ESnet
ENERGY SCIENCES NETWORK

Segment Routing Experience at ESnet

Nick Buraglio, Chin Guok
Energy Sciences Network (ESnet)
Lawrence Berkeley National Laboratory

J-RENS
DePaul University
Oct 10, 2018

What is Segment Routing?

- Segment routing is a networking technology that combines the features of MPLS with the flexibility of SDN. It allows for controller augmented and source-based routing without the need for maintaining state across a network core and allowing for seamless fallback to traditional network protocols in the case of failures.
- Originally scoped as a way to simplify QoS in MPLS
- Drafts date back to 2004
- Our journey started with a round table discussion on segment routing in June of 2016

ESnet

Motivations for Segment Routing in ESnet:
Full Traffic Engineering Solution

- Fine grain control over network link loading
  - Any path between two (“Low-Touch”) service edge devices can be explicitly traffic engineered across the “Hollow” Core.
  - Central path computation and management for network wide optimization.
- Per service instance service guarantees
  - Each service (e.g. L2/L3VPN) will have a distinct set of LSPs associated with it.
- Faster convergence times during failures
  - Use of Fast Re-Route (FRR) results in quicker restoration vs waiting for entire routing table to converge.
- Support for custom restoration policies
  - Use of (external) controller can support complex and customized restoration schemes.

ESnet

Features

- Segment routing (SR) contains many of the powerful and widely deployed features of MPLS in addition to many functional improvement and extensions
  - Traffic Engineering
  - Compatibility with RSVP
  - Path Engineering
  - On-demand next-hop
  - Failure protection

ESnet

Notable details

- Simplified protocol suite
  - Label distribution within the IGP (i.e. no need for LDP; Leverage ISIS-SR, OSPF-SR)
  - Transparency with current technologies such as L2vpn and L3vpn
- Greater troubleshooting ease
  - Global label space (i.e. label space is deterministic and configurable)
  - Removes state from the network

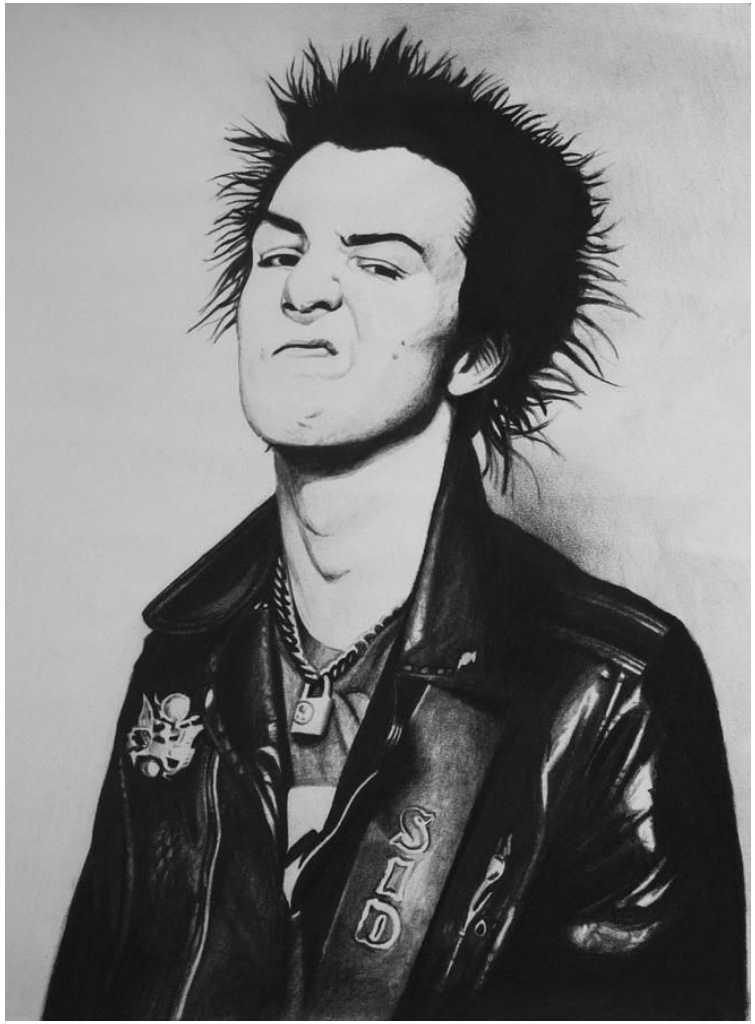
ESnet

Terminology

- Labels
  - SRGB - Segment Routing Global Block
    - Globally unique (Node-SIDS, Anycast-SIDS)
  - SRLB - Segment Routing Local Block
    - Node specific, Locally assigned (Adjacency-SIDS, Binding-SIDS)
- SID: “Segment” Identifier (More about this soon)
- TI-LFA: Topology Independent Loop Free Alternative
- FRR: Fast re-route
- ERO: Explicit Route Object (explicit path)
- PCC - Path Computation Client
- PCE - Path Computation Engine
- CSPF - Constrained Shortest Path First

ESnet

All SIDs are not created equal



Segment ID
32 Bit integer

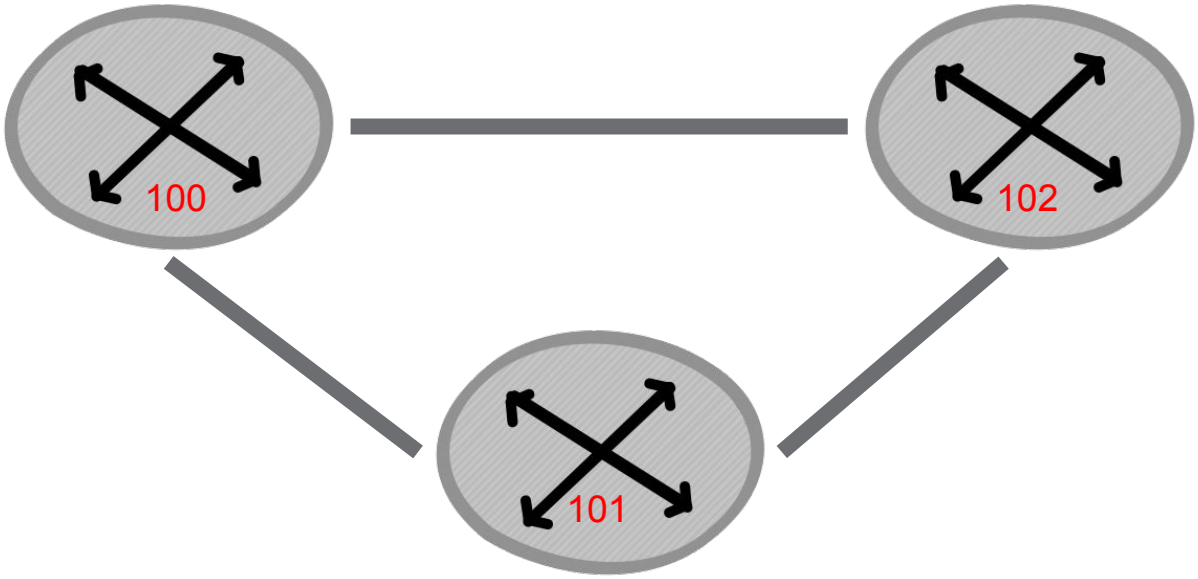
- Prefix/Node SID
- Adjacency-SID
- Anycast SID
- Binding SID

IPv6 SR exists but has little support

ESnet

Node Segment ID

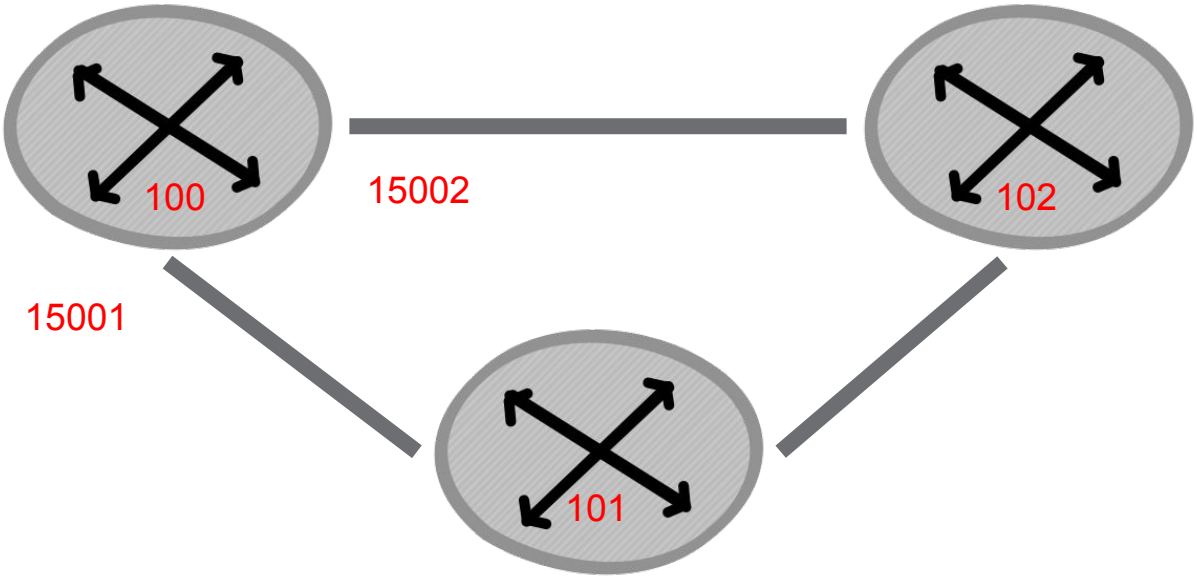
- Node SID
  - Node identifying ID
  - “Globally” unique - unique within the network
  - Advertised by the IGP



ESnet

Adjacency Segment ID

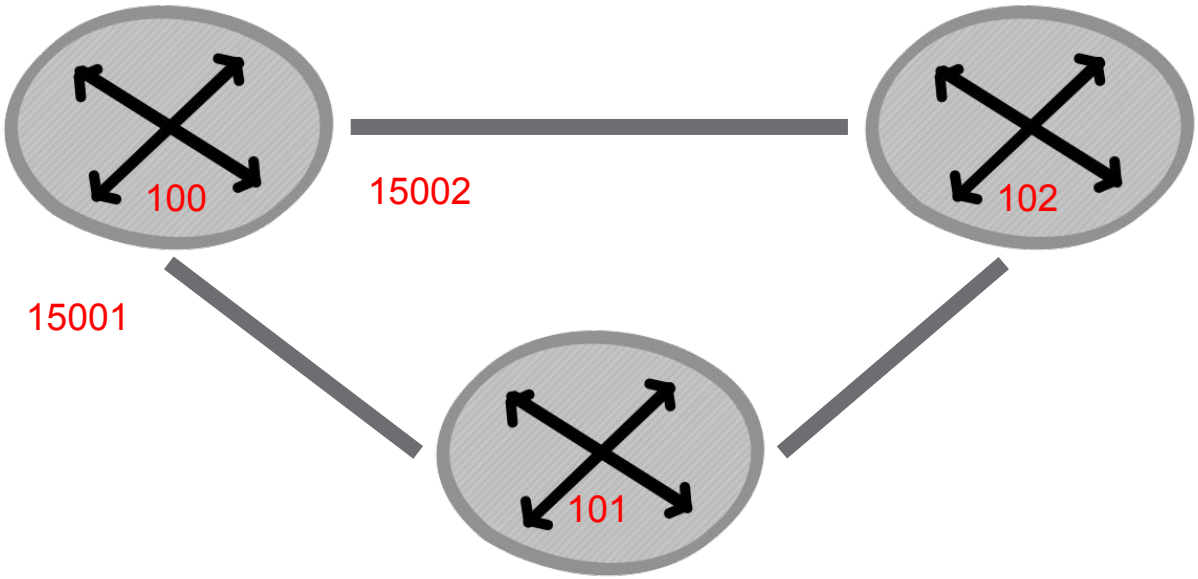
- Adjacency SID
  - Identifies an interface / segment of the network
  - Adj-SIDs are locally scoped to the router (unlike Node-SIDs, Anycast-SIDs which have network wide scope).
  - Adj-SIDs do not have to be (and by default won't be) persistent across reboots.
  - Not globally unique
  - Automatically generated by individual device
  - Only installed into the data plane by adjacent devices



ESnet

Adjacency Segment ID

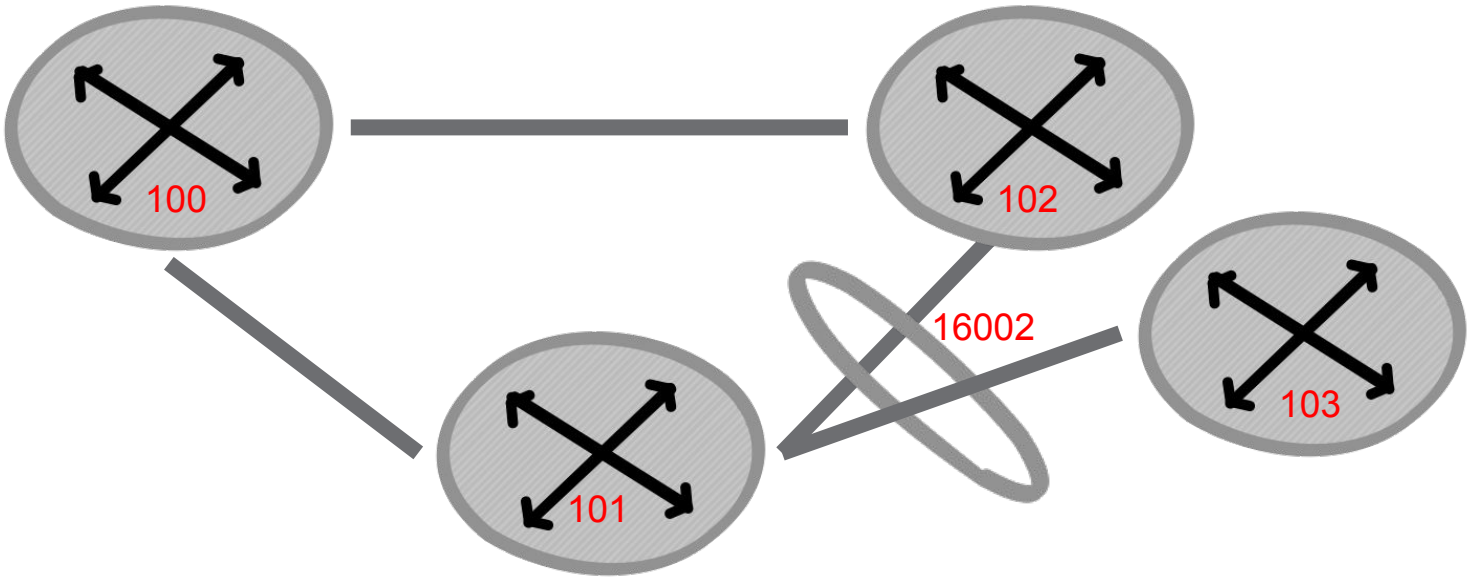
- Some vendor implementation support static assignment of Adj-SIDs, this is taken out of the Segment Router Local Block (SRLB).



ESnet

Anycast Segment ID

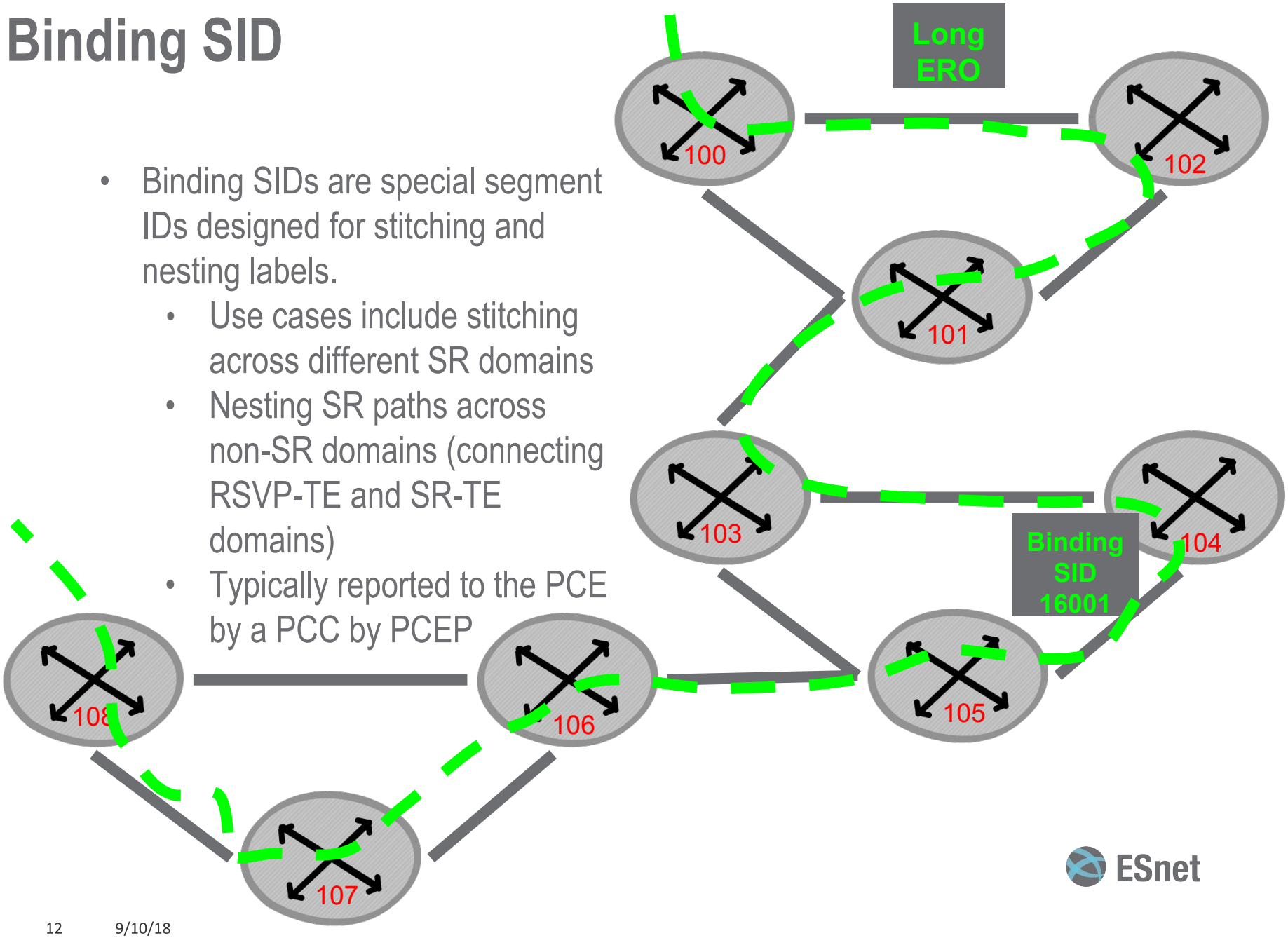
- Similar to adjacency SID
- References a group of devices
- Enforces ECMP shortest path forwarding
- Devices must advertise the same resource (prefix and SID)



ESnet

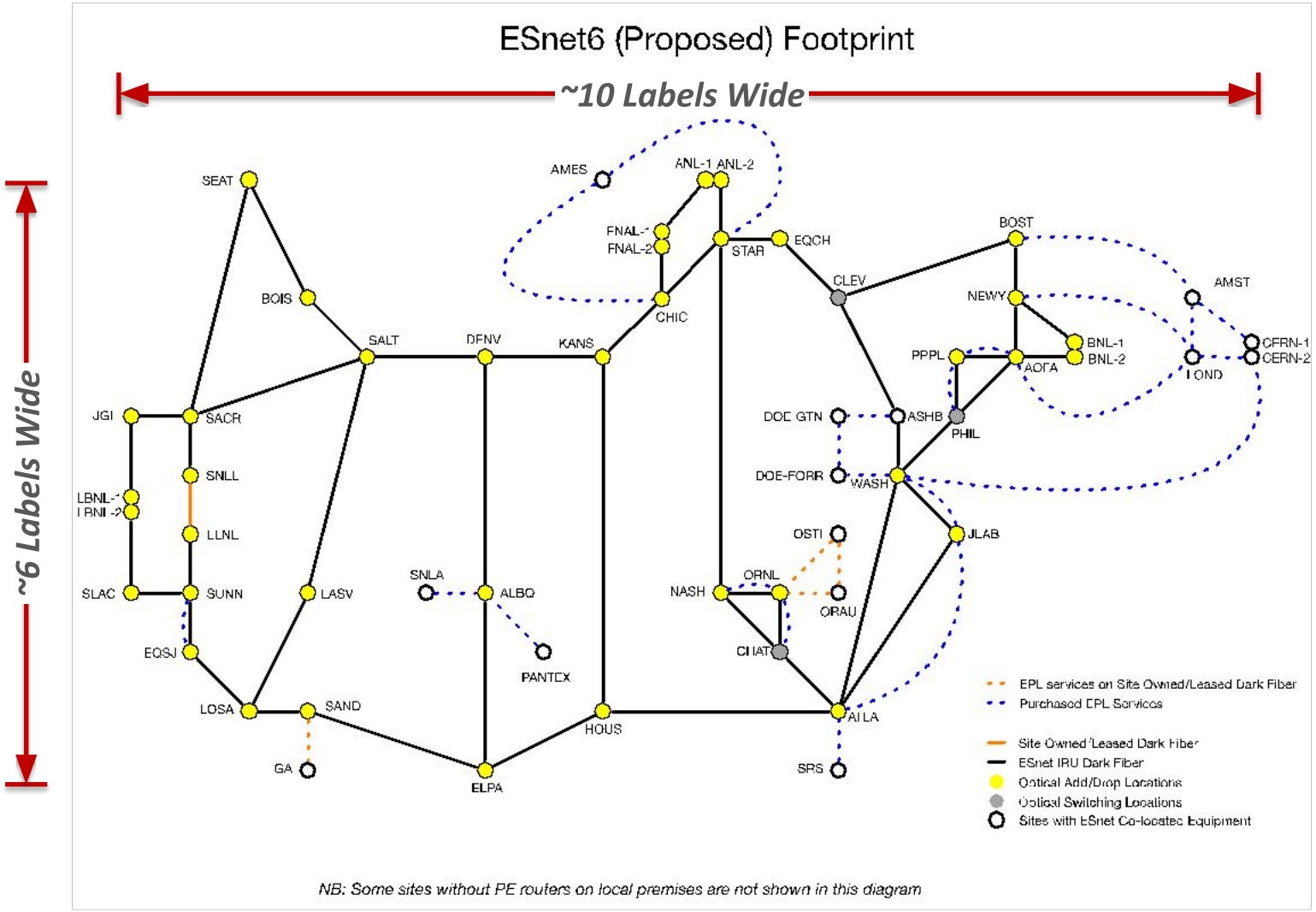
Binding SID

- Binding SIDs are special segment IDs designed for stitching and nesting labels.
  - Use cases include stitching across different SR domains
  - Nesting SR paths across non-SR domains (connecting RSVP-TE and SR-TE domains)
  - Typically reported to the PCE by a PCC by PCEP

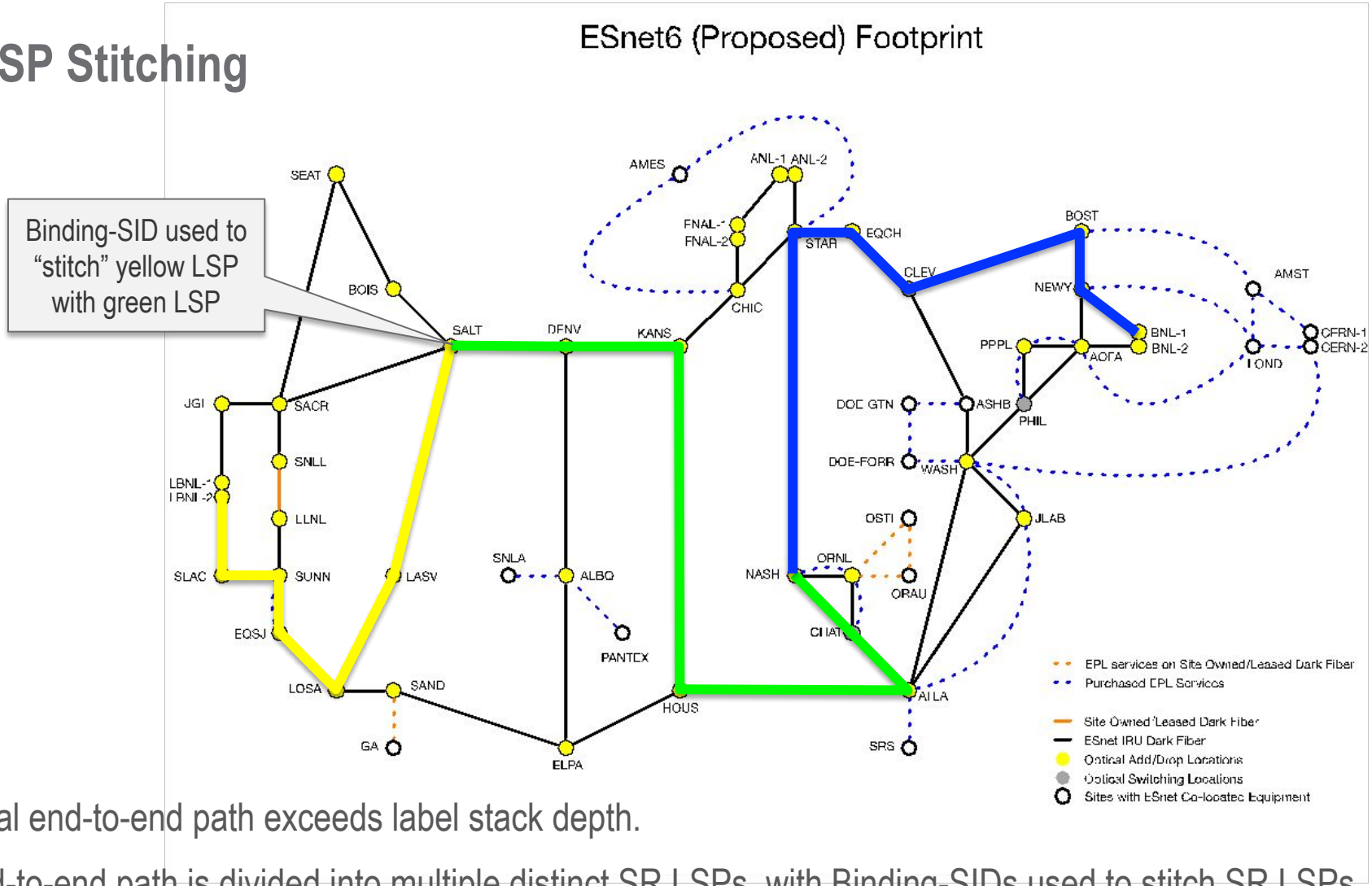


ESnet

ESnet Label Span



SR LSP Stitching



- Total end-to-end path exceeds label stack depth.
- End-to-end path is divided into multiple distinct SR LSPs, with Binding-SIDs used to stitch SR LSPs.
- ** End-to-end path protection/restoration will require either S-BFD (in-skin) or PCE (in conjunction with topology update).

ESnet

SR Labels and Label Stack

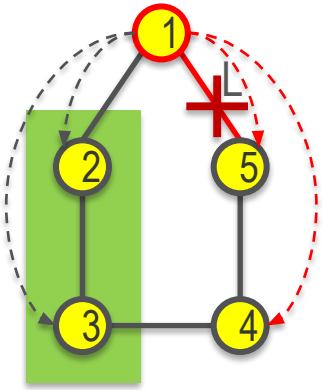
- Typical label “types”
  - Service Label – (1) Service identifier to separate customer data, e.g. VPN Route Target, or
  - FRR Label(s) – (0-n) Used for (TI/r)LFA for fast reroute.
  - Entropy Label – (0, 2) Used for load-balancing (e.g. EL and EL Indicator).
  - Transport Labels – (1-n) Used to define the path of the SR LSP (e.g. Prefix-SID, Adj-SID, Anycast-SID, Binding-SID).
  - Router Alert Label - (0-1) Used to notify the router for exception label processing (e.g. VCCV).
- Label Stack
  - Typical (minimal) stack – Service (1), FRR (2), Entropy (1), Transport (1)
  - Hardware support:
    - *Major vendor support ranges from 6-16 labels deep depending on code and silicon versions*

ESnet

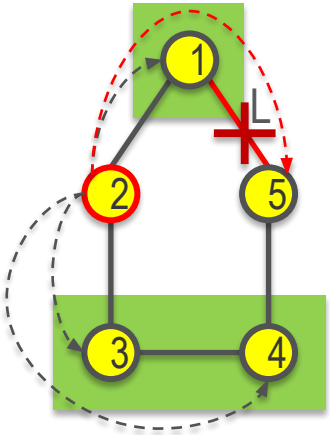
Loop-Free Alternative (LFA) Terminology

(From Remote LFA RFC7490)

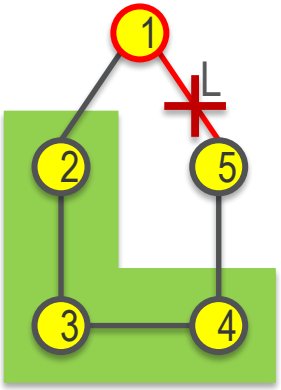
- **P-space(*S, L*)**: set of nodes reachable (using pre-convergence paths) from node *S* without using protected link *L*
- **Extended P-space (*PLR, L*)**: Union of the P-space of the neighbors of PLR
- **Q-space(*D, L*)**: Set of nodes that can reach (using pre-convergence paths) destination D without using protected link *L*



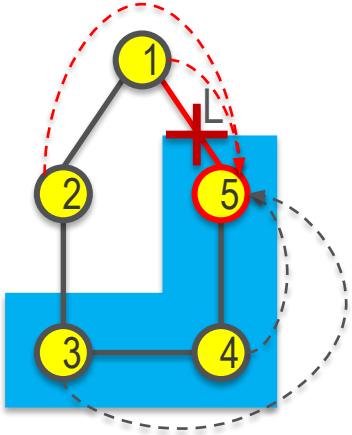
P-space(1,L)



P-space(2,L)



Ext P-space(1,L)



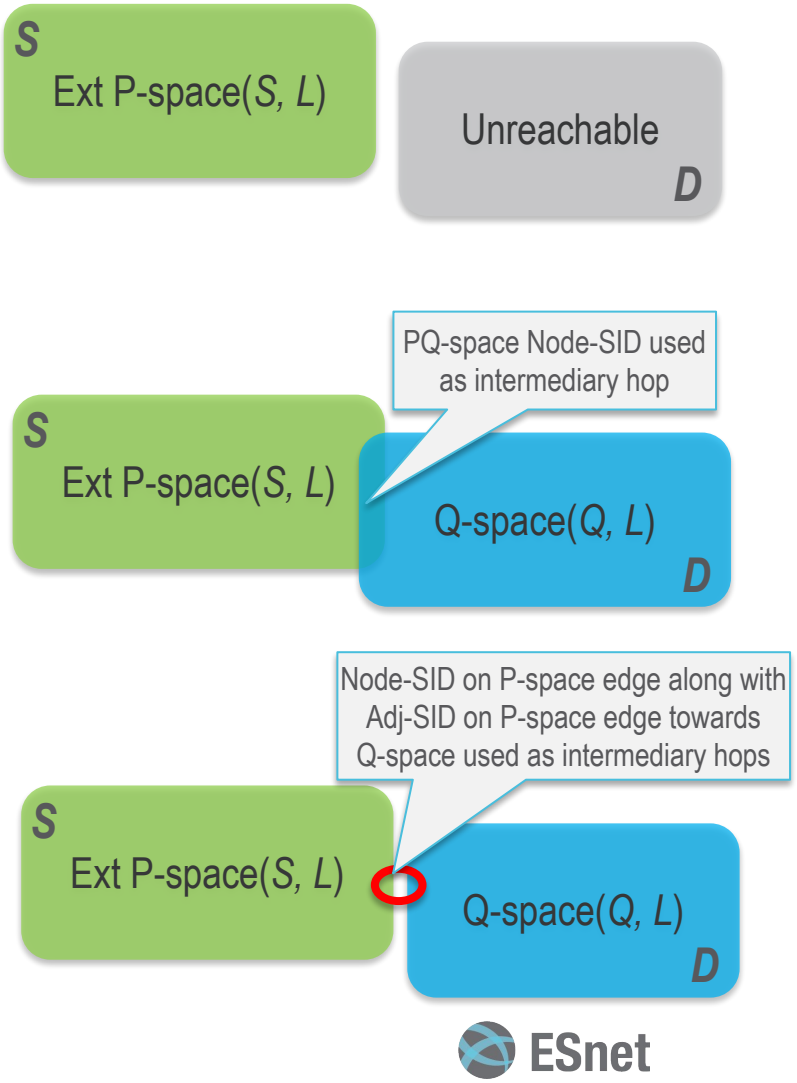
Q-space(5,L)

NB: In the above examples, Source = 1, Destination =5, and all links have the same metrics

ESnet

LFA vs Remote LFA vs Topology Independent LFA

- **LFA** will be able to maintain connectivity to D if it is within the Extended P-space(*S, L*).
- **rLFA** will be able to maintain connectivity to *D* if there is an overlap between the Extended P-space(*S, L*) and Q-space(*Q, L*).
- **TI-LFA** will be able to maintain connectivity to D even if the Extended P-space(*S, L*) and Q-space(*Q, L*) do not overlap. However, the requirement for additional labels is directly related to the number of disjoint spaces.



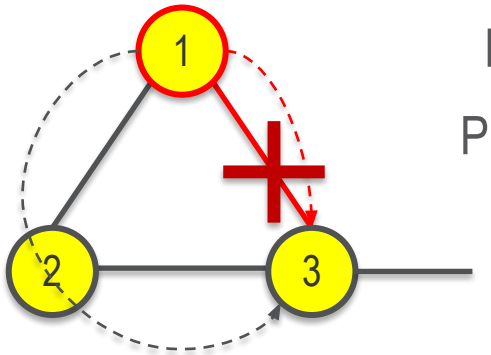
SR Failure Detection Considerations

- Local notification (e.g. failure of directly connected interfaces)
  - Triggers a switch to backup path at the headend node.
  - Triggers FRR on downstream (non-headend) node.
- IGP updates (e.g. topology changes in the network)
  - LSDB is fed back to external controller to compute a restoration path.
- Seamless BFD (e.g. end-to-end BFD on SR LSP(s))
  - Triggers a switch to pre-computed backup path (that was pre-configured on the router).
  - *Major vendor support forthcoming as of our testing*
- Timing of changes and updates reporting back to controller and requiring action needs to be further explored

ESnet

Protection and Restoration Considerations

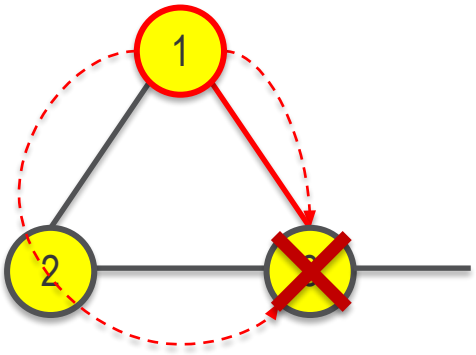
- **FRR is local to failure**.
- For explicit hop-by-hop Adj-SID path definitions, FRR can support against link failure, but not node failure.
  - FRR for Adj-SID link failure is supported by translating the Adj-SID to the next-hop Node-SID.



Link 1-3 failure results in successful FRR to path 1-2-3

Primary path uses Adj-SID_1-3

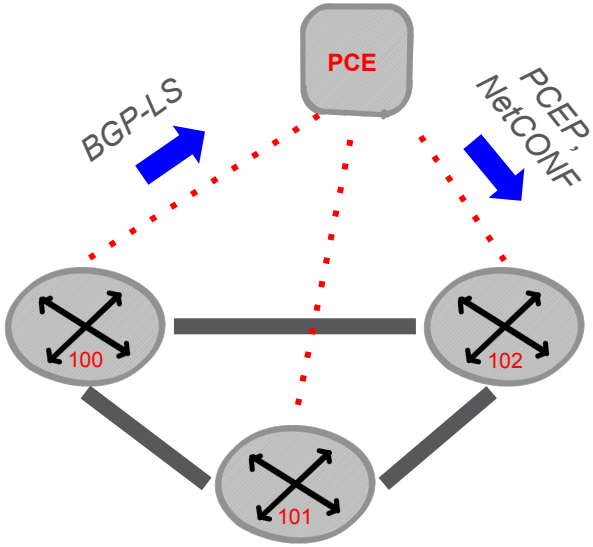
Protection path uses Node-SID_3



Node 3 failure results in unsuccessful FRR

ESnet

Path Computation Engine (PCE) aka Controller



- The Good
  - Removes (configuration) state from the network.
  - Reduces computation overhead from the NE.
  - Centrally controlled with global optimization.
    - Solves bin packing problem.
    - Can support customized PCE algorithm.
  - Large, commercial network adoption.
- The Bad
  - PCE needs to peer with every (head-end) router node to initiate LSP.
  - Needs real-time deep knowledge of network state for PCE initiated protection/recovery.
- The Ugly
  - Limited PCE choices.
  - Some protocols still under development.
  - PCE HA and connectivity become a critical dependency.

ESnet

Platform dependent details

- BGP-LS is typically used to retrieve network state.
- PCEP is used for LSP control.
  - PCEP-LSP ID (local to the router) is how the controller/router identifies the LSP
- “Start Weight” constraint in PCE path computation can be based off a proprietary algorithm to distribute reservation bandwidth (similar to RSVP bandwidth) across various links.
- Some platforms have a signaling mechanism function to tag a link for “maintenance” and resignal all LSPs on the lisk to reroute around it, essentially draining the link.
- Stitching of LSPs performed differently across different controllers
- Controller redundancy models varies
  - Latency requirements may define geographic placement

ESnet

Common controller details

- BGP-LS is used to retrieve network state.
- PCEP is used for LSP control.
  - By default, PCEP provisioned LSPs are also protected (have to choose “Route on Protected IP Link” to make it use TI-LFA).
  - Use Adj-SID by default to define the LSP path.
- Some proprietary-ish mechanisms such as resynchronizing with routers.
  - Not clear what conditions would cause the controller and network to be out of sync and warrant this
- Controller redundancy warrants further testing
  - Cluster can be distributed but must but has latency requirements synchronization to work properly
  - Cluster scaling past 3 instances needs to be tested

ESnet

High level conclusions

- Next step in evolution of MPLS networking
  - Lots of conceptual overlap
- A dizzying number of acronyms and subtle technologies comprise a larger super-set that is “Segment Routing”
- Controller options are limited
  - Protocol support is there
  - Both commercial and FOSS options do exist
- Redundancy is hard / complicated / growing in support for:
  - Loop free paths
  - Controller elements
- Examples of large networks running controller based SR in production is hard to find
- Simplification is achievable but trade offs are necessary

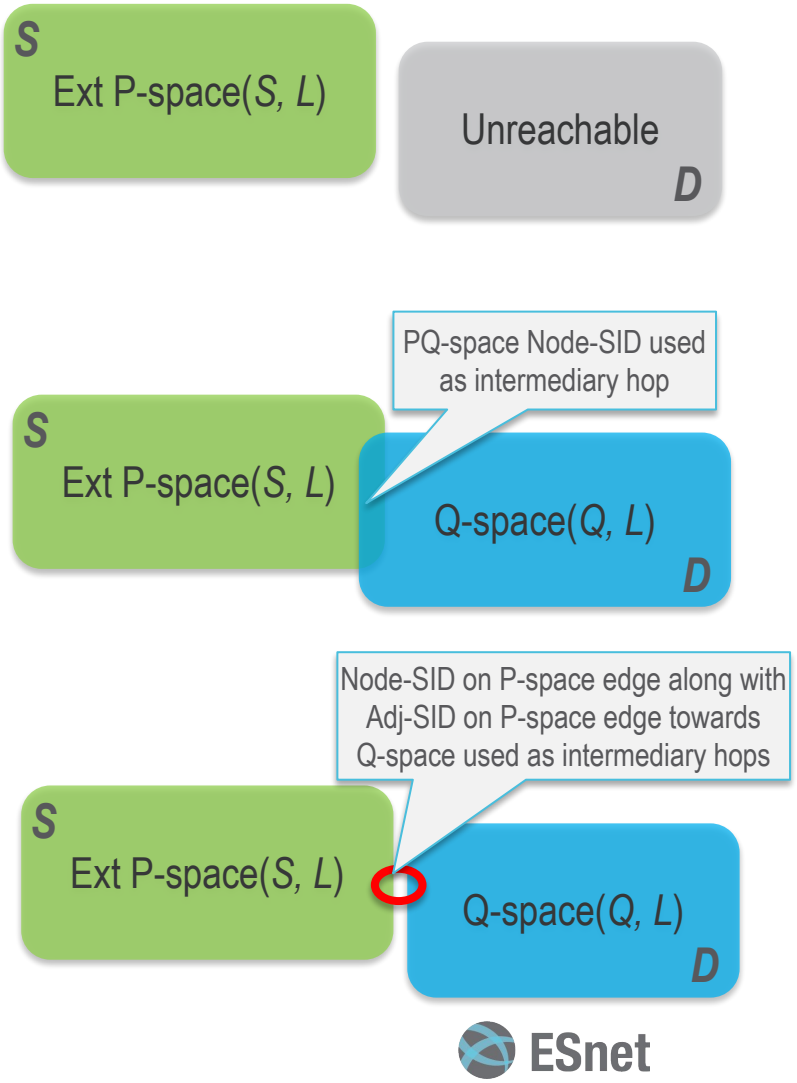
ESnet

Questions...

ESnet

LFA vs Remote LFA vs Topology Independent LFA

- **LFA** will be able to maintain connectivity to D if it is within the Extended P-space(*S, L*).
- **rLFA** will be able to maintain connectivity to *D* if there is an overlap between the Extended P-space(*S, L*) and Q-space(*Q, L*).
- **TI-LFA** will be able to maintain connectivity to D even if the Extended P-space(*S, L*) and Q-space(*Q, L*) do not overlap. However, the requirement for additional labels is directly related to the number of disjoint spaces.



SR LSPs Types and differences

- ISIS-SR LSPs
  - Governed by IGP metrics.
- SR-TE LSPs
  - PCC initiated / PCC controlled.
    - LSP is configure on the router, with path computation done in-skin.
  - PCC initiated / PCE controlled.
    - LSP is configured on the router, with path computation and done by an external controller (via PCEP or BGP-LU) along with any path updates.
  - PCE initiated / PCE controlled.
    - External controller computes and management LSP (using PCEP or BGP-LU), including any path updates. (There is no LSP configuration on the router.)
  - Other proprietary solutions

ESnet