Segment Routing

A tutorial

- Paresh Khatri

- Jan 20, 2021

Agenda

1. Introduction

2. Use cases and applicability

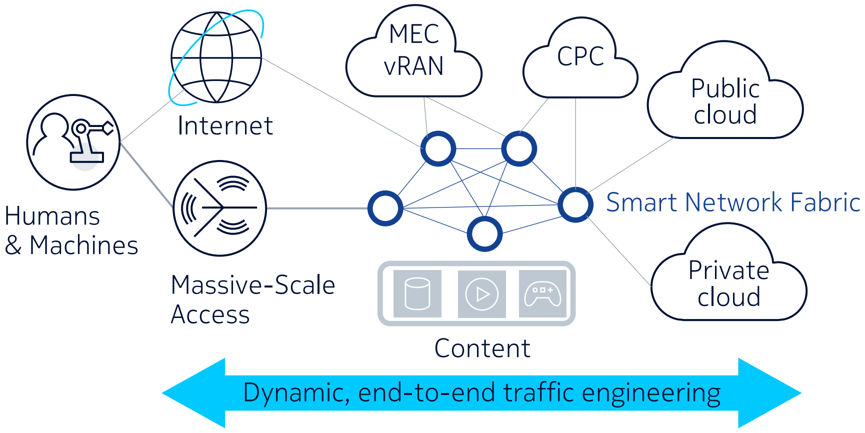
3. Deployment options

introduction

  Public

Tackling network complexity
Segment routing is key to future network design

- Evolution to cloud-centric design
- Future IP networks will need finer routing control to enable new applications
- RSVP-TE and LDP are de-facto in most networks but fall short:
  - Protocol-related scaling issues
  - Traffic engineering is complex to implement and has limitations
  - Source-based steering is complex
  - Load balancing and fast reroute support have limitations

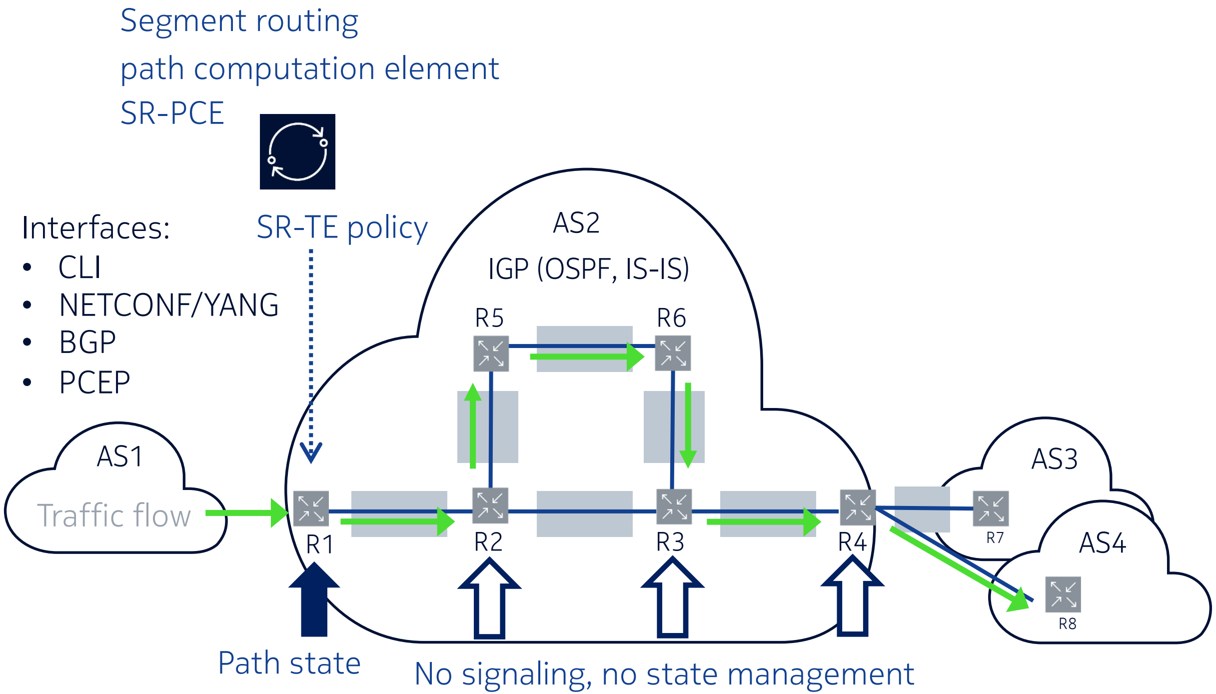


NOKIA

What is segment routing?

A source-based routing protocol to steer IP traffic along a specific path

- Predefines paths to meet desired constraints
- Based on extensions to IGP protocols
- Routing done with a list of IDs (SIDs) for segments to traverse, appended to the packet
- Only the ingress router maintains policy and state information about the path



Complementing segment routing with an SDN controller
Benefits

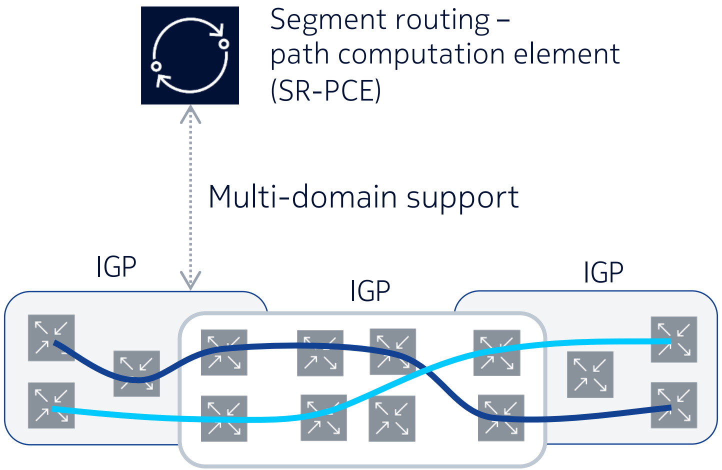
✓ Real-time network **visualization**

✓ Improved **SLA adherence** by engineering paths end-to-end with routing decisions based on
- Actual **utilization**
- **Bandwidth** (bandwidth management)
- **Latency** requirements

✓ **Automated** path [re-]routing
- to distribute traffic for optimal network **capacity usage**
- to trigger path **re-optimization**
- to prevent issues from **maintenance** actions

✓ **Resiliency** enhancements

✓ Offline **simulation**



Optimize end-to-end paths to avoid congestion, latency and SLA impact

MPLS: a historical perspective (1)

- Two main protocols: LDP or RSVP-TE
  - LDP for scale and simplicity
    - extensions for fast re-route (loop-free alternates [LFAs])
  - RSVP-TE for TE and FRR for some time
- To scale MPLS we enabled:
  - LDPoRSVP
  - Seamless MPLS: Labeled-BGP with LDP/RSVP-TE
- Traffic engineering: RSVP-TE based
- Services through:
  - BGP/IGP shortcuts, PW (T-LDP/BGP), VPLS (LDP/BGP), IP-VPN (BGP), MVPN (BGP/mLDP/P2MP RSVP)

- Issues:
  - Traffic-engineering solutions don’t scale when we want more granularity/dynamicity
  - Remote LFA for LDP is considered too complex: requires dynamic T-LDP signaling

NOKIA

MPLS: a historical perspective (2)

	LDP	RSVP-TE
Overview	Multipoint to point	Point to point
Operation	Simple	LSP per destination/TE-path
Dependencies	Relies on IGP	Relies on IGP TE
Label allocation	Locally significant per node (interface)	Locally significant per node (interface)
Traffic Engineering	No	Yes
Scaling	1 label per node (interface)	Nx(N-1)
Fast Reroute	LFA, LFA Policies, RLFA - <100% coverage	Link/Node protection (detour/facility) – 100% coverage
Multicast	mLDP	P2MP RSVP
IPv6	Extensions required	Extensions required

What problem are we trying to solve ?

SCALE

- Increasing network growth with granular traffic engineering (TE) requirement
- RSVP-TE is the only widely-spread solution to provide TE
- No LDP-TE available
- LDP Fast ReRoute (FRR) can be used in some parts of the network but is topology dependent

RSVP-TE

- Pros:
  - Source Routed protocol ; ingress Label Edge Router (iLER) has full control to setup LSP to destination
  - Presence of strong FRR and TE capabilities
- Cons:
  - Soft-state ; refresh mechanism required : refresh reduction (RFC2961) aggregates messages but not # soft-states
  - Mid-point state presence in network (with FRR) consumes CPU cycles and memory

NOKIA

Objective of Segment Routing

The primary objective for Segment Routing (SR) is source routing: the ability for a node to specify a unicast forwarding path, other than the normal shortest path, that a particular packet will traverse …

…without requiring mid-point state .

  Public

IETF SPRING working group

- SPRING (**S**ource **P**acket **R**outing **I**n **N**etworkin**G**) Working Group addresses the following:
  - IGP-based MPLS tunnels without the addition of any other signaling protocol
    - The ability to tunnel services (VPN, VPLS, VPWS) from ingress PE to egress PE with or without an explicit path, and without requiring forwarding plane or control plane state in intermediate nodes.
  - Fast Reroute
    - Any topology, pre-computation and setup of backup path without any additional signaling.
    - Support of shared-risk constraints, support of link/node protection, support of micro-loop avoidance.

IETF SPRING working group

- SPRING (**S**ource **P**acket **R**outing **I**n **N**etworkin**G**) Working Group addresses the following:
  - Traffic Engineering
    - The soft-state nature of RSVP-TE exposes it to scaling issues; particularly in the context of SDN where traffic differentiation may be done at a finer granularity.
    - Should include loose/strict options, distributed and centralised models, disjointness, ECMP-awareness, limited (preferably zero) per-service state on midpoint and tail-end routers.
  - All of this should allow incremental and selective deployment with minimal disruption

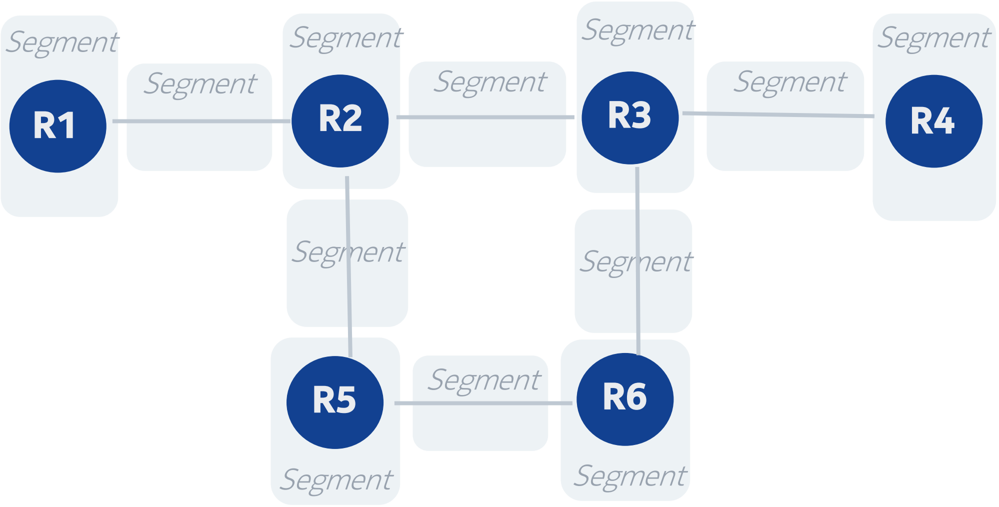
IETF SPRING working group

- Data plane support required:
  - Leverage the existing **MPLS** dataplane without any modification
    - MPLS label stack imposition
    - MPLS label operations: pop, swap, push, PHP
  - Leverage the **IPv6** dataplane with a new IPv6 Routing Header Type (Routing Extension Header)

  Public

Introduction to Segment Routing

- Segment Routing provides a tunneling mechanism that enables source routing.
- Paths are encoded as sequences of topological sub-paths called segments, which are advertised by link-state routing protocols (IS-IS and OSPF).



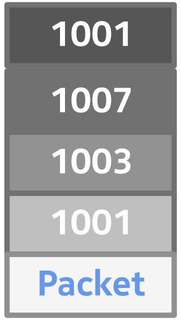
  Public

NOKIA

Encoding Segment Routing tunnels

- A Segment Routing (SR) tunnel, containing a single segment or a segment list, is encoded as:
  - A single MPLS label or an ordered list of hops represented by a stack of MPLS labels (no change to the MPLS data-plane).
  - A single IPv6 address, or an ordered list of hops represented by a number of IPv6 addresses in the IPv6 Extension header (Segment Routing Header).
- The segment list can represent either a topological path (node, link) or a service.

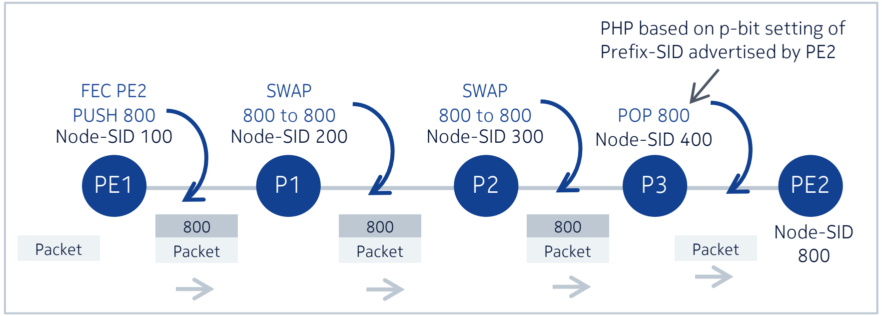
> The segments can be thought as a set of instructions from the ingress PE such as “go to node D using the shortest path”, “go to node D using link/node/explicit-route L”



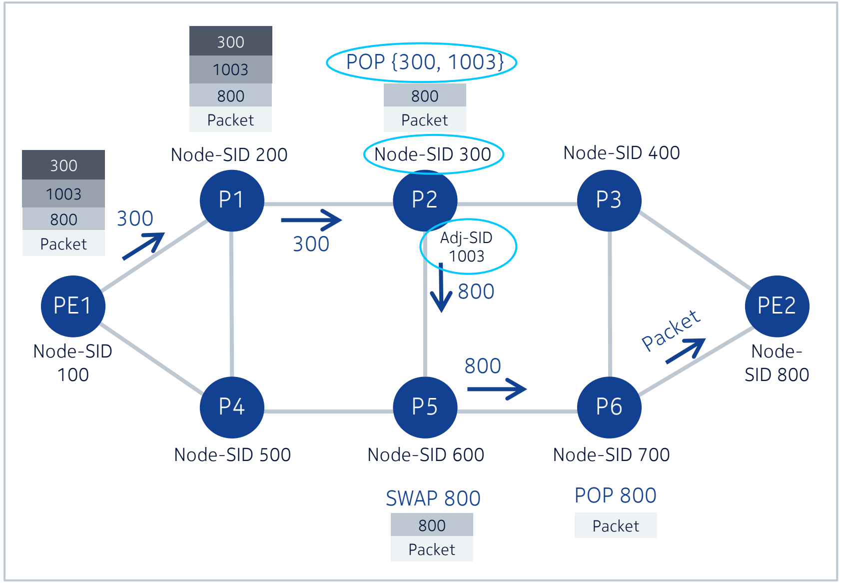
  Public

Segment Routing
A simple operational overview

Tunnel with node segment (NODE-SID)



SR tunnel with node and adjacency (ADJ-SID) segments



  Public

NOKIA

Operations on segments

- Three distinct operations:
  - **PUSH**: the insertion of a segment at the head of the Segment list.
  - **NEXT**: the active segment is completed; the next segment becomes active.
  - **CONTINUE**: the active segment is not completed and hence remains active.

Segment routing with MPLS data plane (1)

- MPLS instantiation of Segment Routing aligns with the MPLS architecture defined in RFC 3031
- For each segment, the IGP advertises an identifier referred to as a Segment ID (SID). A SID is a 32-bit entity; with the MPLS label being encoded as the 20 right-most bits of the segment ID.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+ Label
|                Label                  | Exp |S|       TTL     | Stack
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+ Entry

                    Label:  Label Value, 20 bits
                    Exp:    Experimental Use, 3 bits
                    S:      Bottom of Stack, 1 bit
                    TTL:    Time to Live, 8 bits
```

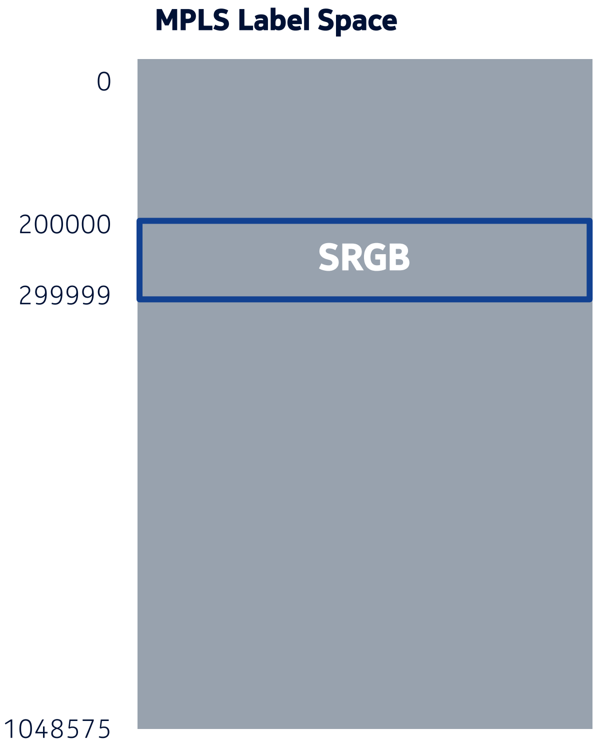
  Public

Segment routing with MPLS data plane (2)

- When Segment Routing is instantiated over the MPLS data-plane, the following actions apply :
  - A list of segments is represented as a stack of labels
  - The active segment is the top label
  - The **CONTINUE** operation is implemented as a **SWAP** operation
  - The **NEXT** operation is implemented as a **POP** operation
  - The **PUSH** operation is implemented as a **PUSH** operation

Segment routing with MPLS data plane (3)
Segment Routing Global Block (SRGB)

- Segment Routing Global Block (SRGB)
  - SRGB is the set of local labels reserved for global segments
  - Local property of an SR node
  - Using the same SRGB on all nodes within the SR domain ease operations and troubleshooting and is expected to be a deployment guideline.



Types of segments
Taxonomy

- **IGP SEGMENTS**
  - **Prefix Segments**
    - **Node Segments**
    - **Anycast Segments**
  - **Adjacency Segments**
- **BGP SEGMENTS**
  - **Prefix Segments**
  - **Egress Peer Engineering (EPE) Segments**
    - **PeerNode Segments**
    - **PeerAdj Segments**
    - **PeerSet Segments**

Types of segments

Prefix Segment

- Globally unique – allocated from SRGB
- Typically multi-hop
- ECMP-aware shortest-path IGP route to a related prefix
- Indexing or absolute SID
- Signaled by IGP

Adjacency Segment

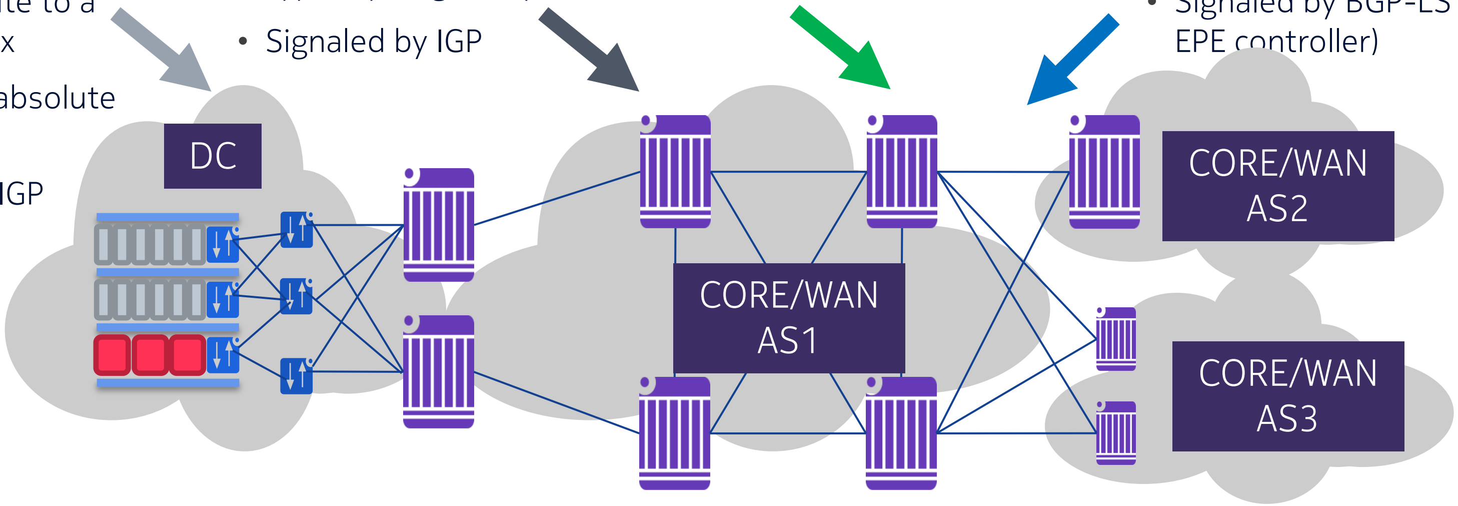
- Locally unique – each SR router in the domain can use the same space
- Typically single-hop
- Signaled by IGP

BGP Prefix Segment

- Example: Prefix Segment in DC environment
- DC GW representation
- Signaled by BGP (in DC)

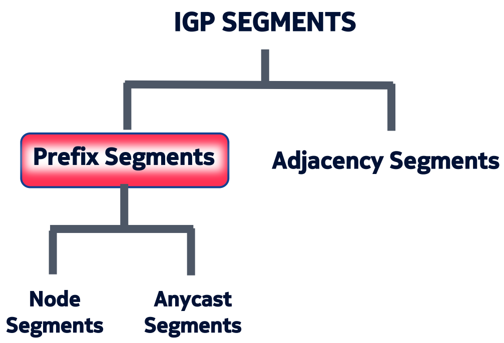
BGP Peer Segment

- EPE ; Egress Peering Engineering
- Influence how to control traffic to adjacent AS
- Signaled by BGP-LS (w/ EPE controller)



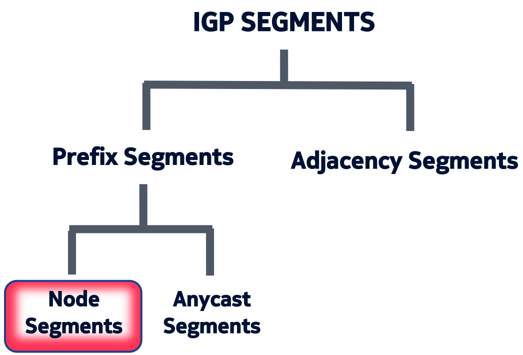
  Public NOKIA

Segment identifiers – prefix segments



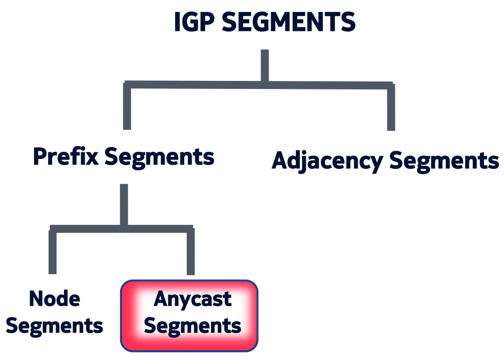
- Prefix Segment (Prefix-SID)
  - Globally unique within the IGP/SR domain – allocated from the SR Global Block (SRGB)*
  - Represents the ECMP-aware shortest-path IGP route to the related prefix
  - Typically a multi-hop path
  - Includes “P” flag to allow neighbours to perform the “NEXT” (pop) operation whilst processing the segment (analogous to Penultimate Hop Popping in MPLS).
  - Two options exist; Indexing or Absolute-SID (described in later slides)

Segment identifiers – node segments



- Node Segment ID (Node-SID)
  - A special prefix segment used to identify a specific router (loopback/system address).
  - Identified by “N” flag being set in advertised segment (Prefix-SID Sub-TLV).
  - Represents the ECMP-aware shortest-path IGP route to the specified node.

Segment identifiers – anycast segments

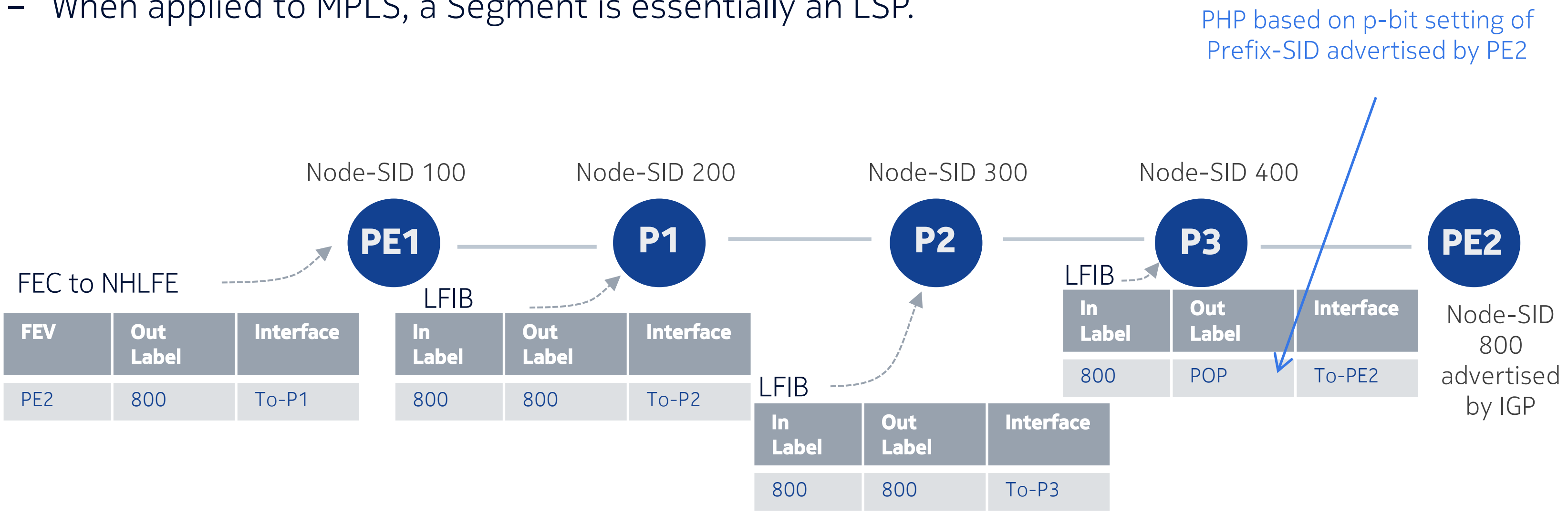


- Anycast Segment ID (Anycast-SID)
  - A prefix segment specifying a set of routers
  - Represents the ECMP-aware shortest-path IGP route to the closest node of the "anycast set".
  - Potentially useful for coarse traffic engineering (i.e. route via plane A of dual-plane network, route via Region B of multi-region network) or node redundancy (i.e. traffic re-routes to shortest path towards any other router that is part of the "anycast set").

NOKIA

Example: SR tunnel with prefix-SID (node-SID) [1]

- PE2 advertises Node Segment into IGP (Prefix-SID Sub-TLV Extension to IS-IS/OSPF)
- All routers in SR domain install the node segment to PE2 in the MPLS data-plane.
  - No RSVP and/or LDP control plane required.
  - When applied to MPLS, a Segment is essentially an LSP.



FEC to NHLFE (PE1)

FEV	Out Label	Interface
PE2	800	To-P1

LFIB (P1)

In Label	Out Label	Interface
800	800	To-P2

LFIB (P2)

In Label	Out Label	Interface
800	800	To-P3

LFIB (P3)

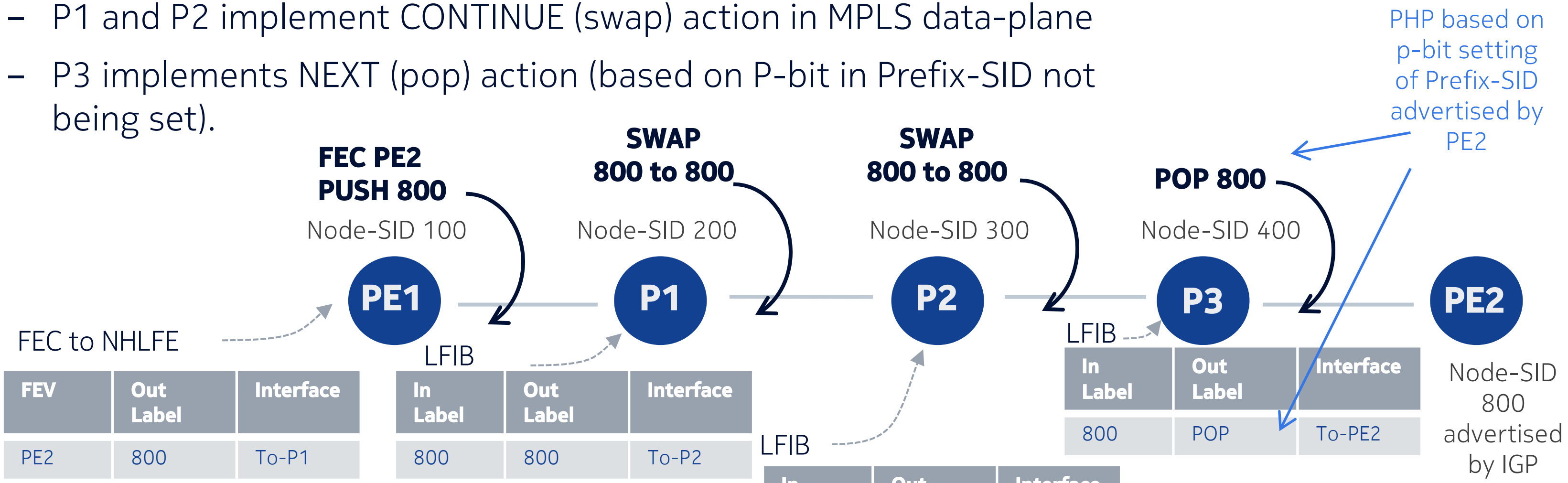
In Label	Out Label	Interface
800	POP	To-PE2

  Public

NOKIA

Example: SR tunnel with prefix-SID (node-SID) [2]

- For traffic from PE1 to PE2, PE1 pushes on node segment {800} and uses shortest IGP path to reach PE2.
- Active segment is the top of the stack for MPLS:
  - P1 and P2 implement CONTINUE (swap) action in MPLS data-plane
  - P3 implements NEXT (pop) action (based on P-bit in Prefix-SID not being set).



FEC to NHLFE

FEV	Out Label	Interface
PE2	800	To-P1

LFIB (P1)

In Label	Out Label	Interface
800	800	To-P2

LFIB (P2)

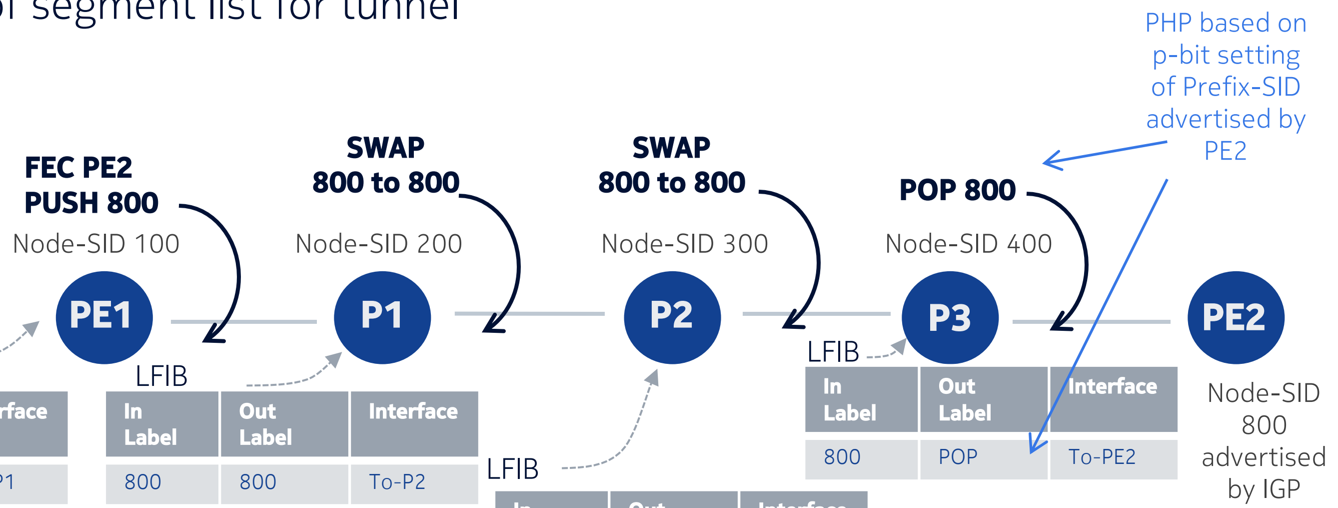
In Label	Out Label	Interface
800	800	To-P3

LFIB (P3)

In Label	Out Label	Interface
800	POP	To-PE2

Example: SR tunnel with prefix-SID (node-SID) [3]

- No per-path state held in network with the exception of segment list for tunnel held at PE1.



FEC to NHLFE

FEV	Out Label	Interface
PE2	800	To-P1

LFIB

In Label	Out Label	Interface
800	800	To-P2

LFIB

In Label	Out Label	Interface
800	800	To-P3

LFIB

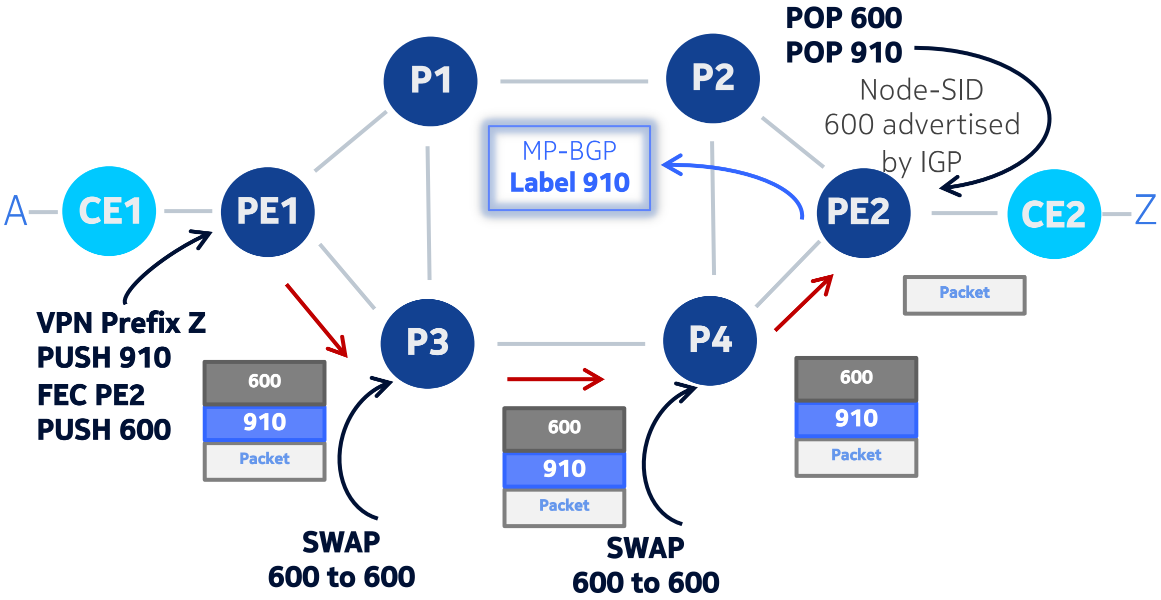
In Label	Out Label	Interface
800	POP	To-PE2

  Public

NOKIA

Prefix segment identifiers – absolute SIDs

- The use of absolute SID values requires a single consistent SRGB on all SR routers throughout the IGP domain.
- Example:
  - PE2 advertises MP-BGP label 910 for VPN prefix Z.
  - To forward traffic to VPN prefix Z, and assuming preferred (non-ECMP) path from PE1 to PE2 is PE1-P3-P4-PE2, PE1 pushes label 910 onto bottom of stack, and label 600 (Node-SID for PE2) on top of stack.
  - Label (SID) does not change hop by hop.



  Public

Prefix SID indexing

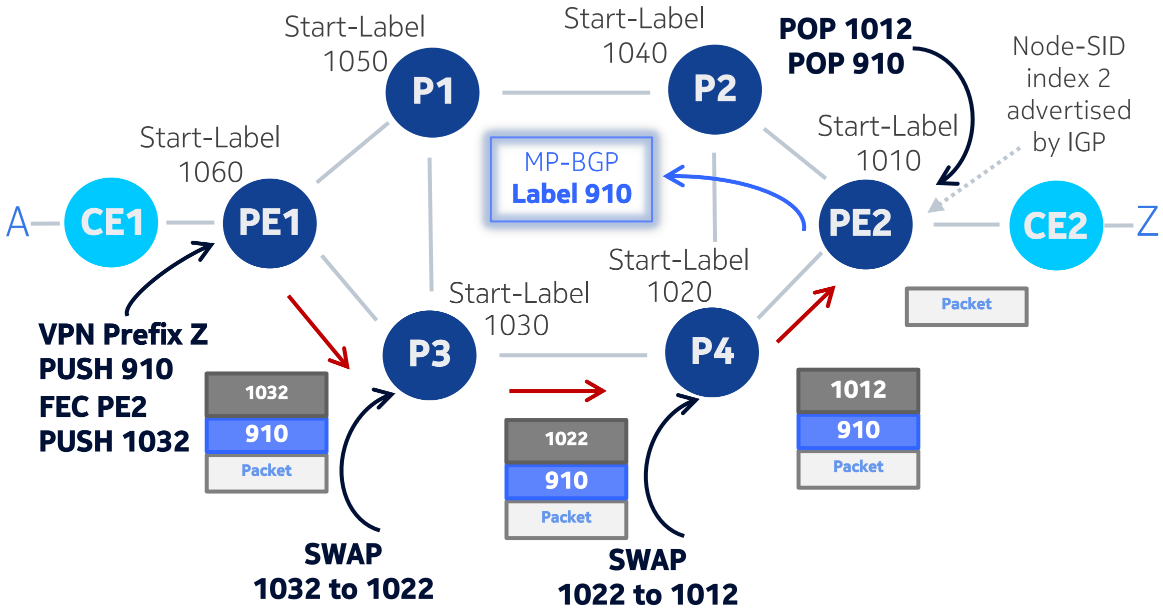
- Why ?
  - SR domain can be multi-vendor with the possibility that each vendor uses a different MPLS label range
  - Prefix SID must be globally unique within SR domain

- How ?
  - Indexing mechanism is required for prefix SIDs. All routers within the SR domain are expected to configure and advertise the same prefix SID index range for a given IGP instance.
  - The label value used by each router to represent a prefix ‘Z’ (= label programmed in ILM) can be local to that router by the use of an offset label, referred to as a start label :

    Local Label (for Prefix SID) = (local) start-label + {Prefix SID index}

Example: Prefix SID indexing

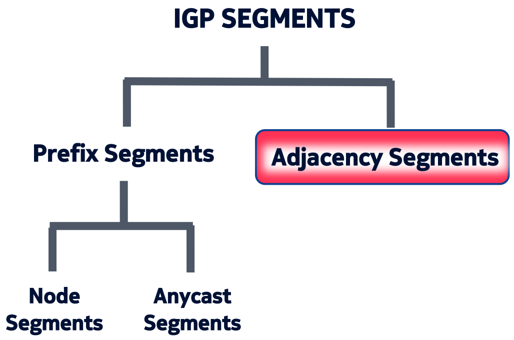
- For example, assume the SID Index Range is {1,100}.
- Each SR router in the domain defines a start point in the SRGB (start-label), and an offset label called an SID index.
  - SR routers sum {start-label + SID index} to obtain a local label for a Prefix SID.
  - Assuming PE2 advertises loopback 192.0.2.2/32 with a prefix index of 2:
  - PE2's SID for itself is {1010+2}= 1012
  - P4's SID for PE2 is {1020+2}= 1022
  - P3's SID for PE2 is {1030+2}=1032
  - PE2 advertises MP-BGP label 910 for VPN prefix Z.



  Public

NOKIA

Segment identifiers – adjacency segments

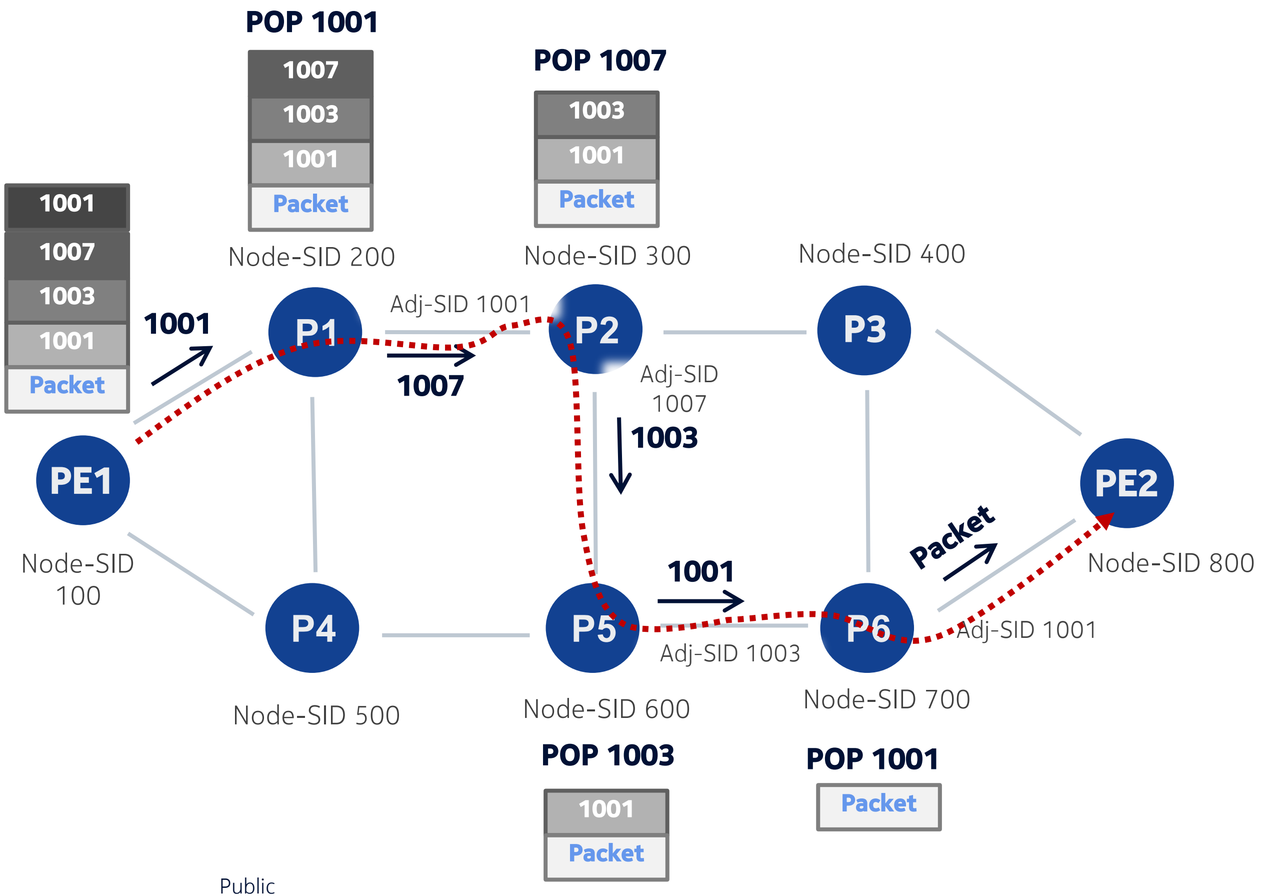


- Adjacency Segment ID (Adj-SID)
  - A segment identifying an adjacency or set of adjacencies that must be in the IGP.
  - Segment Identifier (SID) is local to the router that advertises it (every SR router in the domain can use the same segment space).
  - If:
    - X is the Node-SID of node N, and…
    - Y is an Adj-SID at node N to an adjacency over link L, then….
    - A packet with segment list {X, Y} will be forwarded along the shortest-path to node N, then switched by N towards link L without any consideration of shortest-path routing.
  - If the Adj-SID identifies a set of adjacencies, node N can load-balance the traffic over the members of that set.

Segment identifiers – adjacency segments

- All SR routers advertise Adjacency segment(s) into IGP (Adjacency-SID Sub-TLV Extension to IS-IS/OSPF).
- Adjacency segments may be of local or global significance, but only the advertising SR router installs the adjacency segment into the MPLS data-plane
  - *From a data-path perspective, it is analogous to a label-swap to implicit-null.*
- Provides for end-to-end source-routing capability where the Adjacency segments may determine the explicit hop-by-hop path through the network.
- Beware however, that label stack depth has implications on hardware.

Example: SR tunnel with adjacency segments



  Public

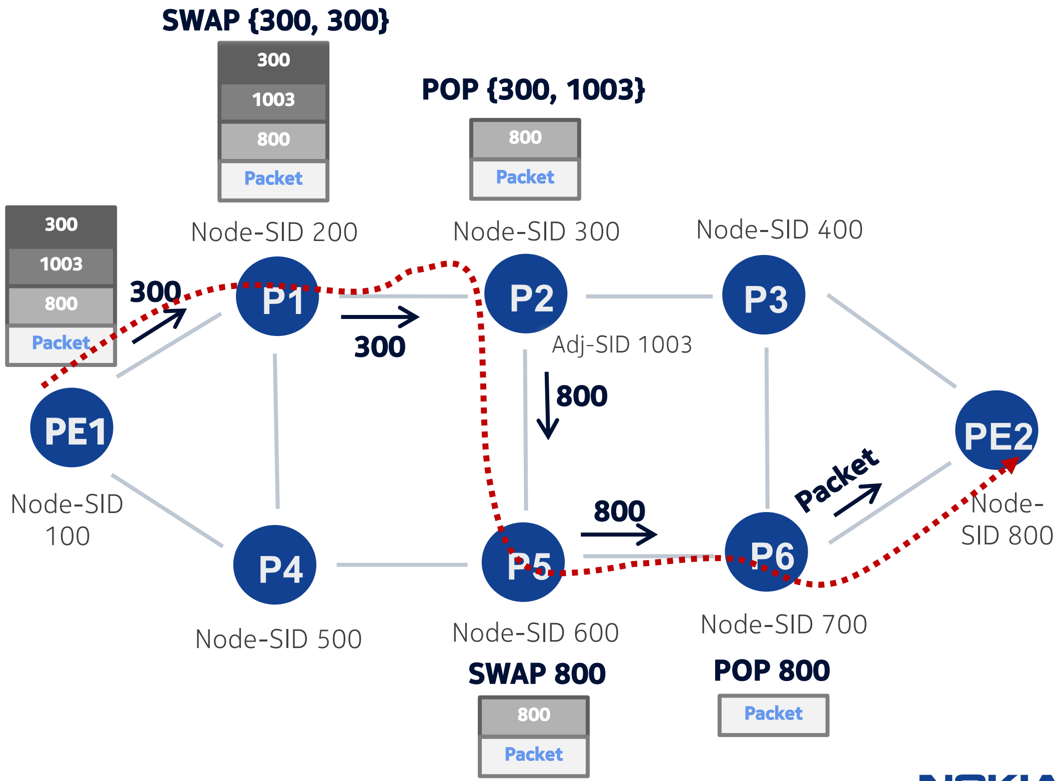
NOKIA

Example: SR tunnel with node and adjacency segments

- A combination of node and adjacency segments is also possible.
- This provides the ability to exercise ECMP paths to the next specified node segment, but enforce the use of a particular link (or links) from that node.

Example: SR tunnel with node and adjacency segments

- In this example, PE1 wants to traverse the link P2-P5 on the way to PE2, as it is under-utilised.
- PE1 therefore imposes the segment list {300, 1003, 800} representing the Node-SID for P2, the Adj-SID for link P2-P5, and finally the Node-SID for PE2.



  Public

NOKIA

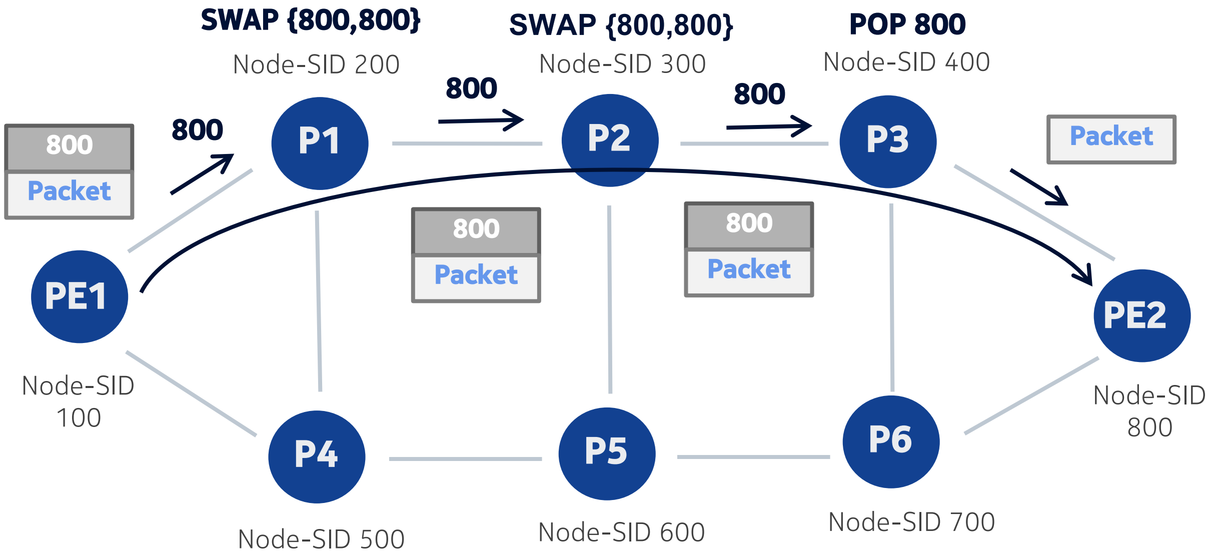
Comparison with LDP and RSVP-TE

	LDP	RSVP-TE	SR
Overview	Multipoint to point	Point to point	Multipoint to point
Operation	Simple	LSP per destination/TE-path	**Simple**
Dependencies	Relies on IGP	Relies on IGP TE	Relies on IGP + offline TE
LBL allocation	Local significant per node (interface)	Local significant per node (interface)	**Global**
Traffic Engineering	No	Yes	**yes**
Scaling	1 LBL per node (interface)	Nx(N-1)	**1 LBL per node/ local interface**
Fast Reroute	LFA, LFA Policies, RLFA - <100% coverage	Link/Node protection (detour/facility) – 100% coverage	LFA, LFA Policies, RLFA/DLFA - can get to **100% coverage** (better than LDP with RLFA)
Multicast	mLDP	P2MP RSVP	TBD
IPv6	Extensions required	Extensions required	**Native**

use cases and applicability

Use case 1: Shortest path routing (1)

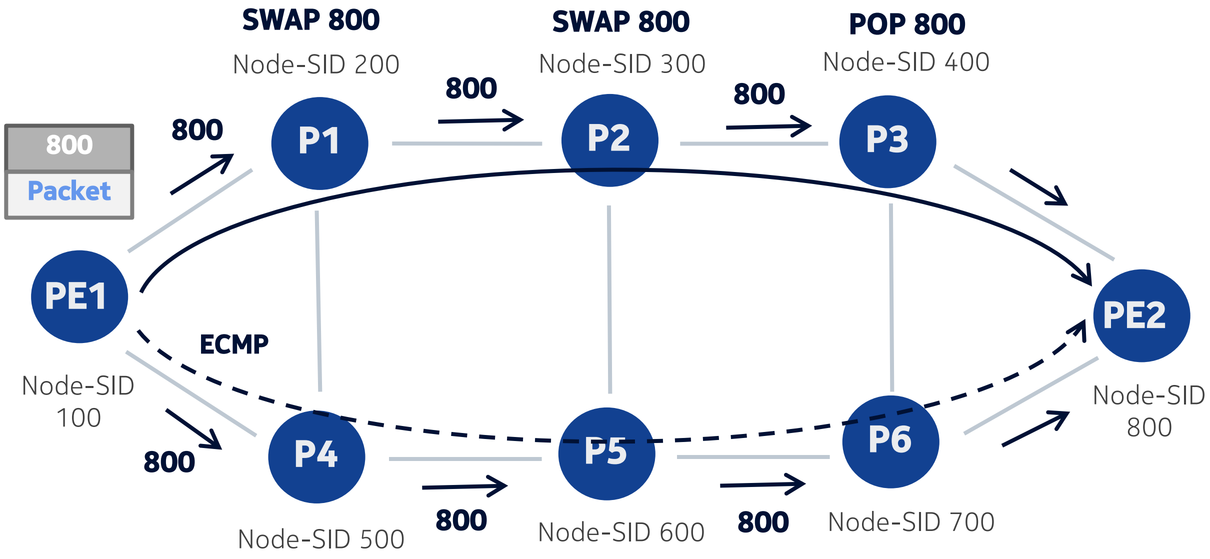
- All nodes advertise a unique node segment into the IGP.
- For traffic from PE1 to PE2, PE1 pushes on segment list {800} and uses shortest IGP path to reach PE2 (PE1-P1-P2-P3-PE2)
  - P1 and P2 install ILM CONTINUE entry {label=800, NHLFE=label 800, Next-Hop=shortest path to PE2}
  - P3 installs ILM CONTINUE or NEXT entry {label=800, POP, Next-Hop=shortest path to PE2}



  Public

Use case 1: Shortest path routing (2)

- If PE1 has ECMP>=2, and equal-cost paths to the SR tunnel tail-end exist, all equal-cost paths can be exercised:
  - Based on hash output, flows m routed PE1-P1-P2-P3-PE2 with segment list {800}
  - Based on hash output, flows n routed PE1-P4-P5-P6-PE2 with segment list {800}

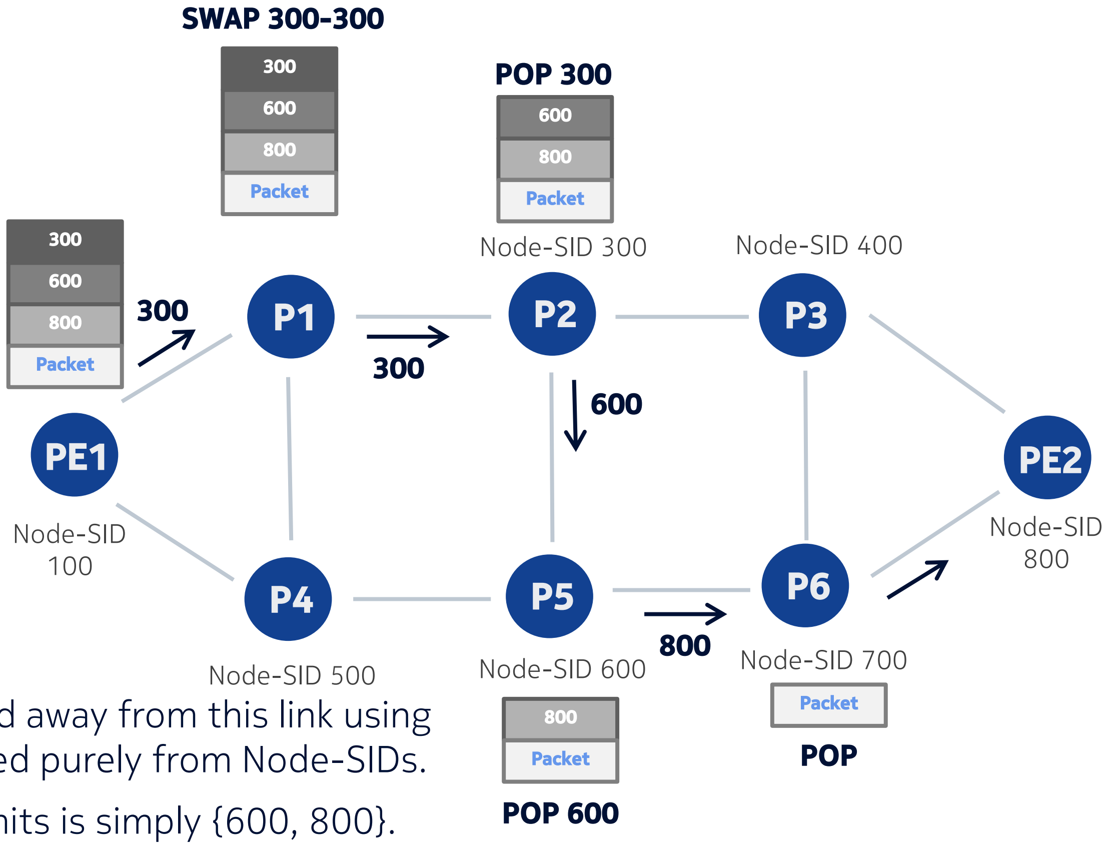


  Public

NOKIA

Use case 2: Source-routing with node-SID

- Adj-SID provides the capability to explicit-route on a hop-by-hop basis, but has the potential to create a deep label stack-depth if all hops are explicitly listed.
- Assume we have a requirement to engineer traffic away from the P2-P3 link (due to high utilisation or link degradation) to some other under-utilised link(s).



- Traffic from PE2 to PE1 can be re-routed away from this link using segment list {300, 600, 800} constructed purely from Node-SIDs.
- Alternative option if link utilisation permits is simply {600, 800}.

  Public

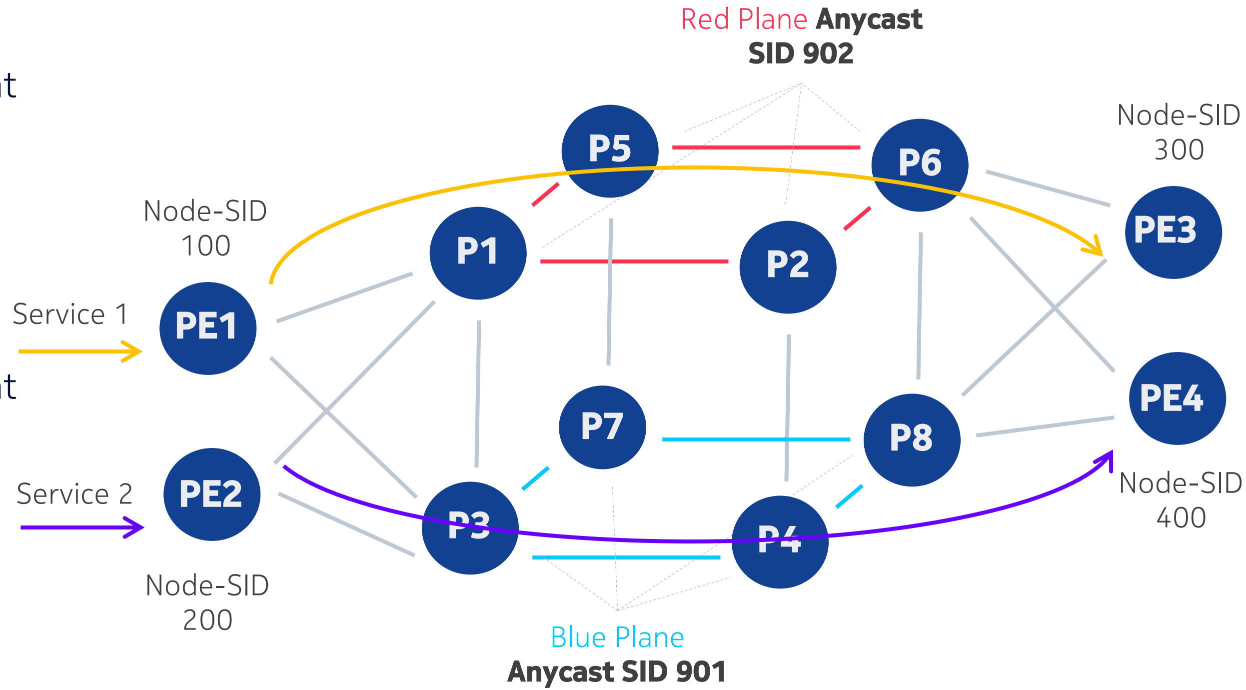
NOKIA

Use case 3: Disjointness (1)

- Disjointness describes two (or more) services that must be completely disjoint of each other. They should not share common network infrastructure – i.e. if one fails, the other must always be active.
- Many networks employ the ‘dual-plane’ design, where inter-plane links are configured such that the route to a destination stays on that plane during a single failure scenario.
- Disjointness can broadly be achieved using Anycast segments.

Use case 3: Disjointness (2)

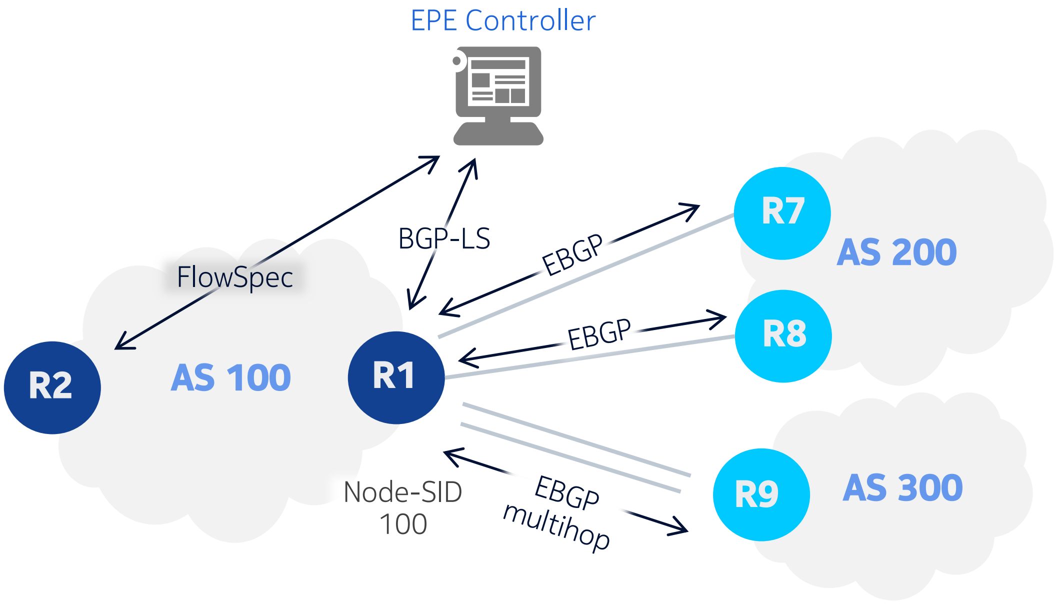
- Assume service 1 between PE1 and PE3 must be disjoint from service 2 between PE2 and PE4:
  - Service 1 at PE1 has segment list {902, 300} including Anycast SID 902 and traverses the red plane before reaching PE3.
  - Service 2 at PE2 has segment list {901, 400} including Anycast SID 901 and traverses the blue plane before reaching PE4.



  Public NOKIA

Use case 4: Egress peer engineering (EPE) (1)

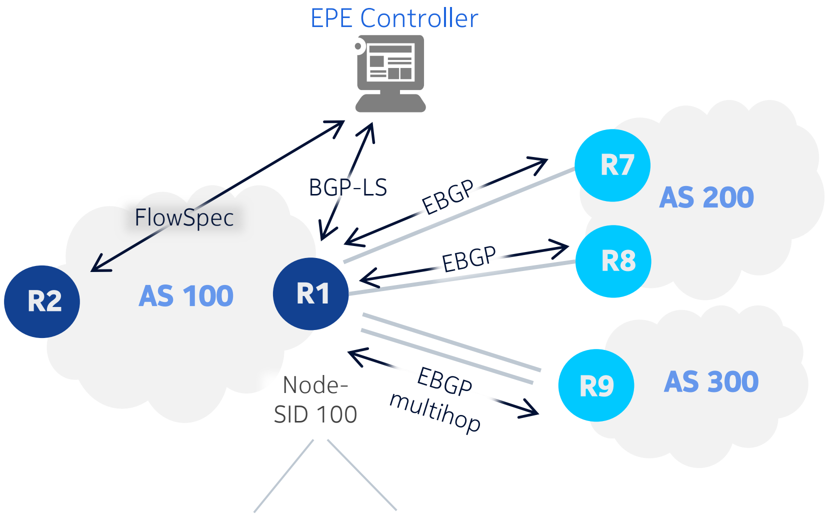
- Egress Peer Engineering defines three BGP Peering SIDs, that allow for programming of source-routed inter-domain paths; PeerNodeSID, PeerAdjSID, and PeerSetSID.
- R1 is an EPE-enabled egress router and allocates the following:
  - PeerNode segment for each of its defined peers (R7, R8, and R9)
  - PeerAdj segment for each recursive interface to a multi-hop peer (R9)
  - PeerSet segment to a set of peers (R7 and R8) (AS200)



NOKIA

Use case 4: Egress peer engineering (EPE) (2)

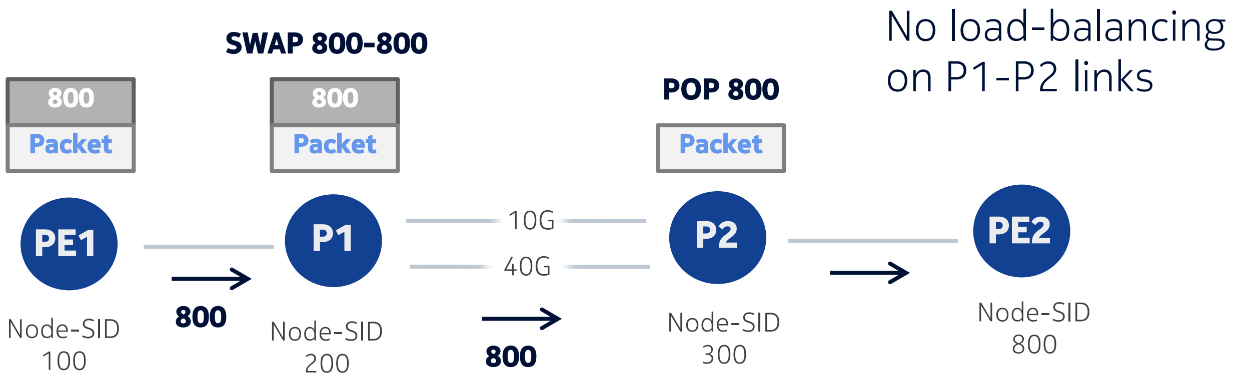
- BGP-LS (BGP Link State) session established between EPE-enabled border router (R1) and the EPE controller:
  - R1 advertises PeerNode, PeerAdj, and PeerSet SIDs using SR extensions to BGP-LS, and programmes FIB accordingly.
- EPE Controller programmes source-routes from ingress routers to EBGP peers using FlowSpec/OpenFlow; i.e.
  - 80% traffic to AS 300 with segment list {100, 1005}
  - 20% traffic to AS 200 with segment list {100, 1006}
  - Prefix <NLRI/Length> segment list {100, 1003}
  - Prefix <NLRI/Length> segment list {100, 1004}



Incoming Label	Operation	Outgoing Interface
1001	POP	Link to R7
1002	POP	Link to R8
1003	POP	Upper link to R9
1004	POP	Lower link to R9
1005	POP	Load-balance on any link to R9
1006	POP	Load-balance on any link to R7 or R8

Use case 5: Adjacency segment load-balancing (1)

- In this example, two adjacencies exist between P1-P2.
- Assuming capacity-based metrics are in use, the 10G link between P1 and P2 is unused for shortest path forwarding.

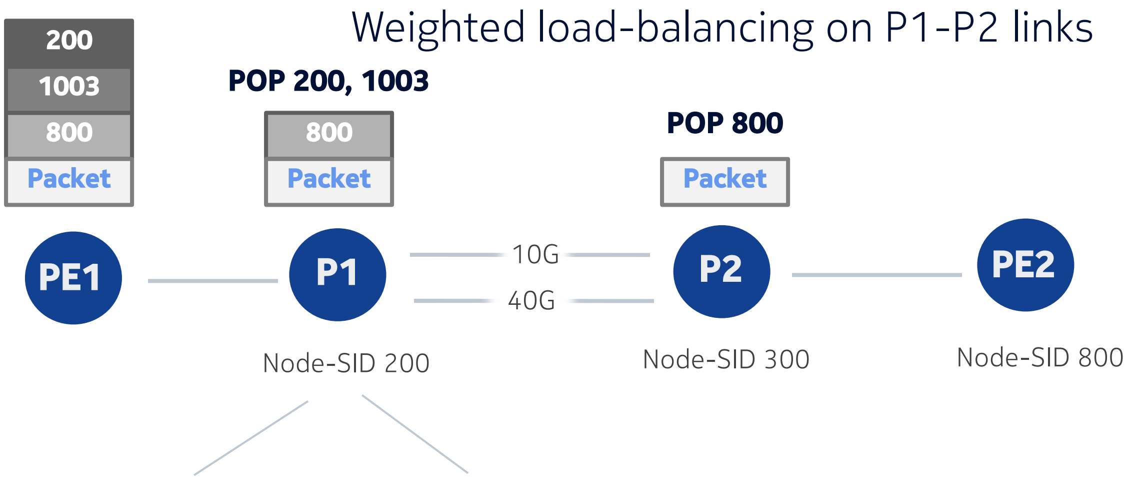


  Public

NOKIA

Use case 5: Adjacency segment load-balancing (2)

- Adj-SID TLV provides the capability to load-balance across multiple adjacencies.
  - P1 advertises individual Adj-SIDs for the 10G link (1001) with weight 1, and 40G link (1002) with weight 4.
  - P1 also advertises an Adj-SID for the adjacency set (1003)
  - PE1 pushes segment list {200, 1003, 800}. Node-SID 200 gets the traffic to P1, while Adj-SID 1003 load-balances the traffic to P2 on a weighted 4:1 basis.

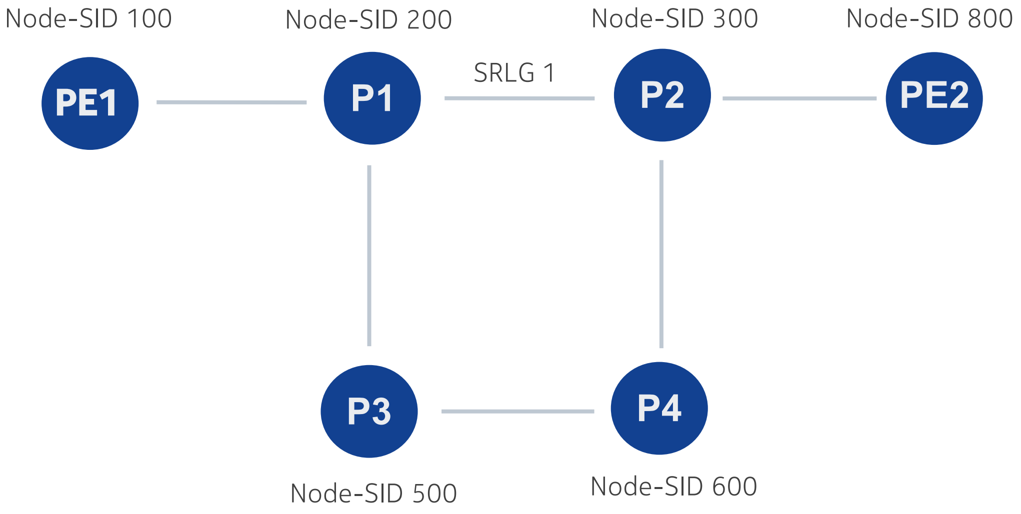


Link	Adj-SID	Adj-Set	Weight
10G	1001	1003	1
40G	1002	1003	4
Both	1003	-	-

NOKIA

Use case 6: Distributed cspf-based traffic engineering (1)

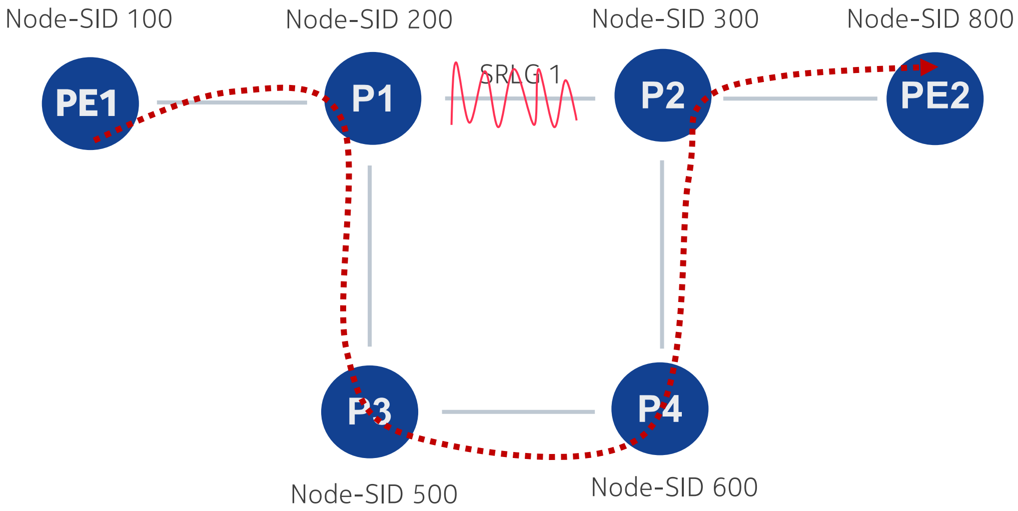
- Traffic Engineering information made available to CSPF for RSVP-TE based LSPs can also be made available to SR tunnels
  - Includes available link bandwidth, admin-groups, shared-risk link groups (SRLGs) etc.
- In the example topology, assume that link P1-P2 is in SRLG 1.
  - The SRLG information is flooded into IS-IS (RFC 4874) or OSPF (RFC 4203).



  Public

Use case 6: Distributed cspf-based traffic engineering (2)

- If PE1 computes a CSPF to PE2 for a path that should avoid SRLG 1, it first prunes the links signalled as belonging to that SRLG (i.e. link P1-P2) from the topology.
- From the remaining topology, it computes a path – in this simple case, the path PE1-P1-P3-P4-P2-PE2.
- PE1 therefore imposes the segment list {200, 500, 600, 300, 800}, or even {500, 600, 800}.



Use case 7: Seamless MPLS and segment routing (1)

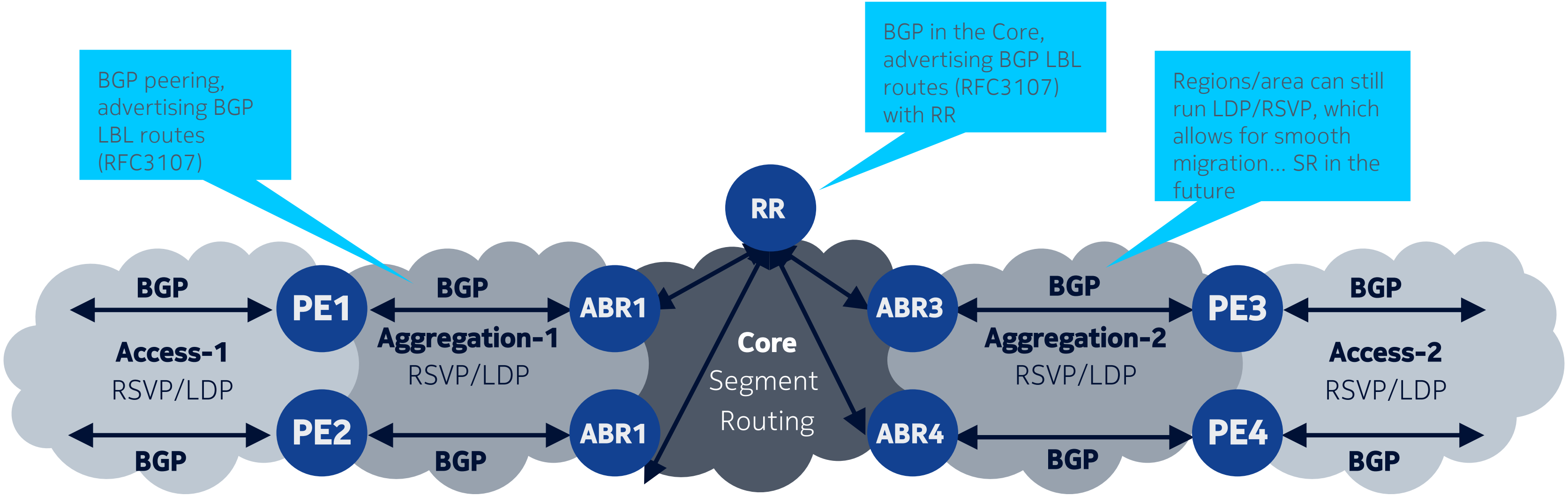
End-to-end scaling integrating SR/LDP/RSVP-TE

- SR can be seen as alternative for LDP and RSVP-TE. This means that the same scaling requirements will remain in case of an E2E MPLS coverage in a multi-area/instance domain.

- Seamless MPLS could be used to cross area or AS boundary, similar to what is available today with LDP and/or RSVP-TE. This approach has some clear advantages:
  - Smooth migration with existing MPLS domains
  - BGP is a field-proven scalable protocol
  - Non-SR nodes can still connect to a SR MPLS domain

Use case 7: Seamless MPLS and segment routing (2)

End-to-end scaling integrating SR/LDP/RSVP-TE

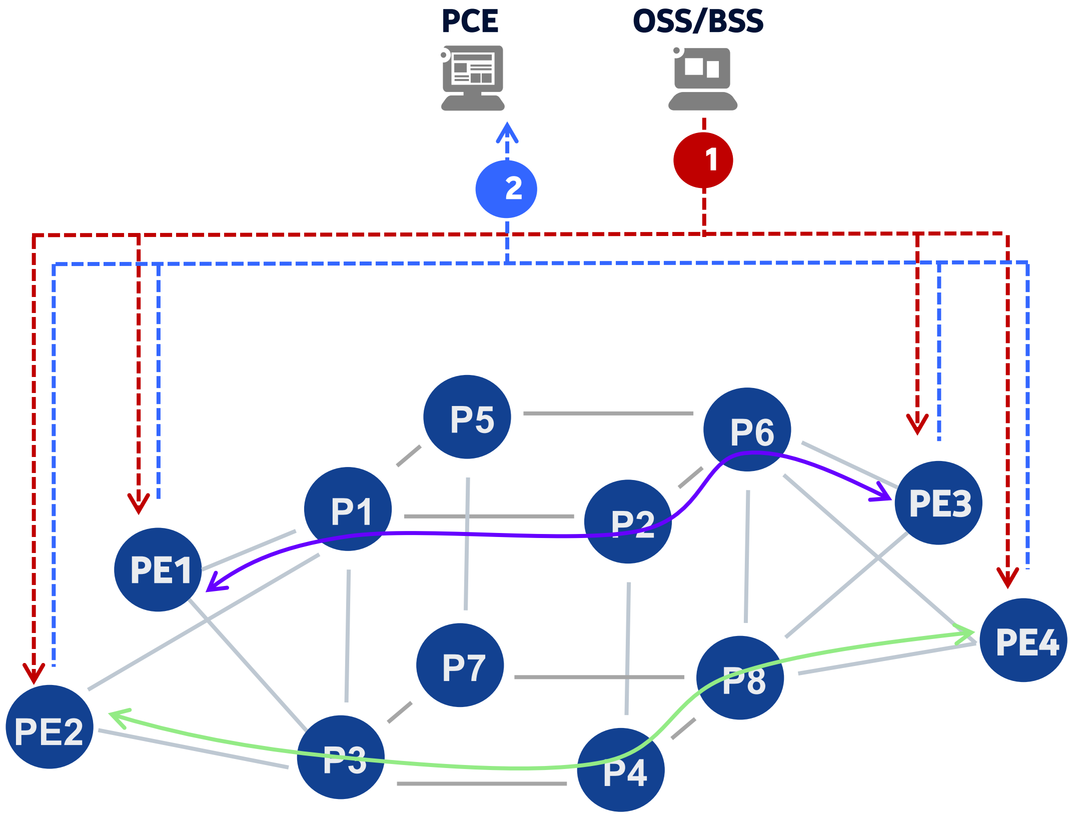


  Public

NOKIA

Use case 8: Service creation with a path computation element (PCE)

Co-routed service node provisioning



Step 1

OSS provisions **diverse** services on PE's

a. Type of service: VPWS
b. Local attachment circuits
c. Tunnel endpoints: Remote and Local
d. Tunnel type: Segment Routing, RSVP
e. Path constraints: Bandwidth, Co-routed, service diversity, bi-directionality

Step 2

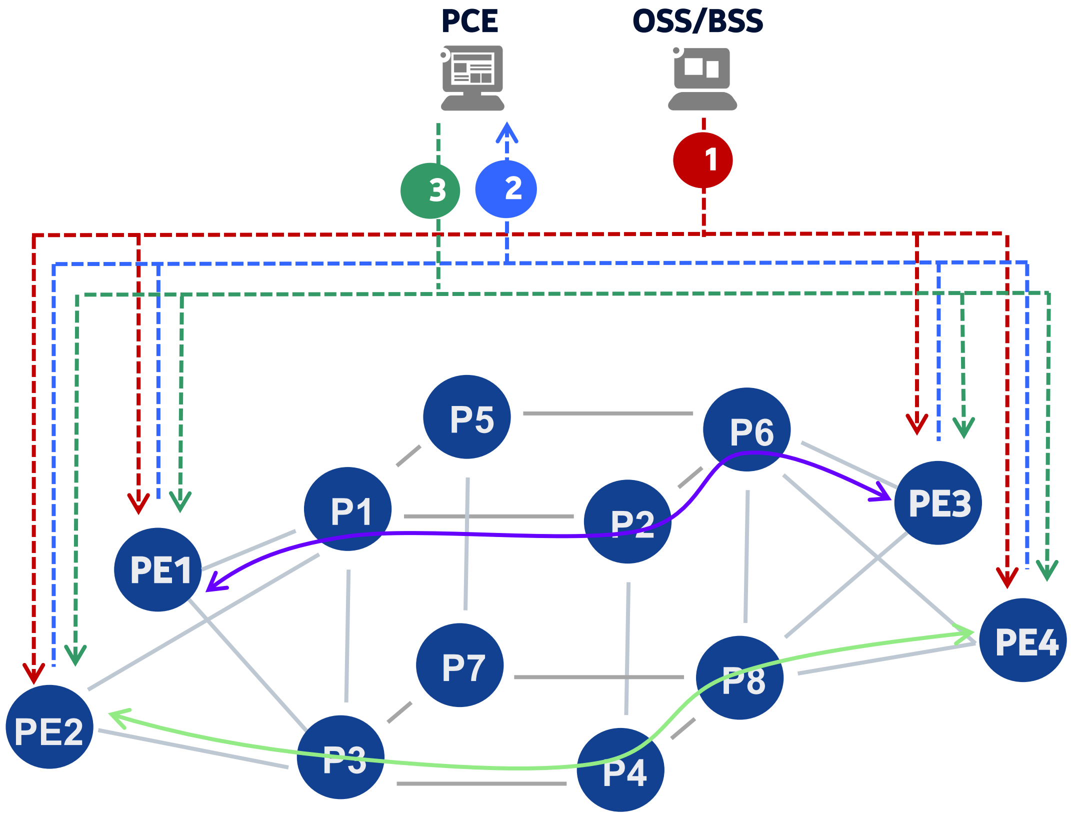
a. PE makes path computation request (PCReq), or path computation status report (PCRpt) to the PCE server
b. Requires PCE extensions to signal path diversity with other services.

  Public

NOKIA

Use case 8: Service creation with a path computation element (PCE) (cont.)

Co-routed service node provisioning



Step 3

a. PCE computes and downloads the paths for the tunnel set.
b. PEs bind service to paths

Step 4

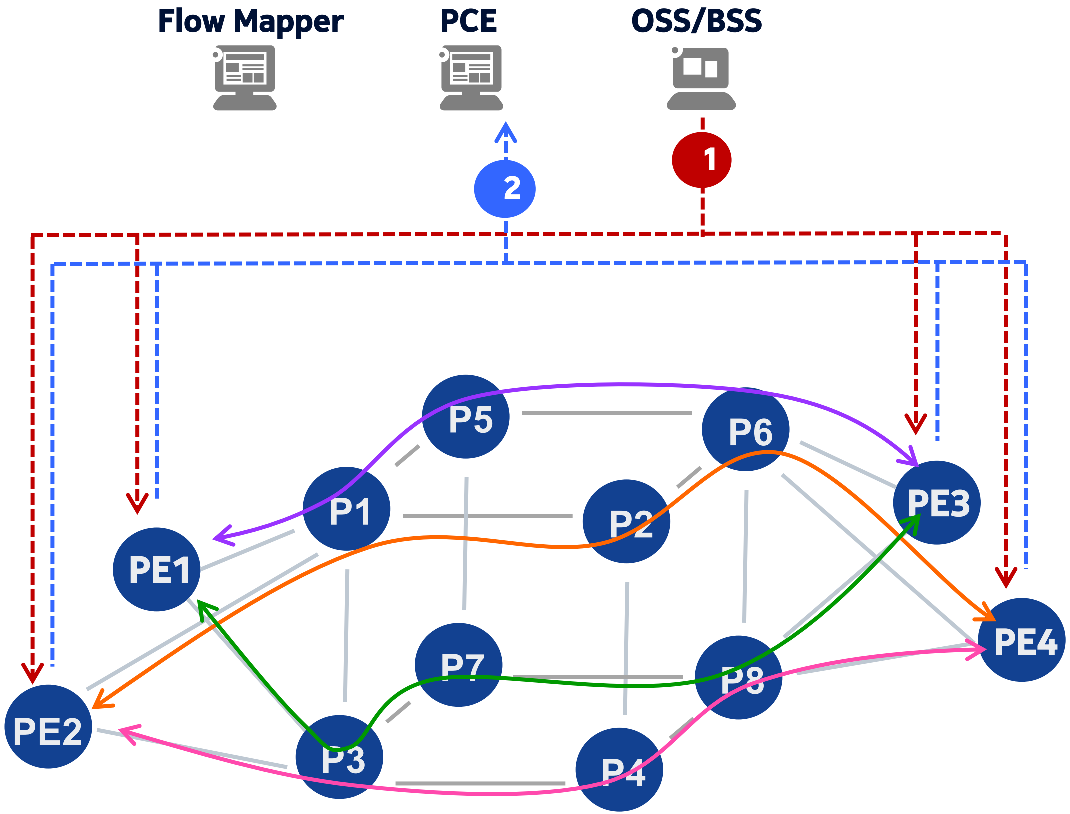
a. PCE monitors LSP stats and re-optimises tunnels as required, downloading new paths to PE routers (same PLSP-ID)
b. PE performs make-before-break and moves to the new path.

  Public

NOKIA

Use case 9: Service creation with a path computation element (PCE)

Global bandwidth optimisation



Step 1

OSS provisions parallel infrastructure tunnels between a pair of PE nodes

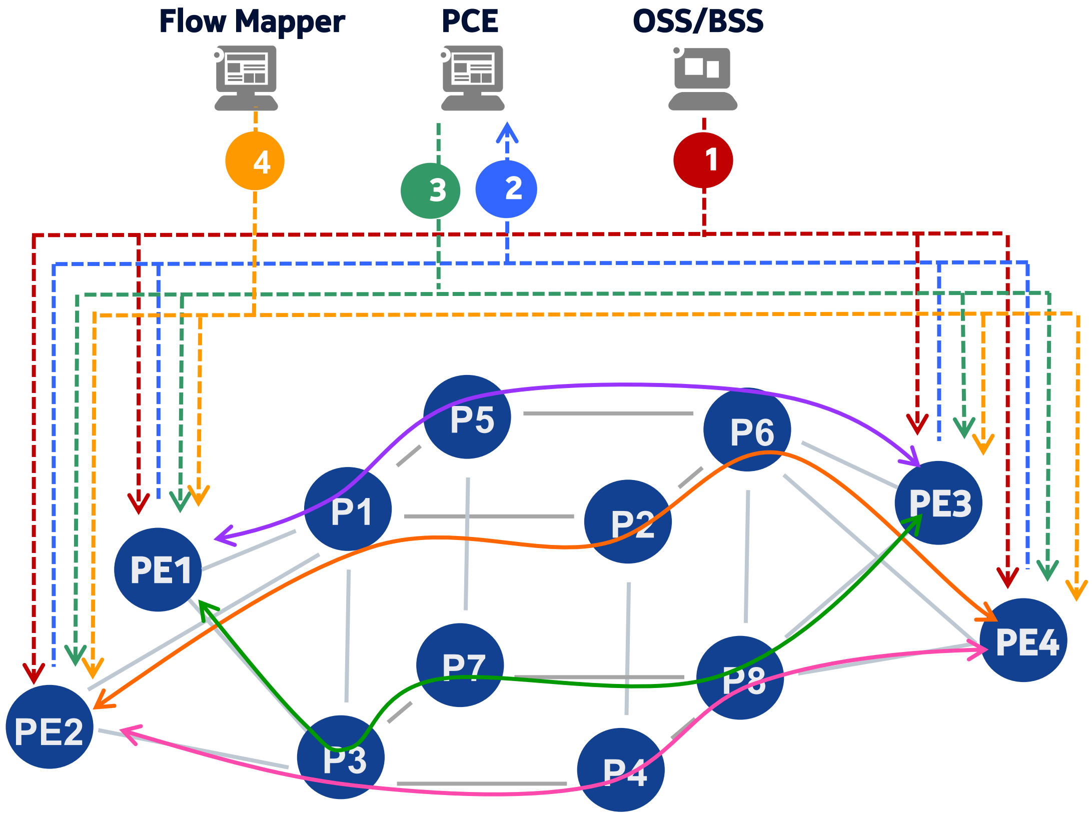
a. Type of service: VPRN, VPLS VPWS
b. Local attachment circuits
c. Tunnel endpoints: Remote and Local
d. Tunnel type: Segment Routing, RSVP
e. Path constraints: min/max bandwidth, diversity, admin-group

Step 2

a. PE makes path computation request (PCReq), or path computation status report (PCRpt) to the PCE server with path diversity constraints among the parallel set of tunnels.

Use case 9: Service creation with a path computation element (PCE)

Global bandwidth optimisation



Step 3

a. PCE computes and downloads the path.

b. PE node informs the external flow mapper of the set of LSP-ID values created between endpoints.

Step 4

a. External flow mapper pushes down the mapping of flow/prefix/destination to the set of parallel tunnels using OpenFlow, XMPP etc.

b. PE instantiates the flow policies to map each flow to the designated LSP-ID.

NOKIA

deployment options

Deployment options

Two broad categories

- Greenfields:
  - Relatively straightforward
  - Requires “new” software with segment routing capabilities
  - Opportunity to bypass LDP or RSVP-TE altogether
  - Care needs to be taken to ensure that all service types, resiliency mechanisms and traffic-engineering capabilities can be supported over segment-routed tunnels

- Existing networks:
  - Similar to greenfields with added considerations:
    - Ability to introduce without disruption to existing services
    - Co-existence with LDP and/or RSVP-TE where deployed; “ships in the night” operation required
    - Option to only build new services with segment routed tunnels, leaving existing services on existing tunnels
    - Migration to an SR-only network

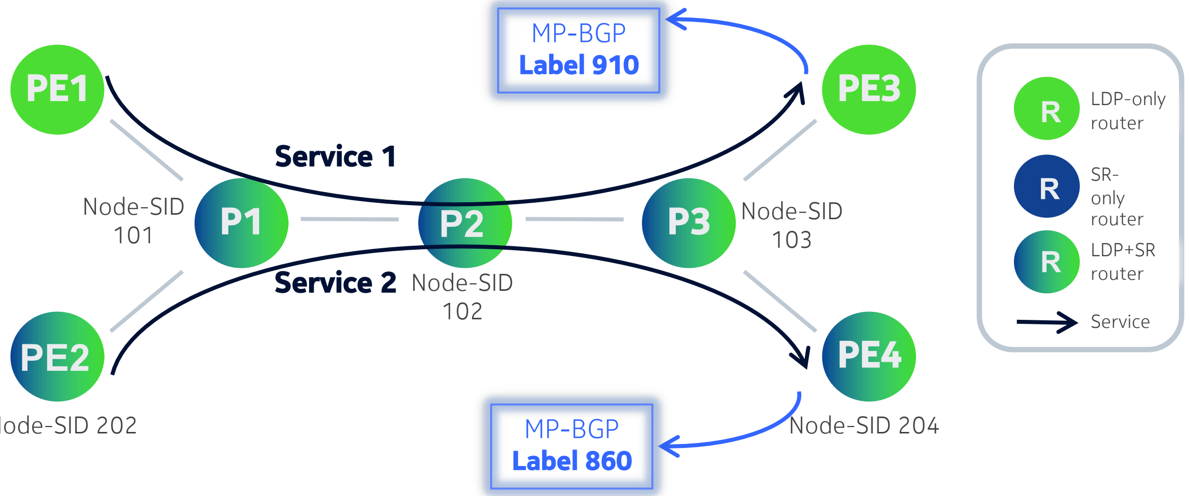
Segment routing and LDP inter-operability

- If an MPLS control plane client (i.e. LDP, RSVP, BGP, SR) installs forwarding entries into the MPLS data-plane, those entries need to be unique in order to function as “Ships in the Night”.
- It’s also likely that these control planes can and will co-exist. For example, LDP and SR could co-exist, where:
  - LDP and SR are present on all routers in the network. Preference for LDP or SR for service tunnels is a local matter at the head-end. SR can also be used to enhance FRR coverage.
  - SR is only present in parts of the network. LDP and SR can be interworked to provide an end-to-end tunnel and/or an FRR tunnel due to the presence of an SR Mapping Server (SRMS).

Segment routing and LDP inter-operability
Scenario 1: Ships-in-the-night co-existence

- Co-existence of LDP-based and SR-based services in the same network

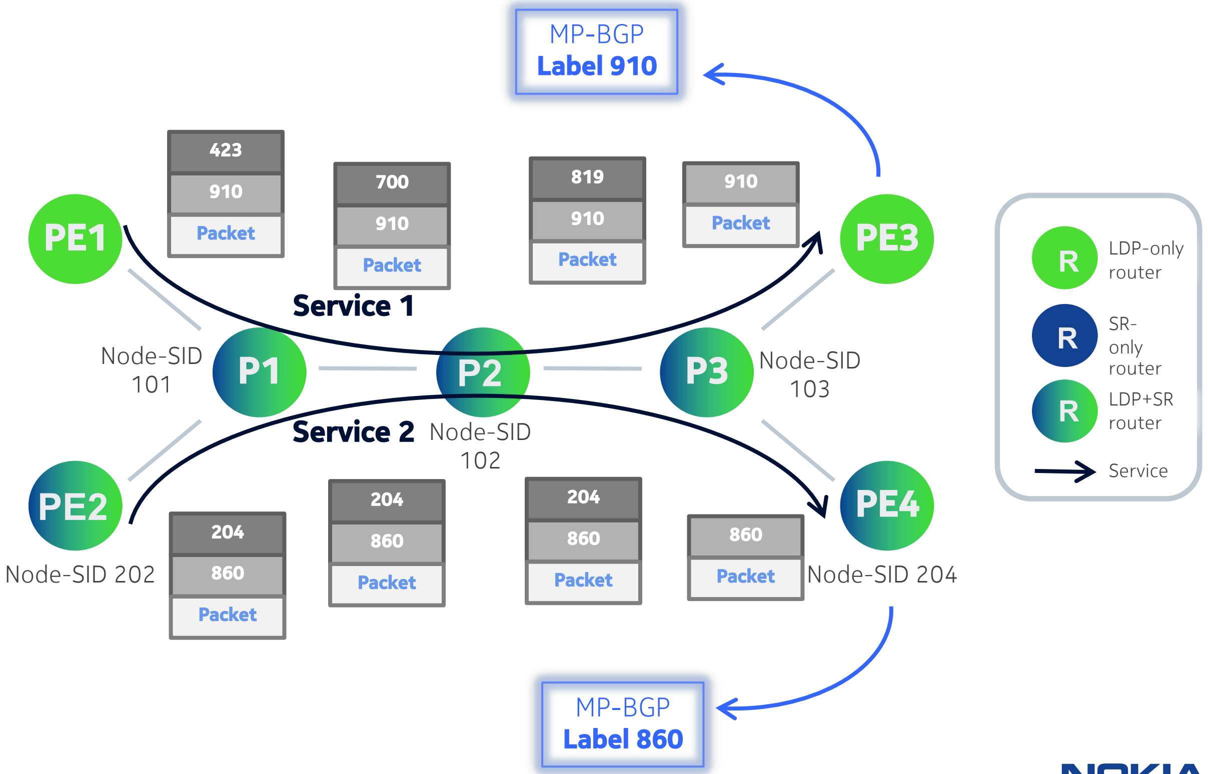
- Requirements:
  - Service 1 to be tunneled via LDP
  - Service 2 to be tunneled via SR
  - Penultimate Hop Popping (PHP) to be used for both services



  Public

Segment routing and LDP inter-operability
Scenario 1: Ships-in-the-night co-existence (cont.)

- Outcome:
  - Service 1 is tunneled from PE1 to PE3 through a continuous LDP LSP traversing P1, P2 and P3.
  - Service 2 is tunneled from PE2 to PE4 through a continuous SR node segment traversing P1, P2 and P3.



  Public

NOKIA

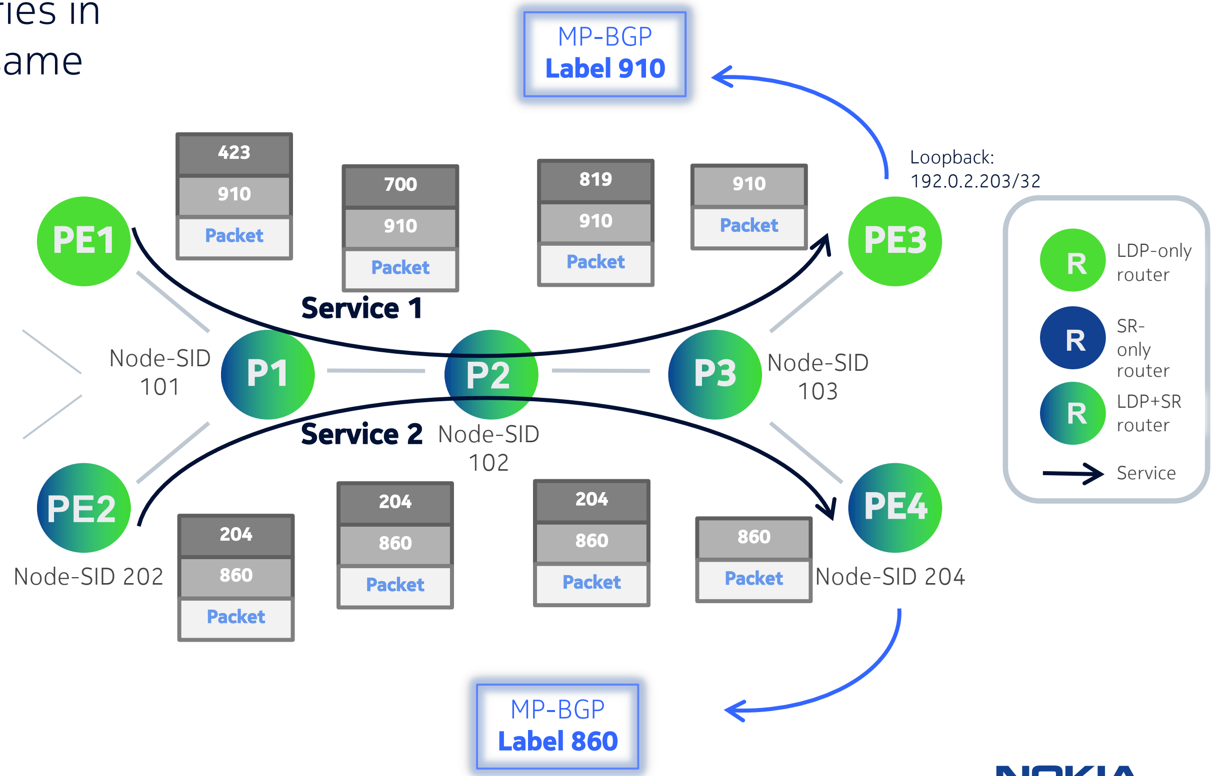
Segment routing and LDP inter-operability

Scenario 1: Ships-in-the-night co-existence (cont.)

- Possible to have multiple entries in the MPLS data plane for the same prefix.

Node P1's MPLS forwarding table

FEC	Incoming Label	Outgoing Label	Next-Hop
192.0.2.204/32 (LDP)	518	612	P2
192.0.2.204/32 (SR)	204	204	P2



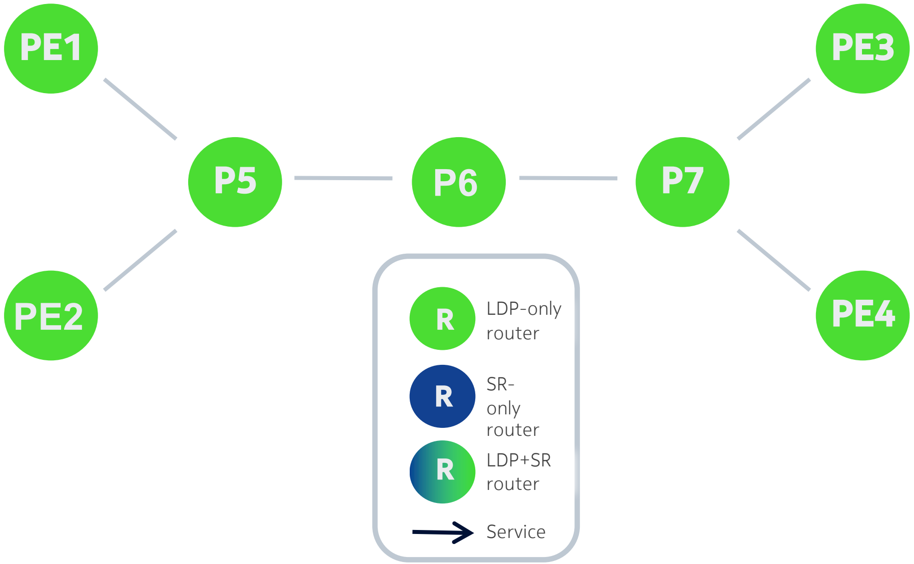
  Public

NOKIA

Segment routing and LDP inter-operability

Scenario 2: Migration from LDP to SR

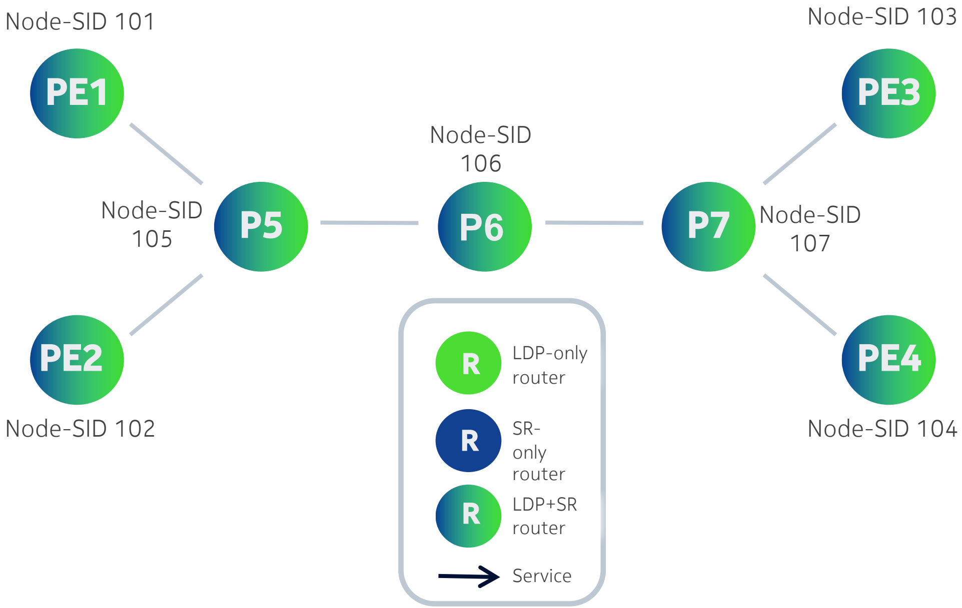
- **Stage 1:**
  - All routers initially run only LDP. All services are tunneled from the ingress PE to the egress PEs over a continuous LDP LSP.



Segment routing and LDP inter-operability

Scenario 2: Migration from LDP to SR (cont.)

- **Stage 2:**
  - All the routers are upgraded to SR. They are configured with the SRGB range [100, 300]. PE1, PE2, PE3, PE4, P5, P6 and P7 are configured with the node segments 101, 102, 103, 104, 105, 106 and 107, respectively .
  - Service traffic is still tunneled over LDP LSPs. For example, PE1 has an SR node segment to PE3 and an LDP LSP to PE3 but the LDP IP2MPLS encapsulation is preferred, by default or via configuration.

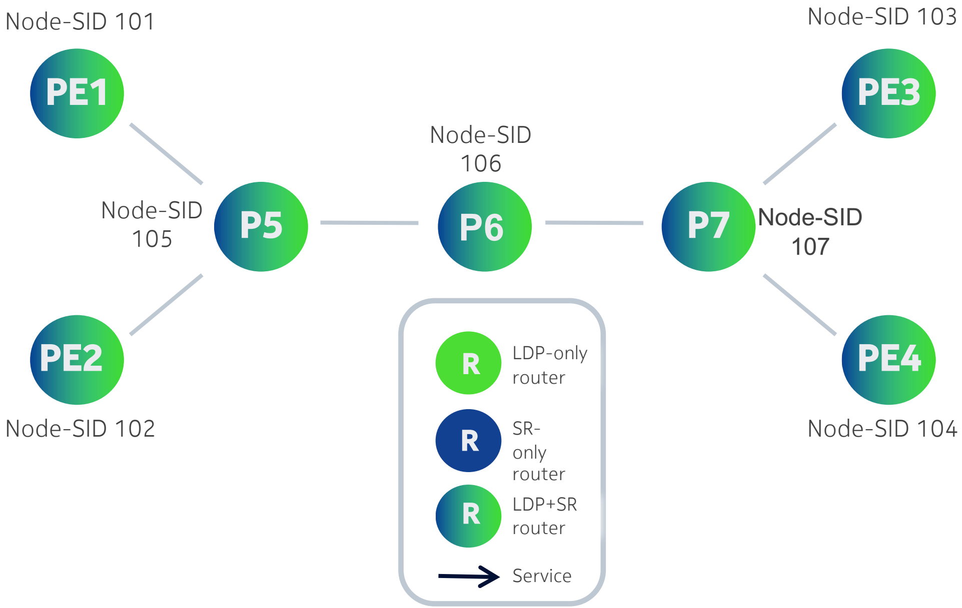


  Public

NOKIA

Segment routing and LDP inter-operability
Scenario 2: Migration from LDP to SR (cont.)

- **Stage 3:**
  - Local policy at PE1 is configured to prefer SR encapsulation over LDP.
  - The service from PE1 to any other PE is now riding over SR. All other service traffic is still transported over LDP LSPs.

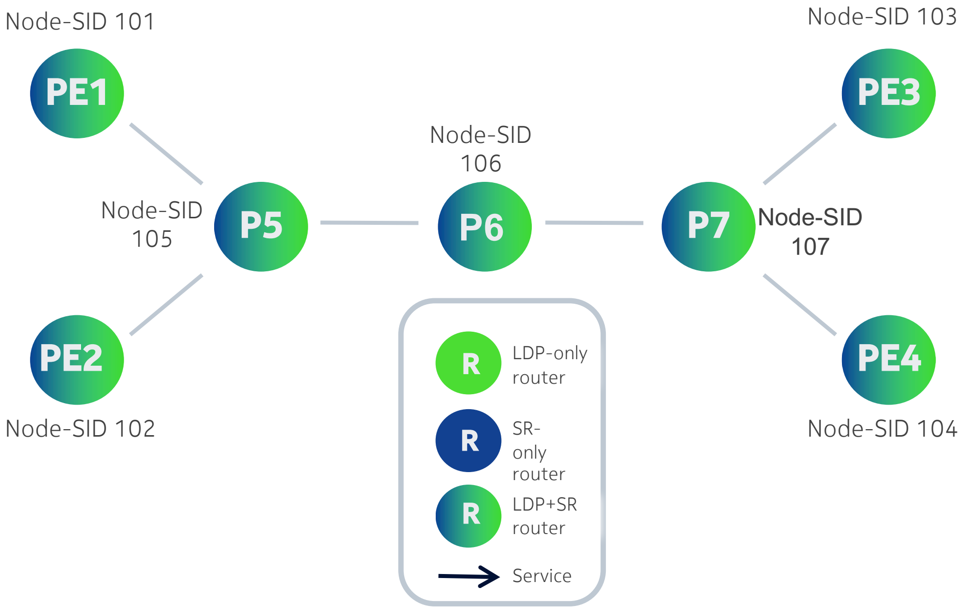


  Public

NOKIA

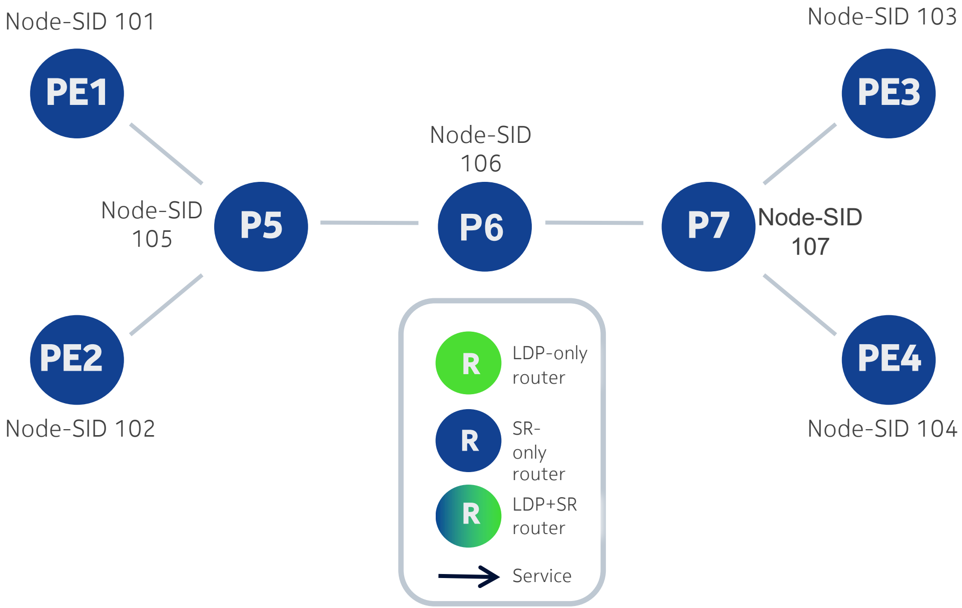
Segment routing and LDP inter-operability
Scenario 2: Migration from LDP to SR (cont.)

- **Stage 4:**
  - Gradually, all edge routers are configured to prefer SR over LDP encapsulation.
  - All the service traffic is now transported over SR.
  - LDP is still operational and services could be reverted to LDP should there be any issues.



Segment routing and LDP inter-operability
Scenario 2: Migration from LDP to SR (cont.)

- **Stage 5:**
  - After a period of smooth operation, LDP can be de-configured from all routers.
  - All routers now solely run SR



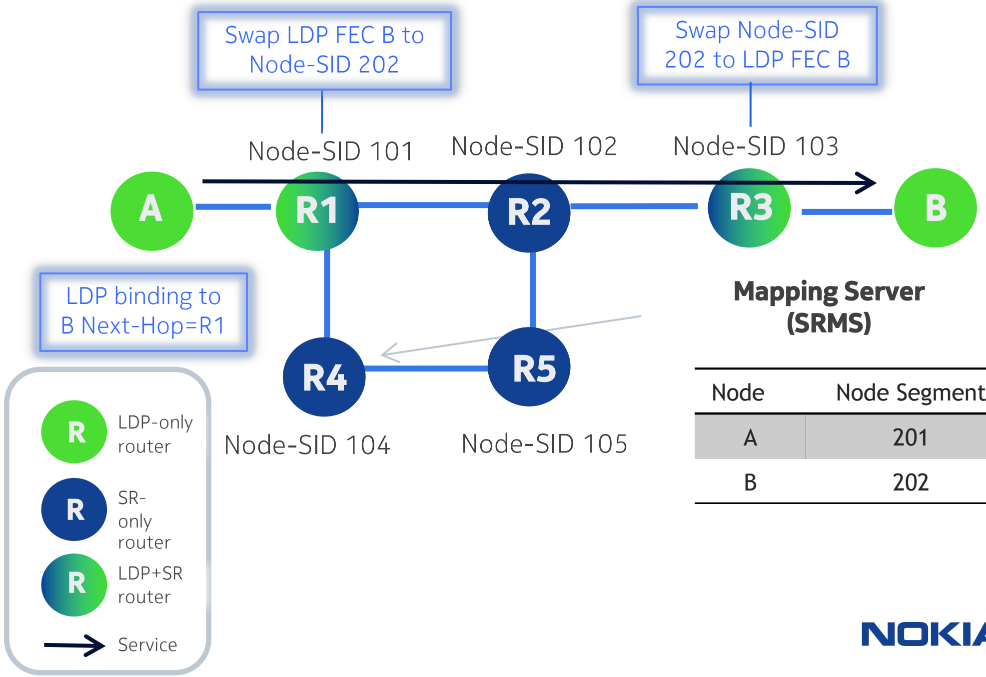
  Public

NOKIA

Segment routing and LDP inter-operability

Scenario 3: Mix of SR-only and LDP-only routers (SR and LDP inter-working)

- One or more Segment Routing Mapping Servers (SRMS) are used to advertise Node-SIDs on behalf of non-SR routers. For example, R4 advertises Node-SIDs 201, and 202, respectively for the LDP-only routers A, and B.
  - A forwards to R1 using conventional LDP. R1 does not have a LDP label binding for its next-hop R2, but does have an SR Node-SID, so it swaps its local LDP-label for FEC B to Node-SID 202 and forwards to R2.
  - R3 knows that B is not SR-capable (as B did not advertise SR capability in ISIS/OSPF), so R3 swaps Node-SID 202 for LDP FEC B.



Node	Node Segment
A	201
B	202

  Public

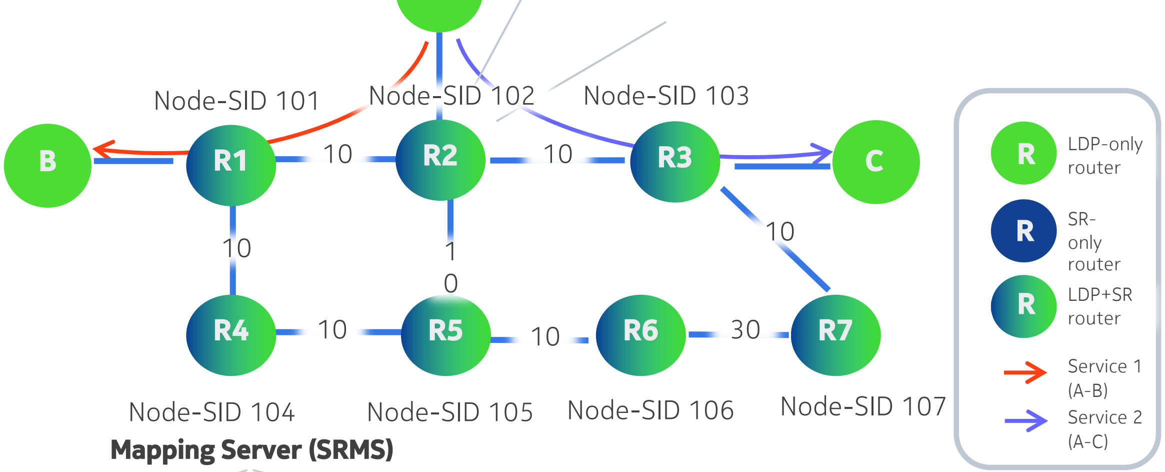
NOKIA

Segment routing and LDP inter-working

Scenario 4: Using SR to provide LDP fast reroute

- A similar methodology to LDP-SR interworking can be used to provide FRR coverage:
  - Potential for increased coverage where SR is present only in parts of the network.
  - Full coverage if SR is present on all routers in the network (in which case no Mapping Server is required).

LDP FEC	Incoming Label	Outgoing Label	Outgoing Next-Hop
B	Advertised by R2	Advertised by R1	R1
C	Advertised by R2	Advertised by R3	R3



Node	Node Segment
A	201
B	202
C	203

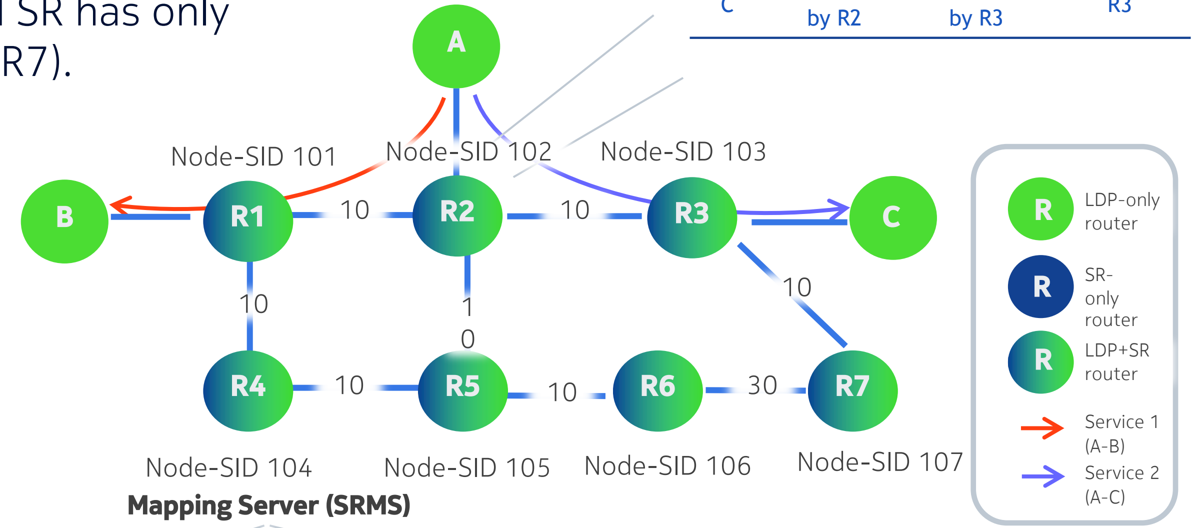
NOKIA

Segment routing and LDP inter-working

Scenario 4: Using SR to provide LDP fast reroute (cont.)

- In the example shown, LDP is used throughout the network, and SR has only partial coverage (routers R1-R7).
  - R4 is SRMS and advertises Node-SID 201, 202, 203 respectively for the LDP-only routers A, B, and C.
  - Router A has services to B and C. LDP is the preferred transport protocol and is used by the head-end, router A (local decision).
  - Objective is to protect link R2-R1 for service 1, and link R2-R3 for service 2.

LDP FEC	Incoming Label	Outgoing Label	Outgoing Next-Hop
B	Advertised by R2	Advertised by R1	R1
C	Advertised by R2	Advertised by R3	R3



Node	Node Segment
A	201
B	202
C	203

  Public

NOKIA

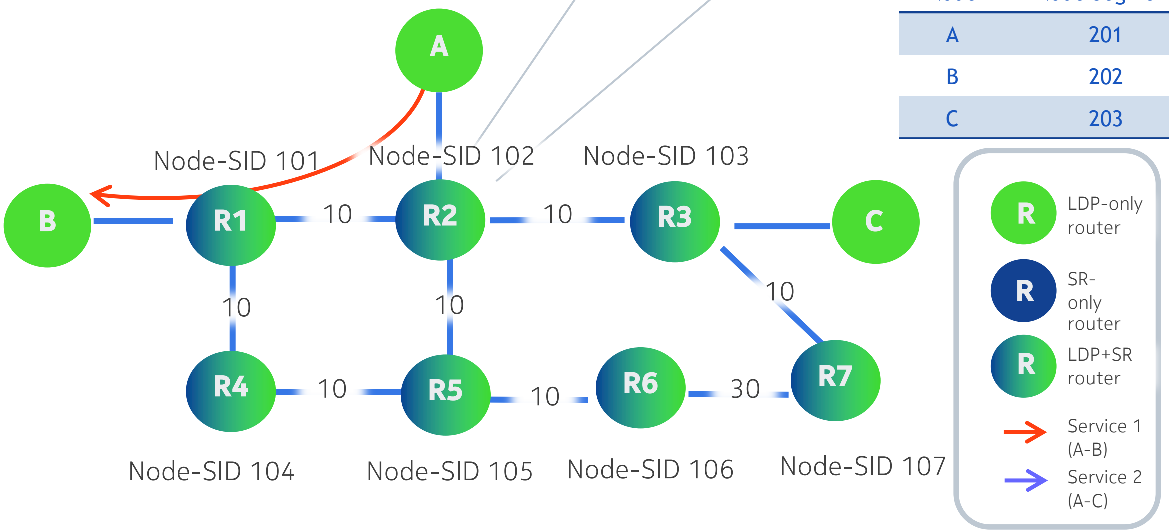
Segment routing and LDP inter-working

Scenario 4: Using SR to provide LDP fast reroute (cont.)

- Protecting service 1
  - Objective is to protect link R2-R1 with a Loop-Free Alternate (LFA) for B (Service 1).
  - Routers R1-R7 advertise Node-SID and Adjacency-SIDs for its IGP adjacencies. R4 is acting as Mapping Server for A, B, and C.
  - In steady-state, LDP is used as the preferred transport tunnel for Service 1 (A-R2-R1-B).

Dest.	Incoming Label	Outgoing Label	Outgoing Next-Hop	Backup Outgoing Label	Backup Outgoing Next-Hop
B	Advertised by R2	Advertised by R1	R1	202 (B N-SID)	Repair tunnel: Node-SID R4 Next-Hop R5

Node	Node Segment
A	201
B	202
C	203



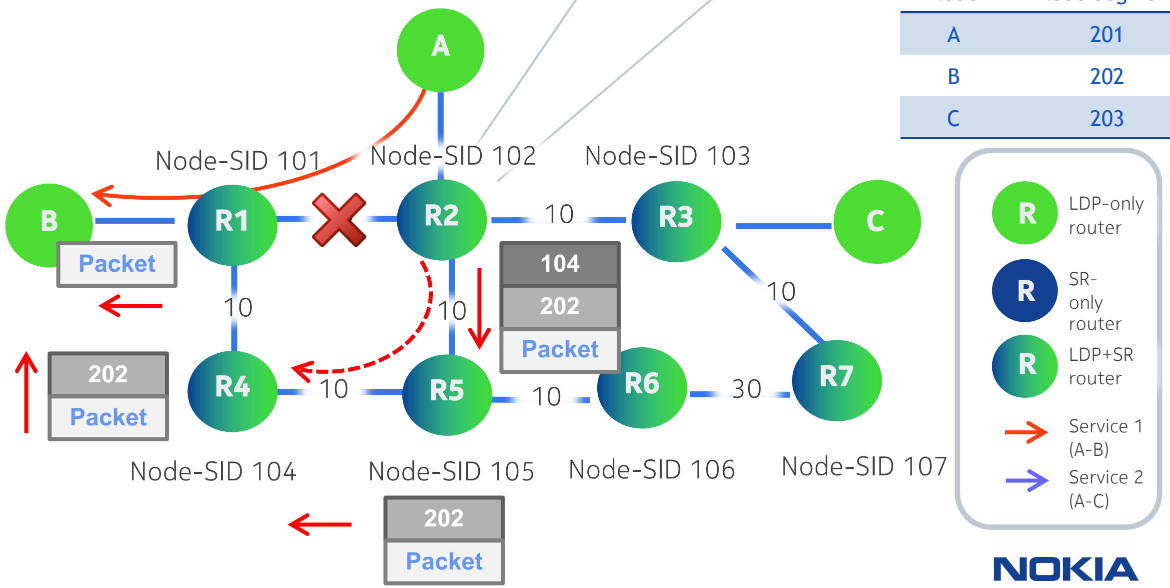
Segment routing and LDP inter-working

Scenario 4: Using SR to provide LDP fast reroute (cont.)

- Protecting service 1 (cont.)
  - Upon failure of link R2-R1, R2 swaps the incoming top (LDP) label with the Node-SID for B (202). R2 then sends the packet into a repair tunnel to R4.
    - R2 forwards the label stack {104, 202} to R5.
    - R5 pops Node-SID 104 (PHP) and forwards the packet to R4.
    - R4 swaps label 202 for 202 and forwards to R1. R1's Next-Hop to B is not SR-capable, so R1 swaps 202 for the LDP label announced by it's Next-Hop (in this case, implicit-null).

Dest.	Incoming Label	Outgoing Label	Outgoing Next-Hop	Backup Outgoing Label	Backup Outgoing Next-Hop
B	Advertised by R2	Advertised by R1	R1	202 (B N-SID)	Repair tunnel: Node-SID R4 Next-Hop R5

Node	Node Segment
A	201
B	202
C	203



NOKIA

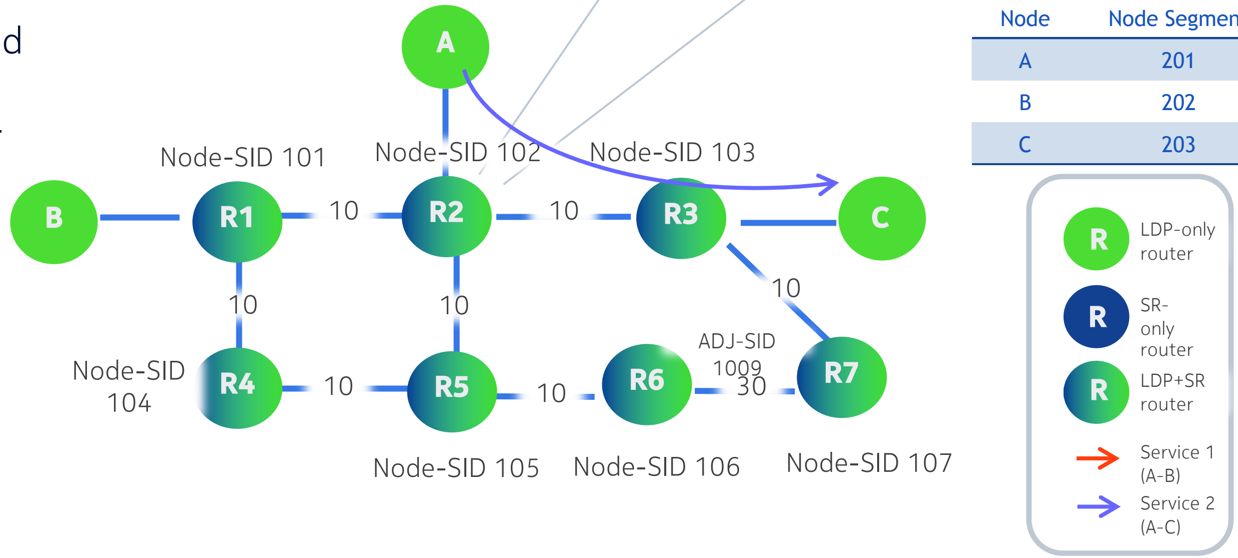
Segment routing and LDP inter-working

Scenario 4: Using SR to provide LDP fast reroute (cont.)

- Protecting service 2
  - Objective is to protect link R2-R3 with an LFA for C (Service 2).
  - In steady-state, LDP is used as the preferred transport tunnel for Service 2 (A-R2-R3-C).

Dest.	Incoming Label	Outgoing Label	Outgoing Next-Hop	Backup Outgoing Label	Backup Outgoing Next-Hop
C	Advertised by R2	Advertised by R3	R3	203 (C N-SID)	Repair tunnel to R6: {106, 1009} Next-Hop R5

Node	Node Segment
A	201
B	202
C	203



  Public

NOKIA

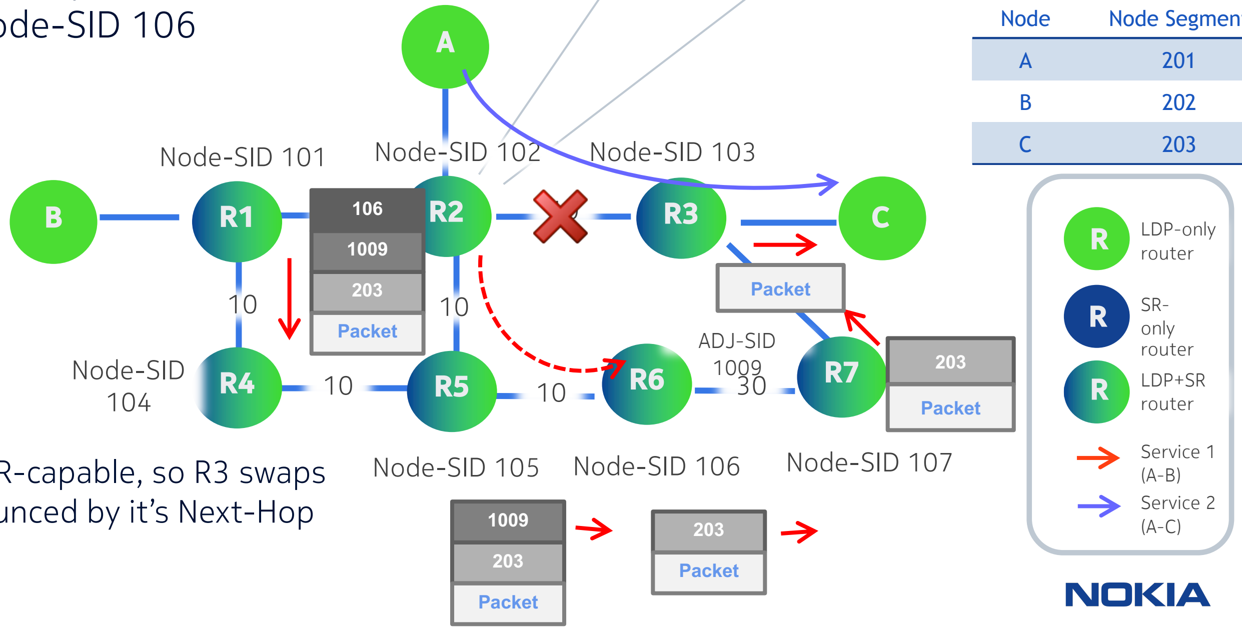
Segment routing and LDP inter-working

Scenario 4: Using SR to provide LDP fast reroute (cont.)

Dest.	Incoming Label	Outgoing Label	Outgoing Next-Hop	Backup Outgoing Label	Backup Outgoing Next-Hop
C	Advertised by R2	Advertised by R3	R3	203 (C N-SID)	Repair tunnel to R6: {106, 1009} Next-Hop R5

- Protecting service 2 (cont.)
  - Upon failure of link R2-R3, R2 swaps the incoming top (LDP) label with the Node-SID for C (203). R2 then sends the packet into a repair tunnel to R6 with Node-SID 106 followed by Adj-SID 1009.
    - R2 forwards the label stack {106, 1009, 203} to R5.
    - R5 pops 106 (PHP) and forwards the packet to R6.
    - R6 pops Adj-SID 1009 and forwards the packet to R7.
    - R7 swaps 203 for 203 and forwards to R3.
    - R3's Next-Hop to C is not SR-capable, so R3 swaps 203 for the LDP label announced by it's Next-Hop (in this case, implicit-null).

Node	Node Segment
A	201
B	202
C	203



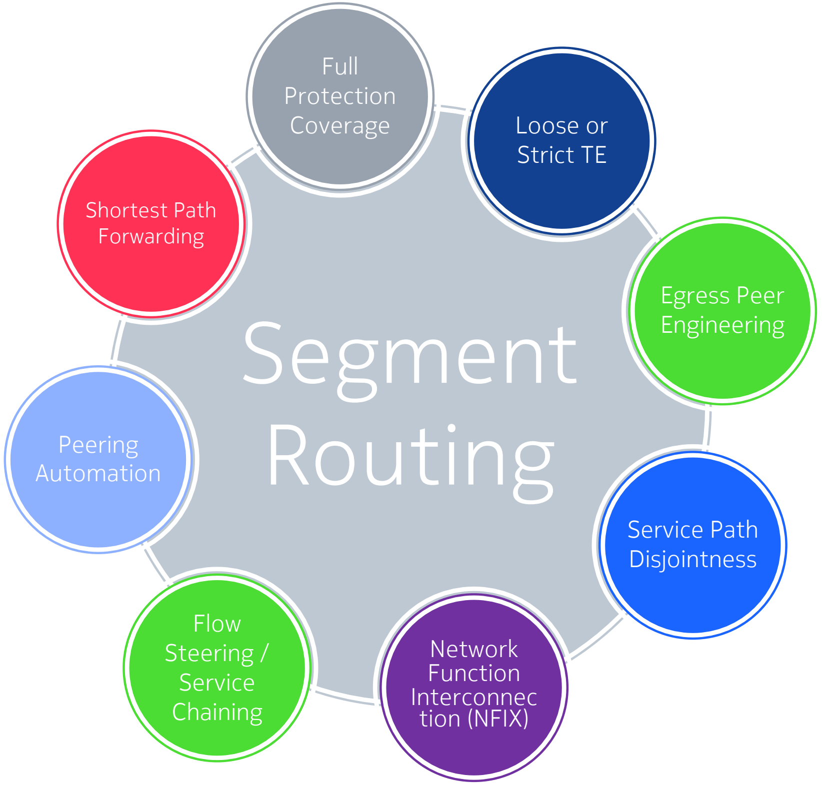
NOKIA

conclusion

  Public

Segment routing in a nutshell

- Reduce control plane protocols
- Improve Scale / Optimize State
- Ease IPv6 Migration
- Programmability / SDN Control



  Public

NOKIA

NOKIA