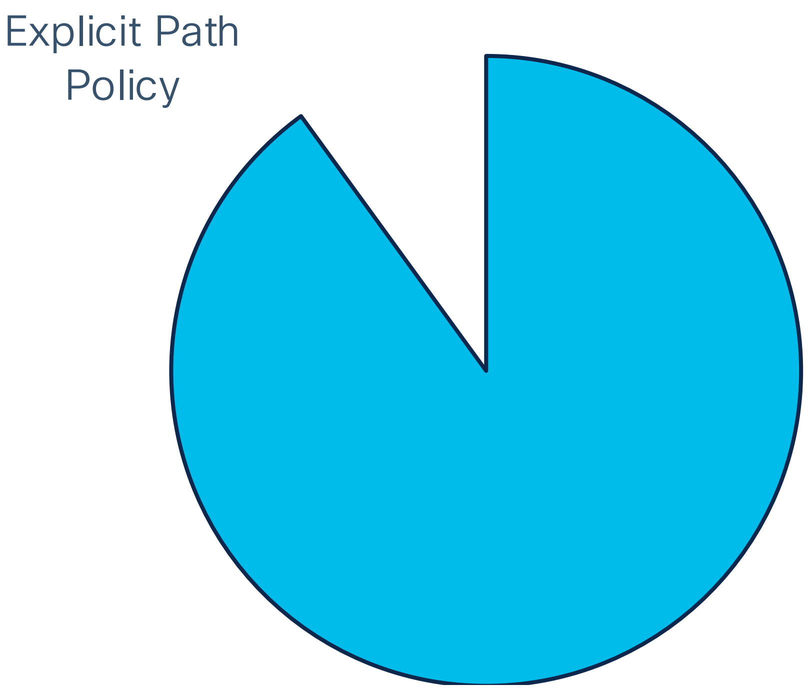
- Colour (C) – numerical value used to differentiate multiple SRTE polices between the same pair of nodes
  - Can be used to indicate treatment of the traffic, SLA that the policy provides
  - Only one with a given colour C can exists between a head-end (H) and end-point (E) –SR policy triplet (H,C,E) is unique
- End-Point (E) – destination of the policy

```
segment-routing
 traffic-eng
  policy POLICY1
   color 20 end-point ipv4 1.1.1.4
   binding-sid mpls 1000
   candidate-paths
    preference 100
     dynamic
      metric type te
     constraints
      affinity
       exclude-any name red
    !
    preference 200
     explicit segment-list SIDLIST1
  !
  segment-list name SIDLIST1
   index 10 mpls label 16002
   index 20 mpls label 30203
   index 30 mpls label 16004
```

Explicit Paths

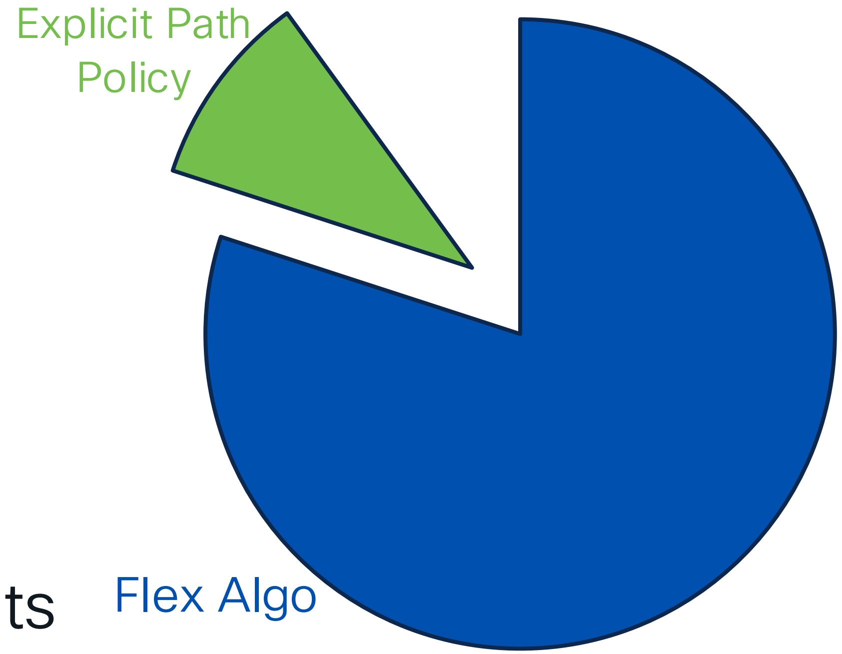
- Majority of RSVP TE policies today
- Explicitly specify strict path through the network
- Protection path can be specified

```
interface tunnel-te1
 path-protection
 path-option 1 explicit name R2-to-R4 protected-by 2
 path-option 2 explicit name backup
!
explicit-path name backup
 1 exclude-address 192.168.91.1
 2 exclude-srlg 192.168.31.2
```



SR Explicit Paths

- Between 5% and 10% of TE use cases
- When you need to explicitly specify strict/loose SID list
- Wighted SID list
- Manual Tactical TE
  - Steer around congestion
  - Steer part of the traffic to avoid congested points



SR Explicit Paths

```
segment-routing
 traffic-eng
  policy POLICY1
   color 20 end-point ipv4 1.1.1.4
   binding-sid mpls 1000
   candidate-paths
❶   preference 100
     dynamic
      metric type te
     constraints
      affinity
       exclude-any name red
    !
❷   preference 200
     explicit segment-list SIDLIST1
  !
  segment-list name SIDLIST1
   index 10 mpls label 16002
   index 20 mpls label 30203
   index 30 mpls label 16004
```

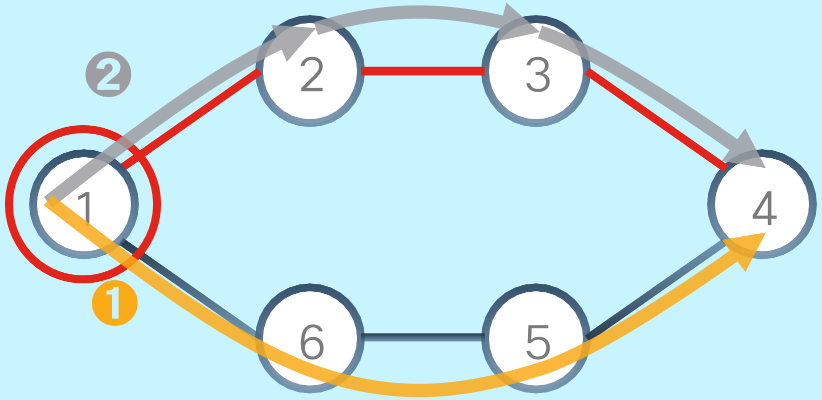
User-defined name

Color and End-point

SR Policy ID: (20,1.1.1.4)

Binding-SID

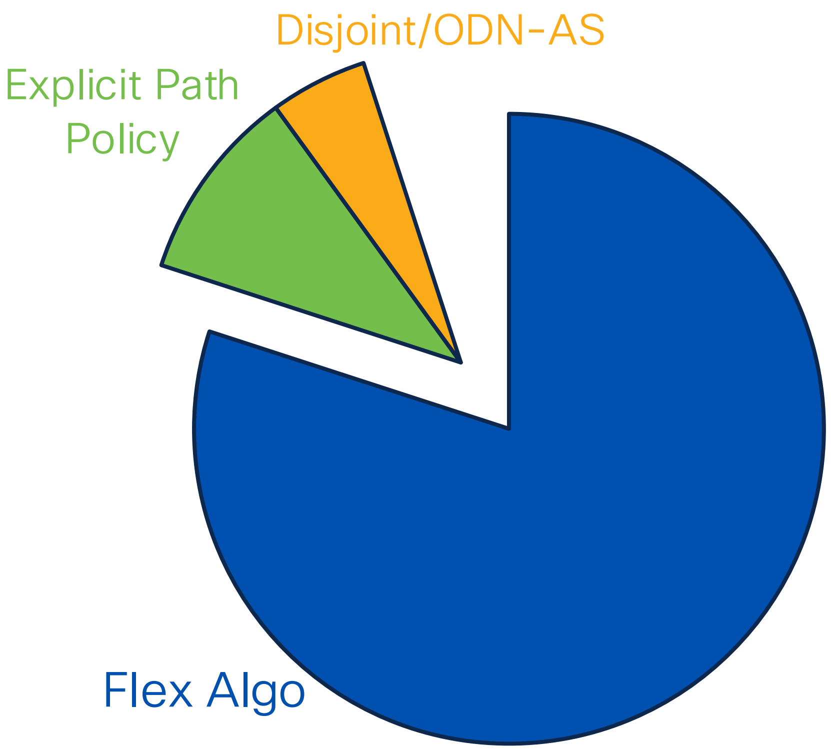
Local Candidate Paths




```
segment-routing
 traffic-eng
  affinity-map
   name red bit-position 0
```

On-Demand Nexthop

- HE automatically instantiates SR policy to BGP next-hop when required
- SR-PCE support intent
  - Min metric (any metric)
  - Include or Exclude any constraints
  - Disjoint
- SLA aware, Simple and Scalable
- Automated Steering (AS)
- Per-flow



95%

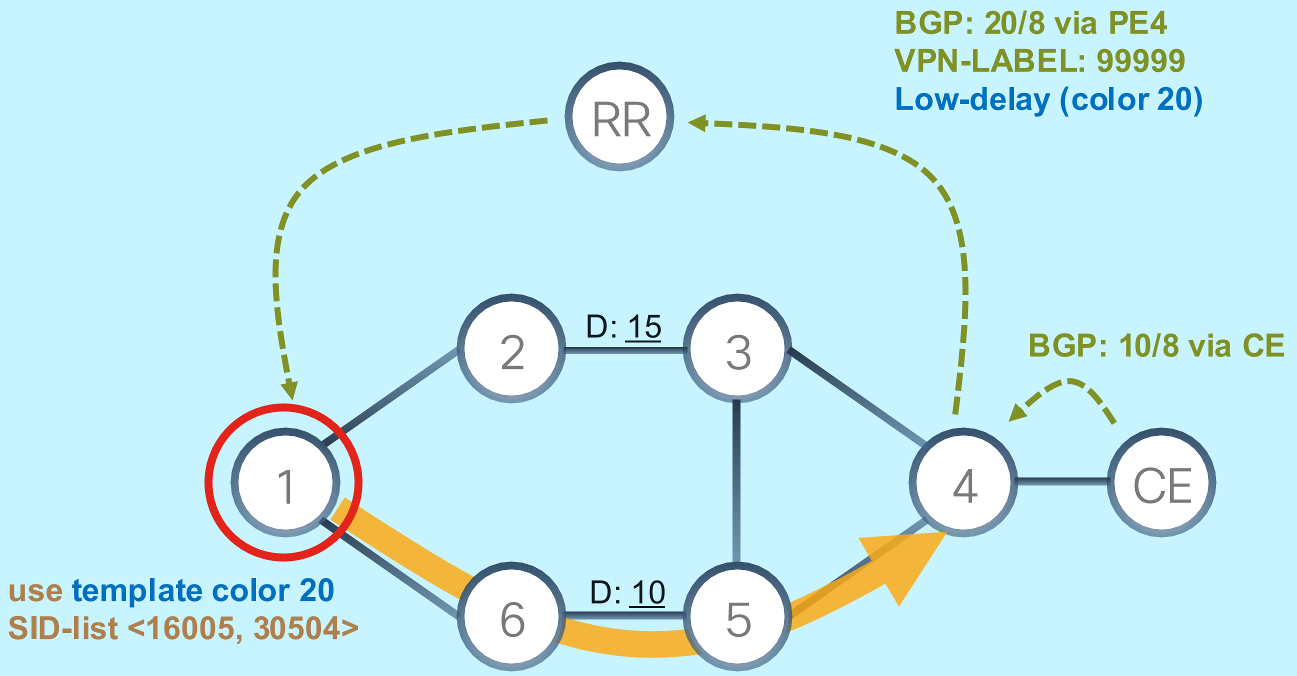
SR ODN

```
segment-routing
 traffic-eng
  on-demand color 10
   dynamic
    metric type igp
  !
  on-demand color 20
   dynamic
    metric type latency

extcommunity-set opaque color10-igp
  10
end-set
!
extcommunity-set opaque color20-ltcy
  20
end-set
!
route-policy SET_COLOR_LOW_LATENCY
  if destination in (10/8) then
  set extcommunity color color20-ltcy
  pass
end-policy
```

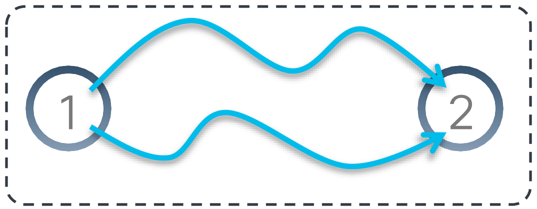
SR Policy template High-BW (color 10)

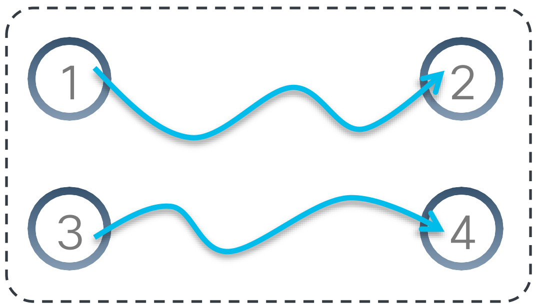
SR Policy template low-delay (color 20)

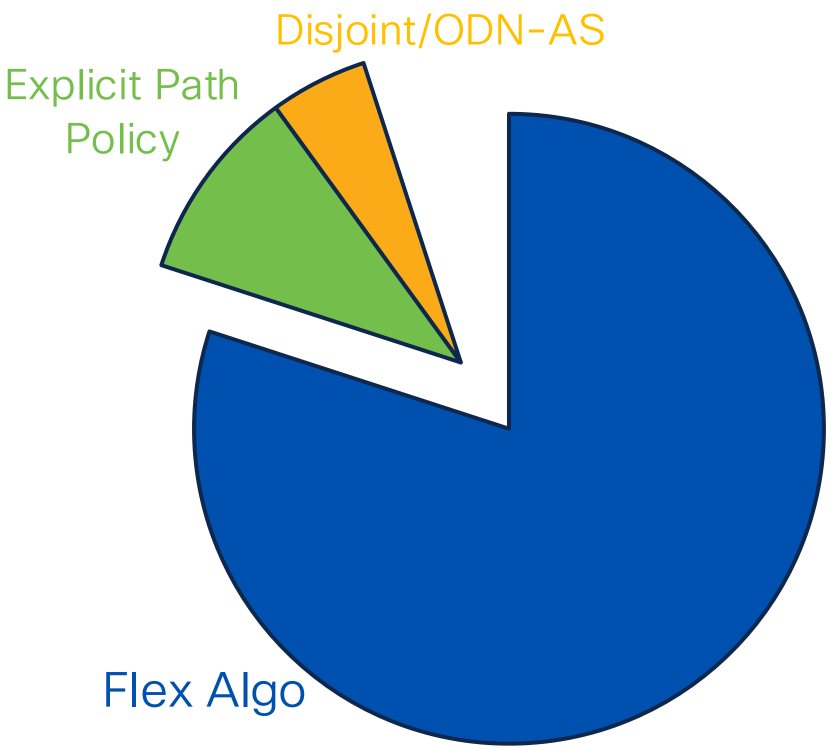


Disjointness

- Same HE – disjointness between a pair of path
  - SRTE can compute a path that is disjoint from another path in the same disjoint-group
- Different HE and Tailends - PCE

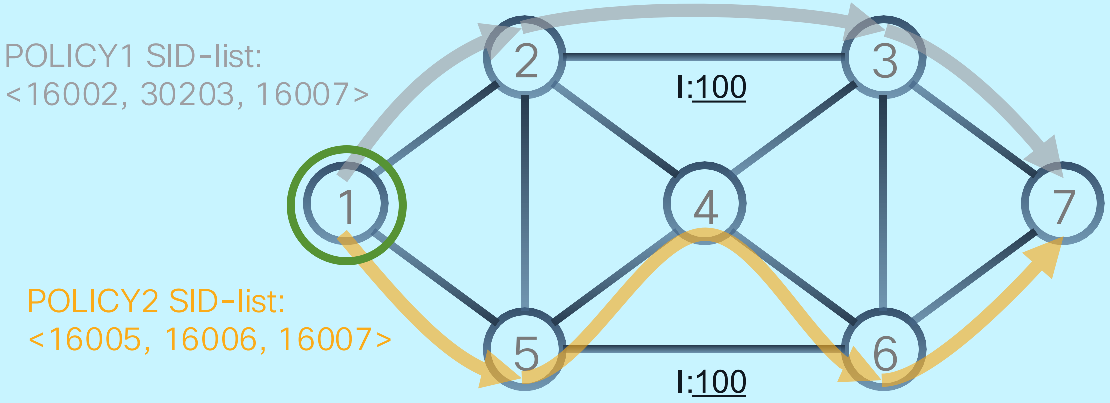






Disjointness, same HE/TE

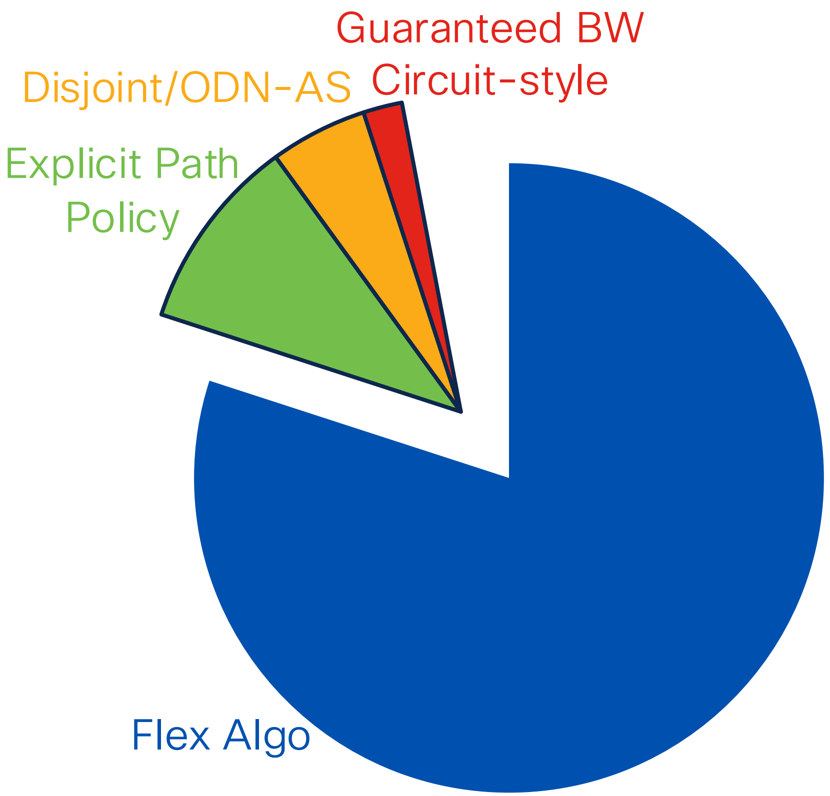
```
segment-routing
 traffic-eng
   policy POLICY1
    color 20 end-point ipv4 1.1.1.7
    candidate-paths
     preference 100
      dynamic
       metric type igp
      constraints
       disjoint-path group-id 1 type node
   !
   policy POLICY2
    color 30 end-point ipv4 1.1.1.7
    candidate-paths
     preference 100
      dynamic
       metric type igp
      constraints
       disjoint-path group-id 1 type node
```



Advance Use Cases

Guaranteed Bandwidth/Circuit-style

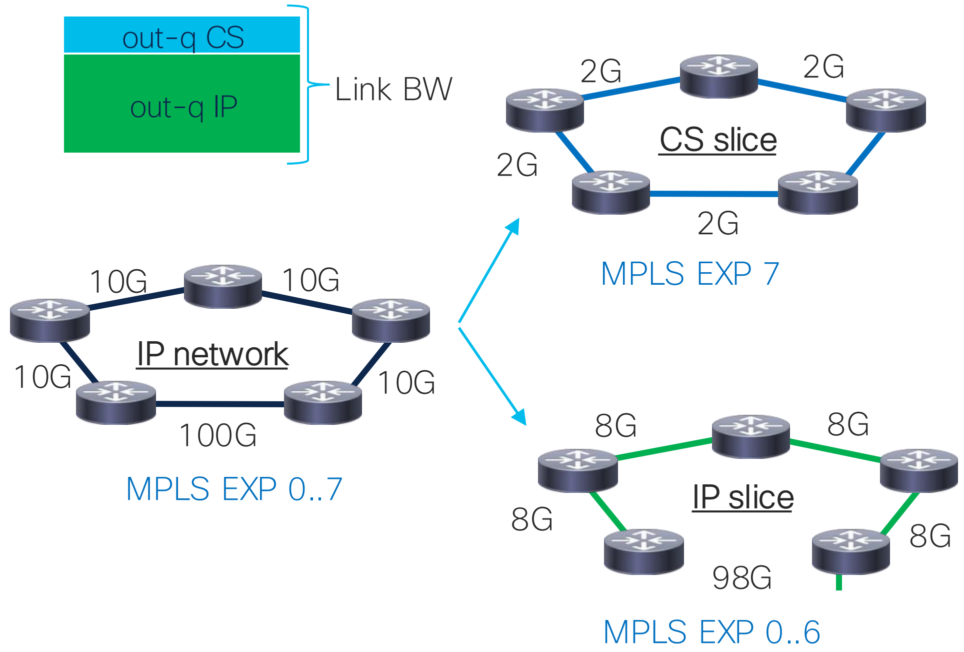
- SR policy that include optional requested BW and priority
- Reserved BW on each link is shared resource managed by an SDN controller
  - Total reservable BW on each link is advertised in the IGP
  - Compute bi-directional, co-routed paths with protection
- QoS mechanisms to unsure BW isolation and no that more than the reserved BW is allowed (admission control)



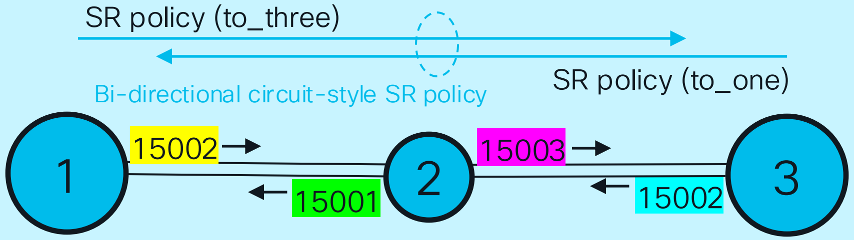
97%

Guaranteed Bandwidth/Circuit-style

- QoS for partitioning and Isolation
- One MPLE-EXP is allocated to CS slice
- Rules
  - CS/Guaranteed traffic is policed to to requested value else -> **dropped**
  - During congestion only other traffic is dropped
  - **Don’t commit what you don’t have**



SR Policy Configuration with static CS-SR




```
segment-routing
 traffic-eng
  segment-list WFlist
   index 1 mpls label 15002
   index 2 mpls label 15003
  !
  segment-list WRlist
   index 1 mpls label 15002
   index 2 mpls label 15001
  !
  policy to_three
   color 10 end-point ipv4 3.3.3.3
   candidate-paths
    preference 20
     explicit segment-list WFlist
      reverse-path segment-list
Wrlist
```

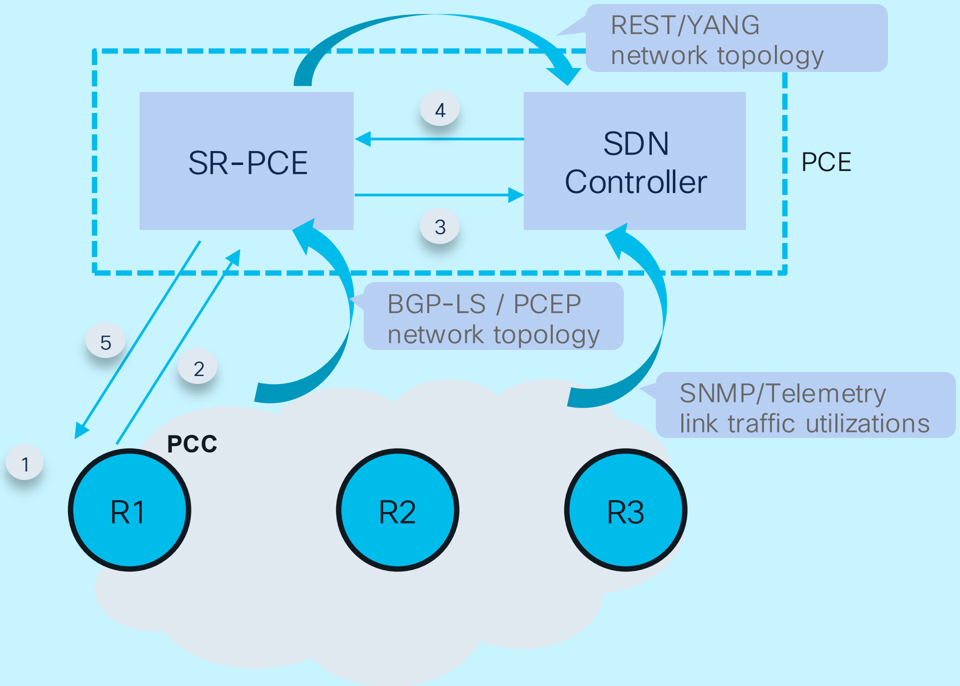
```
router isis rtr1
 ...
 interface HundredGigE0/0/2/0
  address-family ipv4 unicast
   adjacency-sid absolute 15002
```

```
segment-routing
 traffic-eng
  segment-list WFlist
   index 1 mpls label 15002
   index 2 mpls label 15001
  !
  segment-list WRlist
   index 1 mpls label 15002
   index 2 mpls label 15003
  !
  policy to_one
   color 10 end-point ipv4 1.1.1.1
   candidate-paths
    preference 20
     explicit segment-list WFlist
      reverse-path segment-list
WRlist
```

CS-SR Service Policy with BW reservation

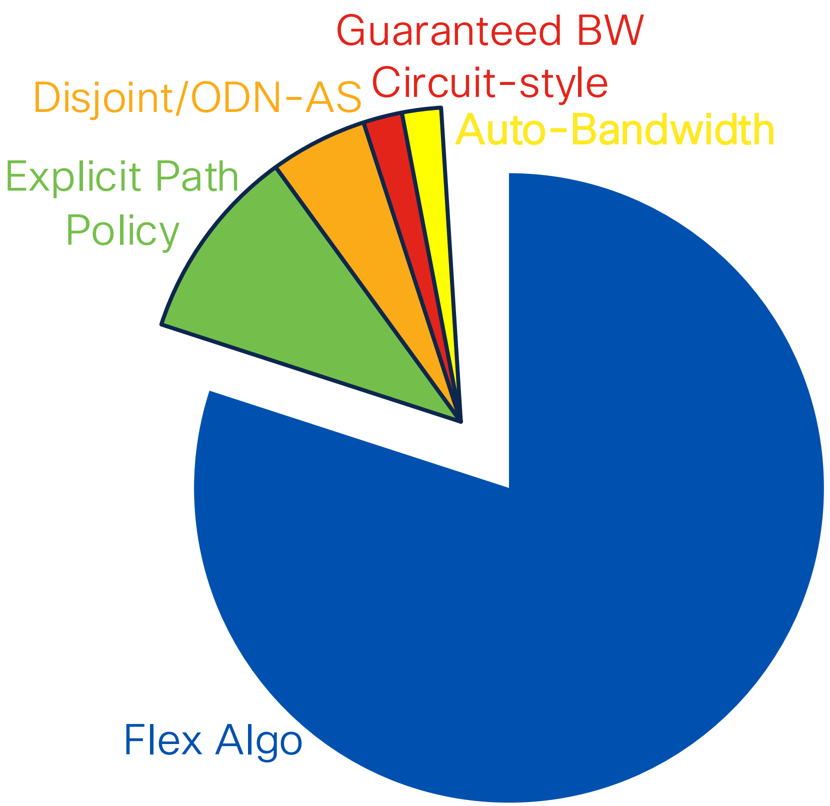
1. Operator configures SR-TE policy with bandwidth constraint
2. PCC sends PCReq to SR-PCE controller
3. SR-PCE requests BW-path from SDN controller
4. SDN controller returns BW-path (or no-path) to SR-PCE
5. SR-PCE sends BW-path (or no-path) to PCC

```
segment-routing
 traffic-eng
  policy to_three
     bandwidth 1000
     path-protection
     color 100 end-point ipv4 3.3.3.3
     candidate-paths
      preference 100
       dynamic pcep
       constraints
        segments protection unprotected-only
        disjoint-path group-id 100 type link
        bidirectional co-routed association-id 100
      preference 50
       dynamic pcep
```



Auto-Bandwidth

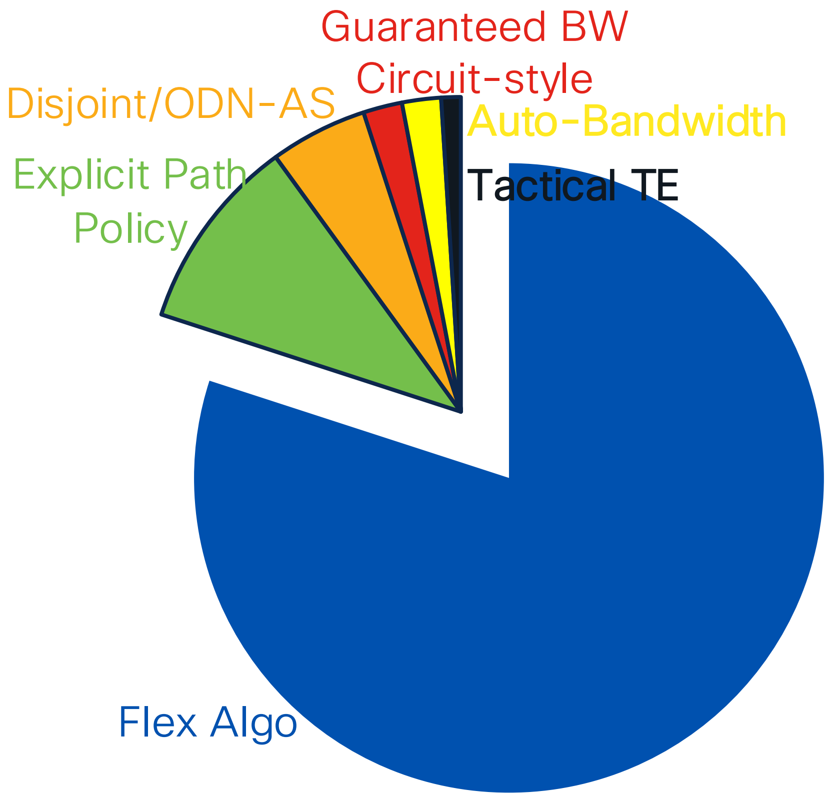
- TE Auto-bw without RSVP
  - Bandwidth is managed centrally (allocated and locked)
- Full mesh
  - HE dynamically and periodically adjust bw based on avg traffic load



99%

Tactical TE (BW optimisation)

- Congestion mitigation
  - Unexpected failure
  - Traffic rise etc.
- SRv6 Deterministic Damand Matrix (DDM)
  - Per locator counters
- Capacity Planning (ACP)
  - Collect DDM counters from all edges
  - Simulates multiple what-if



100%

Traffic Engineering Use Cases

Use Cases	
Flex-Algo	Simplest and most efficient for optimum Intra/Inter-Domain paths for most transport intents (delay, include/exclude affinity, etc.)
Explicit Paths	For cases where you need to statically/explicitly specify loose/strict SID-lists or a set of weighted SID-lists
Disjointness	PCC-initiated ODN/AS for SR policy instantiation for optimum Intra/Inter-Domain paths for special transport intents such as path disjointness
Guaranteed BW	When SLA compliance requires reservation of dedicated BW resources and network is partitioned with high priority QoS queues
Auto-BW	For the RSVP-TE minded. States at the midpoint nodes are eliminated and periodic BW requests from PEs are centrally managed
Tactical TE / BW Opt.	Congestion mitigation solution deployed in exceptional situations such as unexpected failure scenarios, traffic rise and/or poor capacity planning

The Value

Network Availability

Introduce seamlessly
Protect with automatic TI LFA FRR
Stabilize with microloop avoidance
Operate with advanced monitoring and blackhole detection
Monitor with SR Performance Measurement toolkit
...and more — Integrated Performance Measurement (IPM)
...and more — Hop-by-hop packet visibility with Path Tracing

New Revenue Streams

Path Disjointness (Multi-plane)
Real-Time Low Latency Services
Egress Peer Engineering (EPE)
Point-to-Multipoint delivery with Tree-SID
Bandwidth Optimization
...and more — SD-WAN Overlay + SRv6 Underlay Integration

Intent-Based Traffic Engineering

On-Demand Next-Hop (ODN) + Automated steering (AS)
Multi-plane Network Slicing using IGP Flex Algorithms
Multi-Domain intent with SR-PCE
Circuit-Style SR Policies
...and more — Native Stateless Service Chaining
...and more — Drastically Simplified Host Networking
...and even more — Flow Placement in AI Backend/Frontend

Standards

Architecture

- SR Architecture – **RFC 8402**
- SRTE Policy Architecture – **RFC 9256**
- Compressed SRv6 Segment List – **RFC 9800**

Data Plane

- SRv6 Network Programming – **RFC 8986**
- IPv6 SR Header – **RFC 8754**

Control Plane

- SRv6 BGP Services – **RFC 9252**
- SRv6 ISIS – **RFC 9352**
- SR Flex-Algo – **RFC 9350**

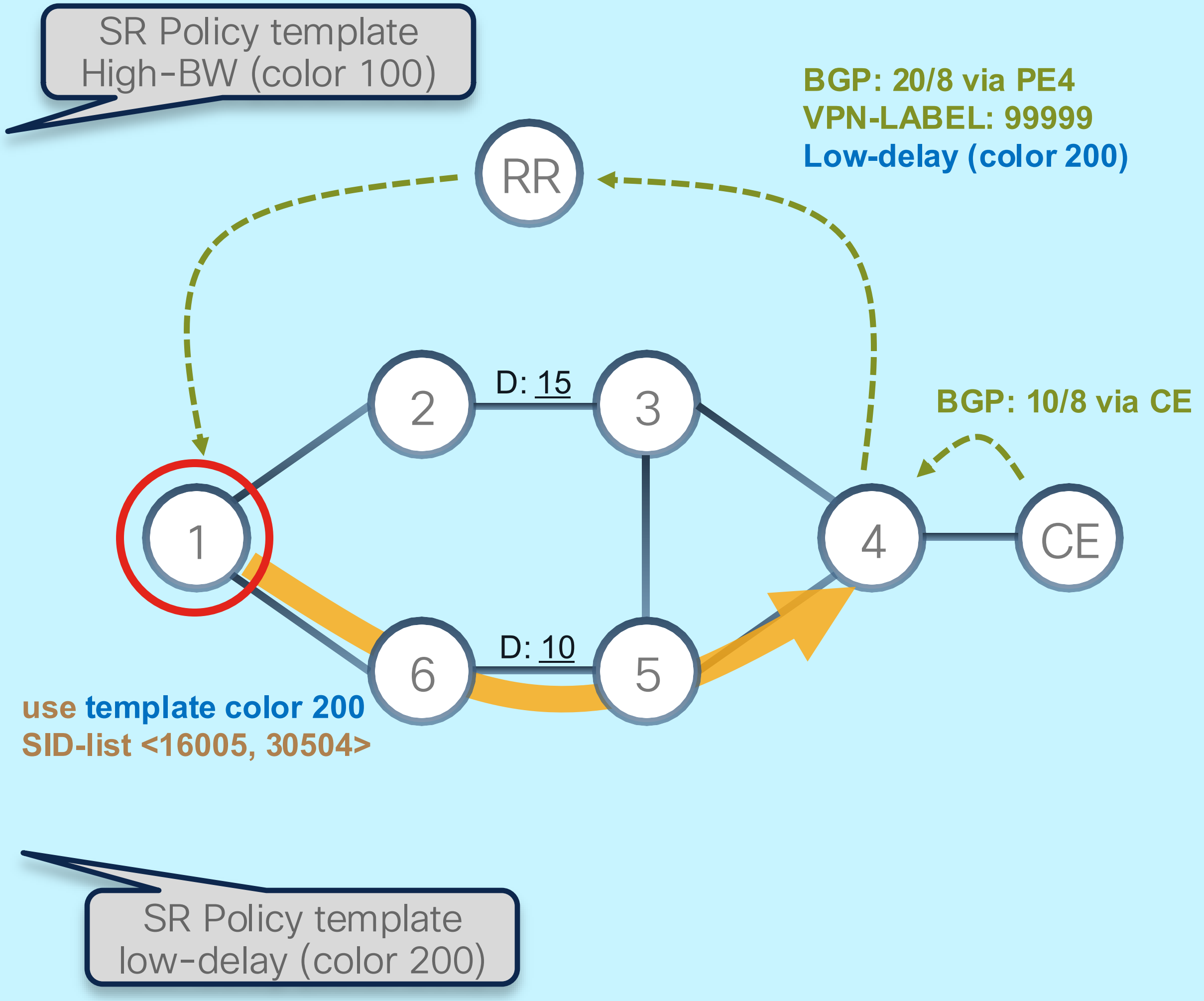
Operation & Management

- SRv6 OAM – **RFC 9259**
- Performance Measurement – **RFC 9503**

Appendix

SRv6 ODN

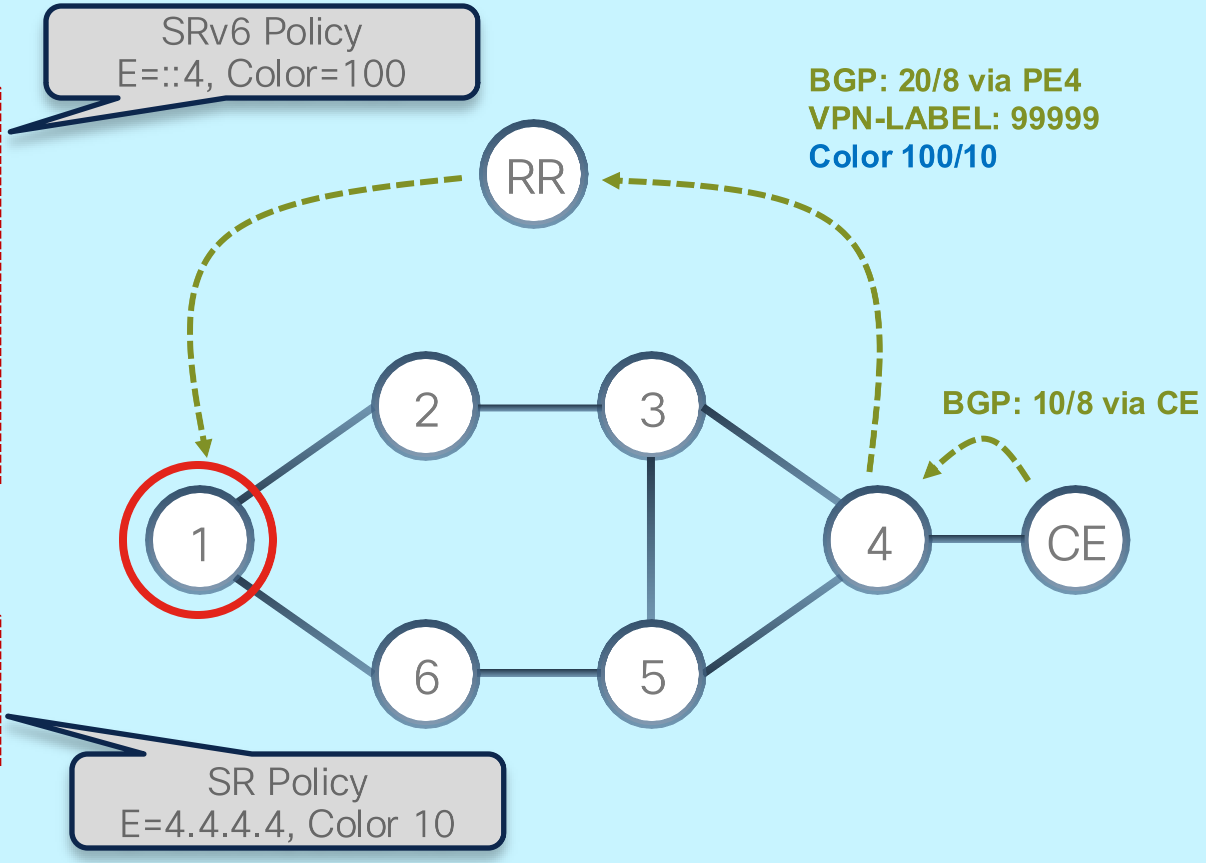
```
segment-routing
 traffic-eng
  on-demand color 100
   srv6
    locator MAIN binding-sid dynamic behavior ub6-insert-
reduced
   !
   dynamic
    !
    metric
     type igp
 !
  on-demand color 200
   srv6
    locator MAIN binding-sid dynamic behavior ub6-insert-
reduced
   !
   dynamic
    !
    metric
     type latency
```



Automated Steering

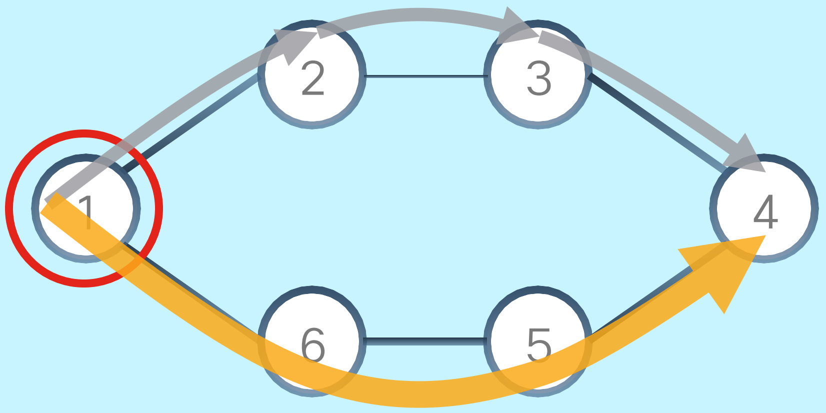
```
segment-routing
 traffic-eng
  color 100 end-point ipv6 fcbb:bb00:107::4
   srv6
    locator MAIN binding-sid dynamic behavior ub6-insert-
reduced
   !
   dynamic
    !
    metric
     type igp
```

```
 segment-routing
  traffic-eng
  color 10 end-point ipv4 4.4.4.4
   dynamic
    metric type igp
```



RSVP TE FRR Configurations
PATH Protection

```
interface tunnel-te0
 destination 4.4.4.4
 path-protection
 path-option 1 explicit name R2-to-R4
 path-option 2 dynamic
!
interface tunnel-te1
 destination 4.4.4.4
 path-protection
 path-option 1 explicit name R2-to-R4 protected-by 2
 path-option 2 explicit name R2-R6-R5
!
explicit-path name R2-to-R4
 index 1 exclude-address ipv4 unicast 6.6.6.6
!
explicit-path name R2-R6-R5
 index 1 exclude-address ipv4 unicast 2.2.2.2
```



SR Explicit Paths - WECMP

```
segment-routing
 traffic-eng
  policy POLICY1
   color 20 end-point ipv4 1.1.1.4
   binding-sid mpls 1000
   candidate-paths
    preference 200
     explicit segment-list SIDLIST1
      weight 1
     !
     explicit segment-list SIDLIST2
      weight 4
     !
  segment-list name SIDLIST1
   index 10 mpls label 16002
   index 20 mpls label 30203
   index 30 mpls label 16004
  !
  segment-list name SIDLIST2
   index 10 address ipv4 1.1.1.4
```

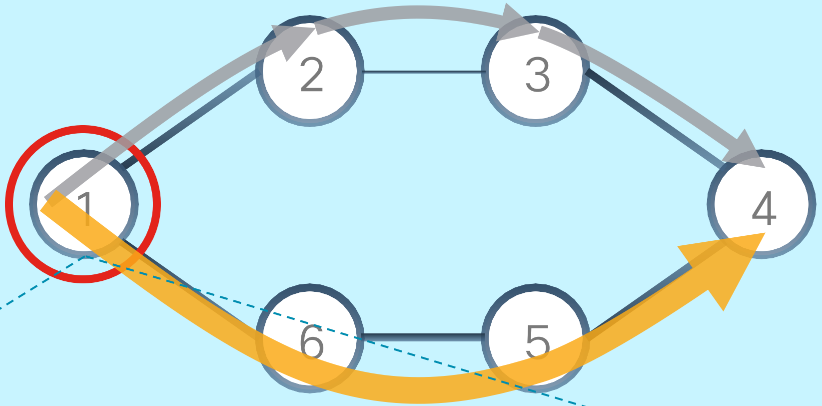
Path preference 200

Explicit SID-list1, Weight 1

Explicit SID-list2, Weight 4

SID-list1

SID-list2



<u>FIB @ head-end Node1</u>
Incoming label: 1000
Action:pop and push <16002, 30203, 16004> (20%)
push <16004> (80%)

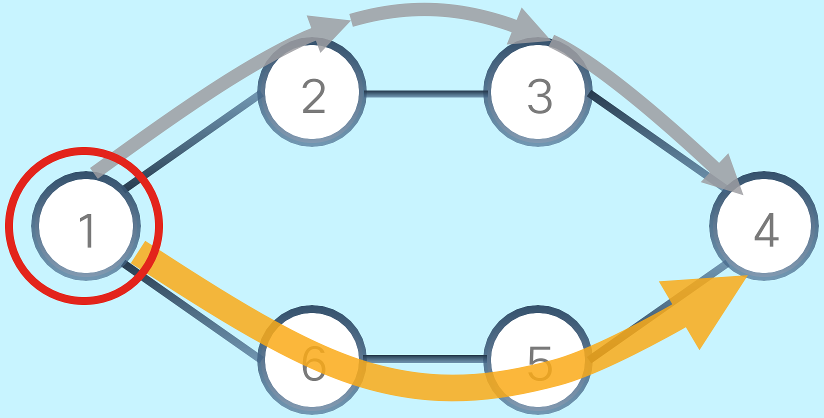
SRv6 Explicit Paths - WECMP

```
segment-routing
 traffic-eng
  srv6
  !
  segment-lists
   srv6
    sid-format usid-f3216
   !
   segment-list r1_to_r4
    srv6
     index 10 sid fcbb:bb00:2:e003::
     index 20 sid fcbb:bb00:4::
    !
   segment-list r1_to_r4_b
    srv6
     index 10 sid fcbb:bb00:2::
     index 20 sid fcbb:bb00:6::
     index 30 sid fcbb:bb00:5::
     index 40 sid fcbb:bb00:4::
```

SID-list1

SID-list2

```
segment-routing
 traffic-eng
  policy P1toR6
   srv6
    locator MAIN binding-sid dynamic behavior ub6-insert-reduced
   !
   color 200 end-point ipv6 fcbb:bb00:4::1
   candidate-paths
    preference 100
     explicit segment-list r1_to_r4
      weight 1
     !
     explicit segment-list r1_to_r4_b
      weight 4
```



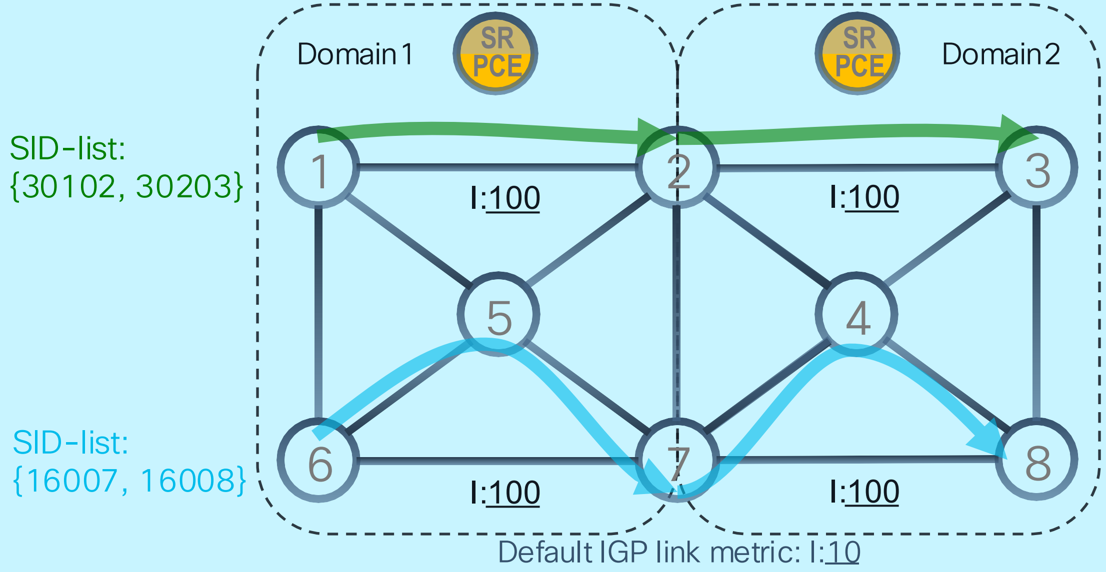
Service Disjointness. Diff HE and TE

```
segment-routing
 traffic-eng
  policy POLICY1
   color 20 end-point ipv4 1.1.1.3
    candidate-paths
     preference 100
      dynamic
       pcep
       metric type igp
      constraints
       association group 1 type node
```

NODE 1

```
segment-routing
 traffic-eng
  policy POLICY2
   color 20 end-point ipv4 1.1.1.8
    candidate-paths
     preference 100
      dynamic
       pcep
       metric type igp
      constraints
       association group 1 type node
```

NODE 6



SRTE/SRv6 in Layer 2 Services

- The SR Policy used to transport Pseudowire traffic can be specified using the preferred-path configuration
- EVPN works similarly
- ODN

```
l2vpn
 pw-class EoMPLS-PWCLASS
  encapsulation mpls
   preferred-path sr-te policy POL1
 !
 xconnect group XCONGRP
  p2p XCON-P2P
   interface TenGigE0/1/0/3
   neighbor ipv4 1.1.1.3 pw-id 1234
    !! below line only if not using LDP
    mpls static label local 2222 remote 3333
    pw-class EoMPLS-PWCLASS
------------------------------------------------
l2vpn
 pw-class evpn-srte
  encapsulation mpls
   preferred-path sr-te policy srte_c_10_ep_4.4.4.4
 !
 xconnect group EVPN-vpws
  p2p EVPN-vpws-400
   interface HundredGigE0/1/0/2.400
   neighbor evpn evi 400 target 333 source 111
    pw-class evpn-srte
```

EVPN RPL Matching

- Route Type – **evpn-routet-ype is (1|2|3|4|5|6|7|8)**
- RD – **rd in (1.1.1.1:0)**
- ESI – **esi in (1110:1110:0101:ffff:1f11)**
  - Net attribute in EVPN RT 1 and 4
  - Path attribute in EVPN RT 2 and 5
- ETAG – **rtag in 1000**
  - Identifies bcast domain
  - Net attribute in RT 1, 2, 3 and 5
  - IP Prefix – destination in (10.0.0.1/32) or destination in (10.0.0.0/24)

EVPN RPL Matching

- Originator – evpn-originator in (4.4.4.4)
  - Only for RT3. Originator’s IPv4 and IPv6
- EVPN gateway – evpn-gateway in (4.4.4.4)
  - Path attribute for RT5. IPv4 or IPv6 address
- MAC address – mac in (1234.5678.9098)
  - RT2