The critical role of Maximum SID Depth (MSD) hardware limitations in Segment Routing ecosystem and how to work around those

NANOG71, October 2017
San Jose, CA

Jeff Tantsura
IAB member, IETF RTGWG Chair
VP Network Architecture, Futurewei, Future Networks

Guedrez Rabah
PhD candidate at Orange Labs
Orange Labs Lannion

Why should you pay attention?

- Segment Routing is becoming mainstream technology, like every other new technology it comes with its own set of limitations.

- Knowing these limitations and understanding how to work around (permanently or temporarily, till new, more capable HW is available) will help you to apply the new technology in your network as well as to make right choices when choosing the next generation HW.

- Ongoing networking disaggregation, where SW and HW might come from different vendors would require additional API's from HW vendors, MSD is one of them.

Vocabulary

- MSD - Maximum SID Depth (defined in IETF OSPF/ISIS/BGP-LS/PCEP drafts)
  - Generic concept defining number of SID's, HW/SW are capable of imposing on a given node
  - Applicable to both, SR-MPLS (labels) and SRv6 (SRH's) data planes
    - Focus of this presentation is SR-MPLS data plane

- SR-MPLS – Segment Routing with MPLS data plane(defined in IETF SPRING drafts)
  - SID instantiated as an MPLS label, context's set by the label value
  - Path (LSP/tunnel) is usually computed by a centralized entity, commonly known as PCE/SDNc
  - PCEP with SR extensions is a commonly used protocol to communicate the path to the ingress
  - MPLS label stack defines at the **ingress** the path a packet will take thru the network
  - Other actions could be defined and applied as packet traverses the network:
    - Apply a service
    - Treat a packet in a special way
    - Set context
    - …

Short SR-MPLS recap – SID types

- Prefix SID
  - Uses SR Global Block (SRGB), must be unique within the SR domain
  - SRGB(‘s) is advertised by an IGP
  - Prefix-SID can be configured as an absolute value or an index (base+offset)

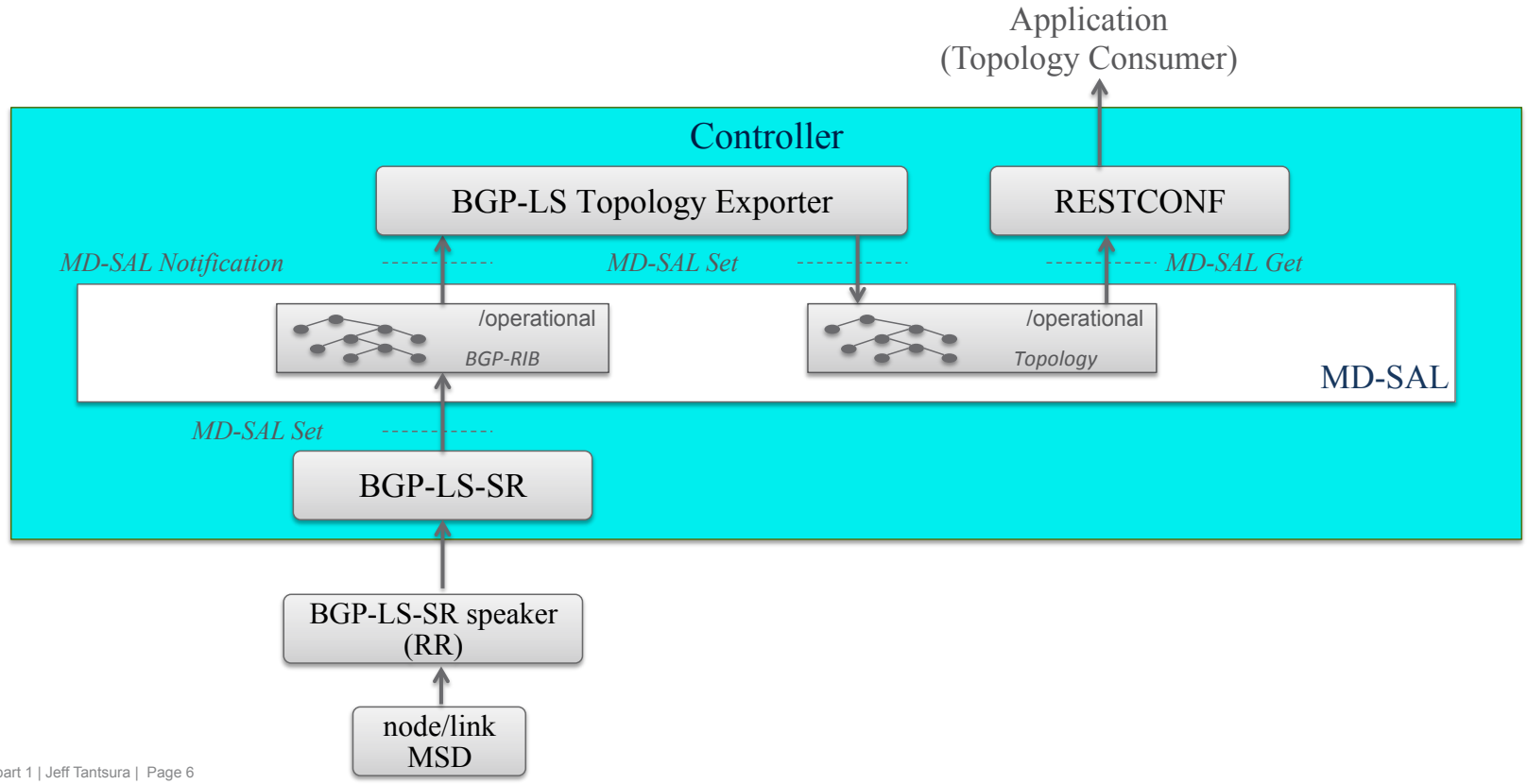
- Node SID
  - Node SID is a prefix SID with ‘N’ (node) bit set, it is associated with a host prefix (/32 or /128) that identifies the node, more than 1 Node SID’s per node can be configured (think router-id)

Short SR-MPLS recap – SID types

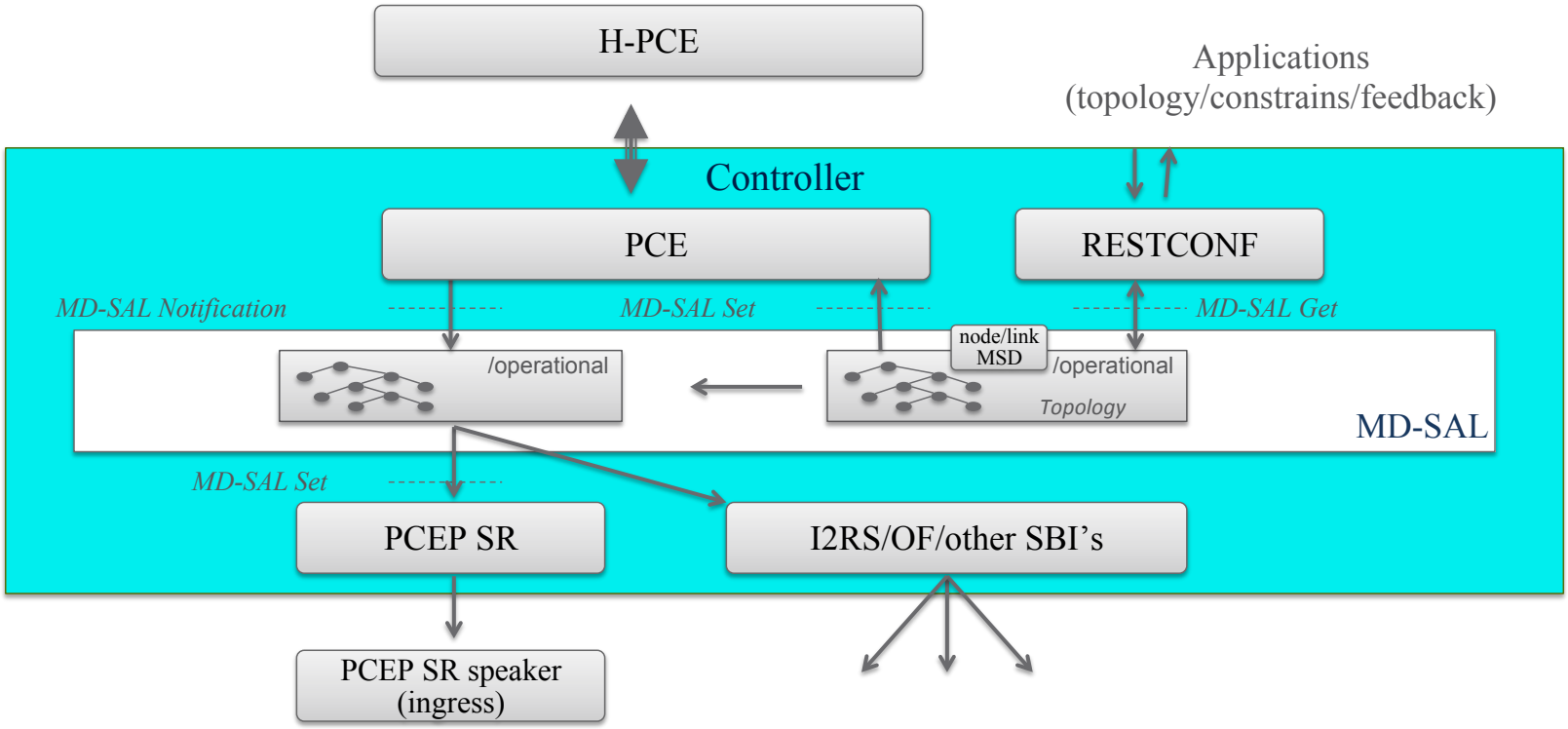
- Adjacency SID
  - Locally significant in most implementations– can be made globally significant thru ‘L’ flag
  - Identifies unidirectional adjacency
  - In most implementations automatically allocated for each adjacency
  - Always encoded as an absolute (not indexed) value

- Binding SID
  - Can be originated by any SR capable device in the SR domain
  - Can be used to instantiate a new label stack at the SID originating node (anchor), hence splitting end2end path into number of sub-paths

SR in SDN world – topology acquisition



SR in SDN world – SID stack provisioning



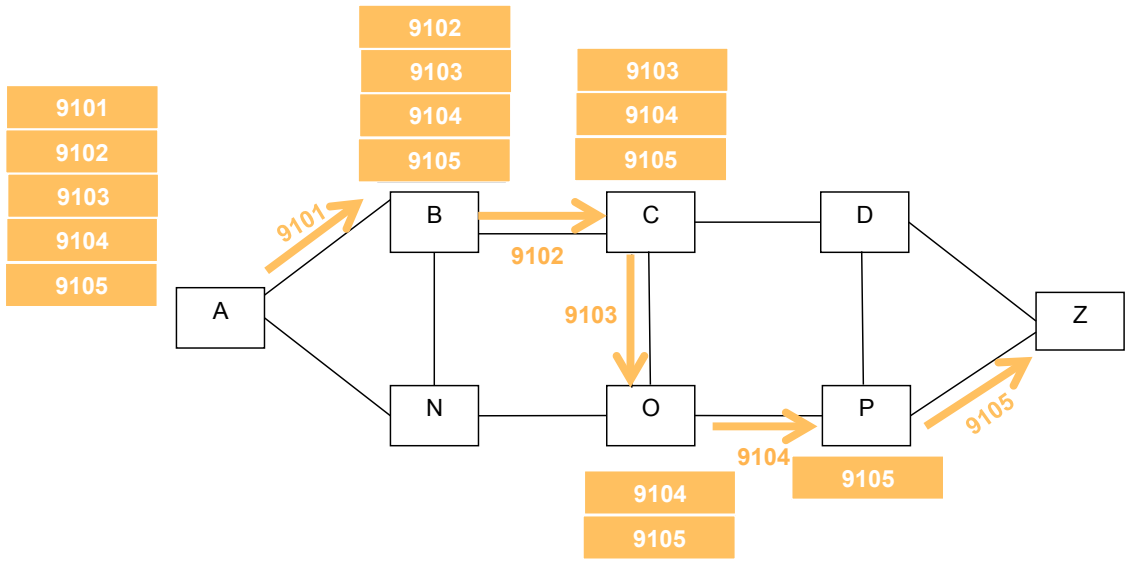
What's the problem?

- MSD supported by different HW/SW differs widely :
  - Linux (kernel 4.10): 2 SID's, some improvements recently (as of 4.11)
  - Low end off the shelf (merchant) silicon, e.g. BCM Trident2: 3-5 SID's
  - High end off the shelf (merchant) silicon, e.g BCM Jericho1 : 4-7 SID's
  - Vendor' silicon, e.g. Juniper' Trio, Nokia' FP3: 4-10+ SID's

- If SID stack > MSD at ingress node
  - Best case:
    - Service can't be provided
  - Worse case:
    - Packet will get dropped somewhere in the network

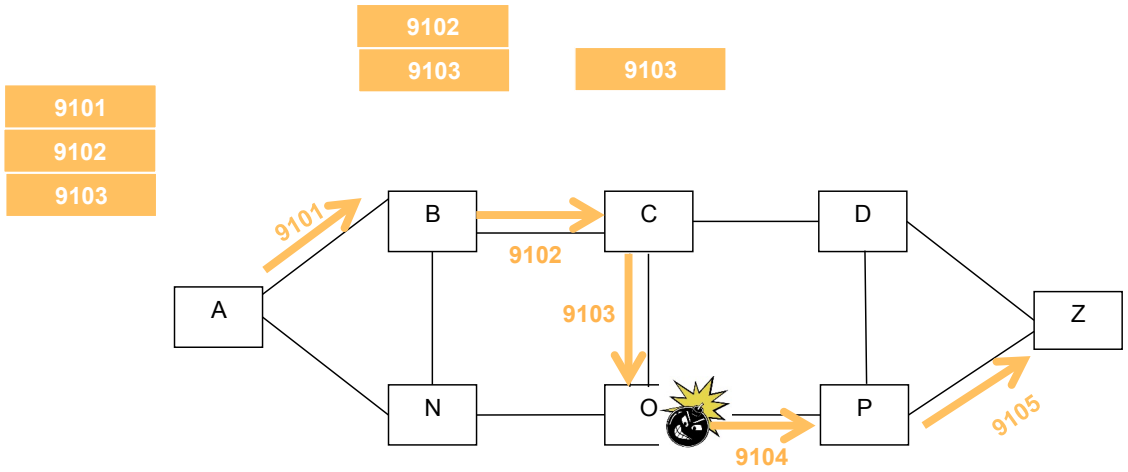
A path with Adjacency SID's (strict encoding)
MSD = 5

› Source routing along an explicit labeled path



A path with Adjacency SID's (strict encoding)
MSD = 3

› Source routing along an explicit labeled path



Possible solutions:
Control plane with exposure to SDNc is the right place to start!

- #1 SID stack compression
  - Efficient path computation algorithms
    - Compressed SID stack that meets MSD limit

- #2 SID stack expansion
  - Instantiate a new SID stack at the anchor node, keeping initial stack within ingress's MSD limits
    - Signaled thru Binding SID

Possible solutions - #1
SID stack compression

- SR-LEA
  - SR paths Label Encoding Algorithm
- SR-LEA-A
  - SR paths Label Encoding Algorithm with global Adj-SID's

Label Encoding Algorithm for MPLS Segment Routing

Rabah Guedrez* Olivier Dugeon* Samer Lahoud† Géraldine Texier‡
*Orange Labs, Lannion, France,
rabah.guedrez@orange.com, olivier.dugeon@orange.com
†*IRISA/Université de Rennes 1/Adopnet Team, Rennes, France, samer.lahoud@irisa.fr
‡IRISA/Télécom Bretagne/Adopnet Team, Rennes, France, geraldine.texier@telecom-bretagne.eu

http://ieeexplore.ieee.org/document/7778603/
http://samer.lahoud.fr/pub-pdf/icin-17.pdf

Possible solutions - #1
SID stack compression

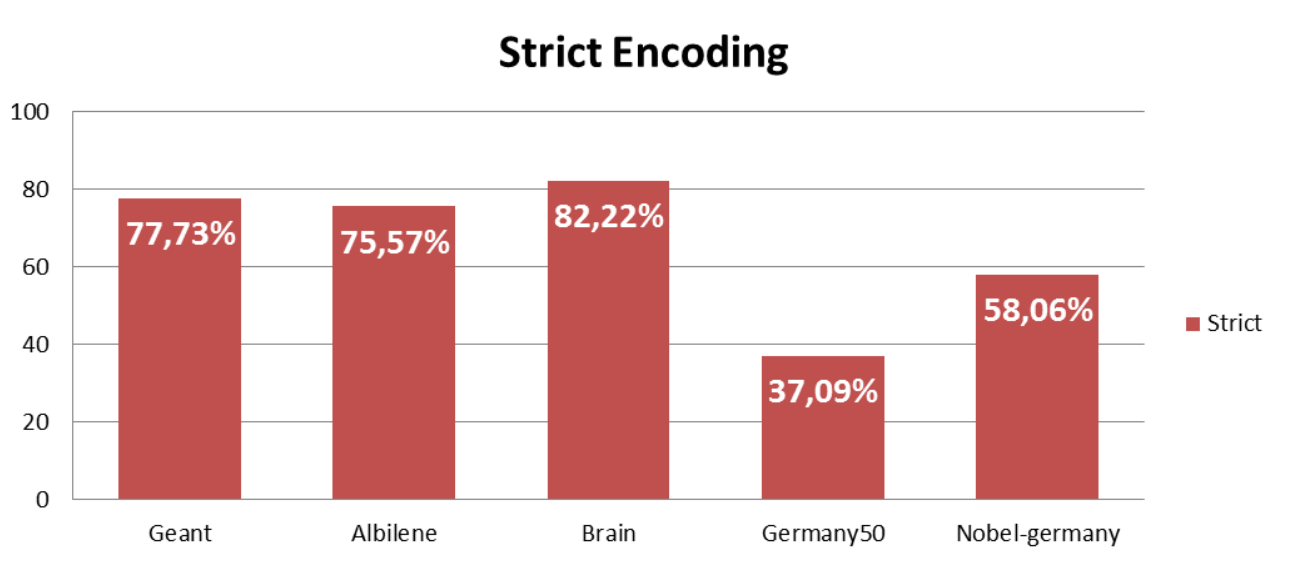
- Analysis based on topologies available from Network Design Library
  - The result is the optimal set of paths to satisfy the demand matrix
    - V: the number of nodes
    - E: the number of links
    - D: number of demands in the demand matrix

Topology	V	E	D
Geant	22	36	431
Albilene	12	18	131
Brain	161	166	9045
Germany50	50	80	1270
Nobel-germany	17	26	248

Possible solutions - #1
SID stack compression

› Analysis based on topologies available from Network Design Library

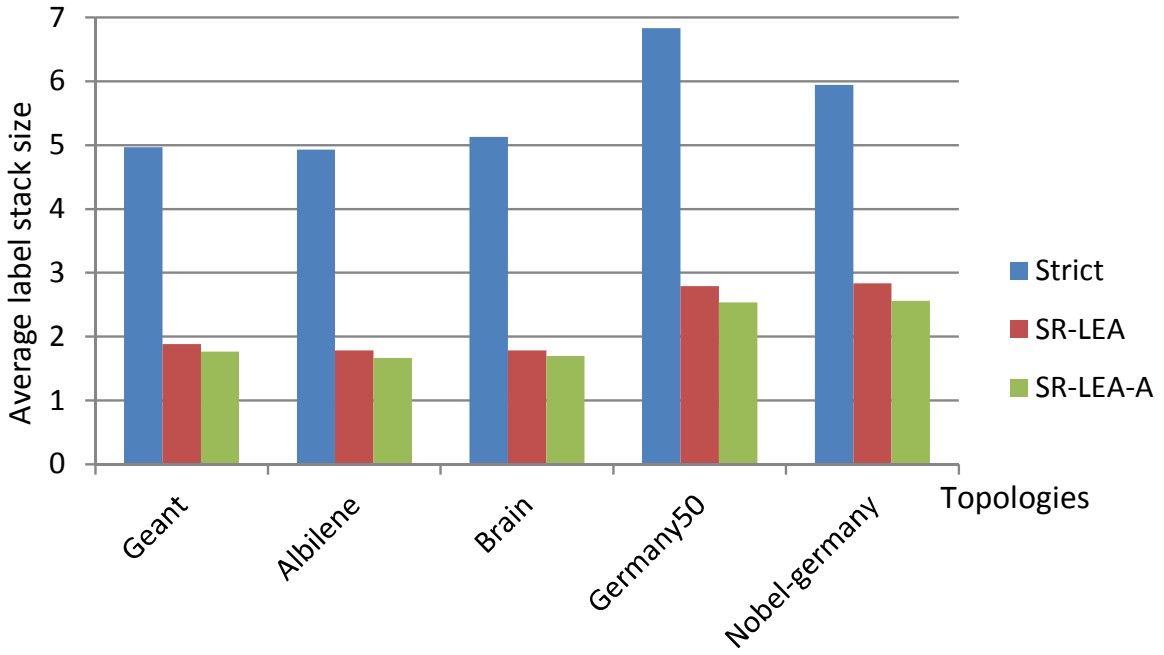
› % of usable paths satisfying service requests with MSD = 5



Possible solutions - #1
SID stack compression

› Analysis based on topologies available from Network Design Library

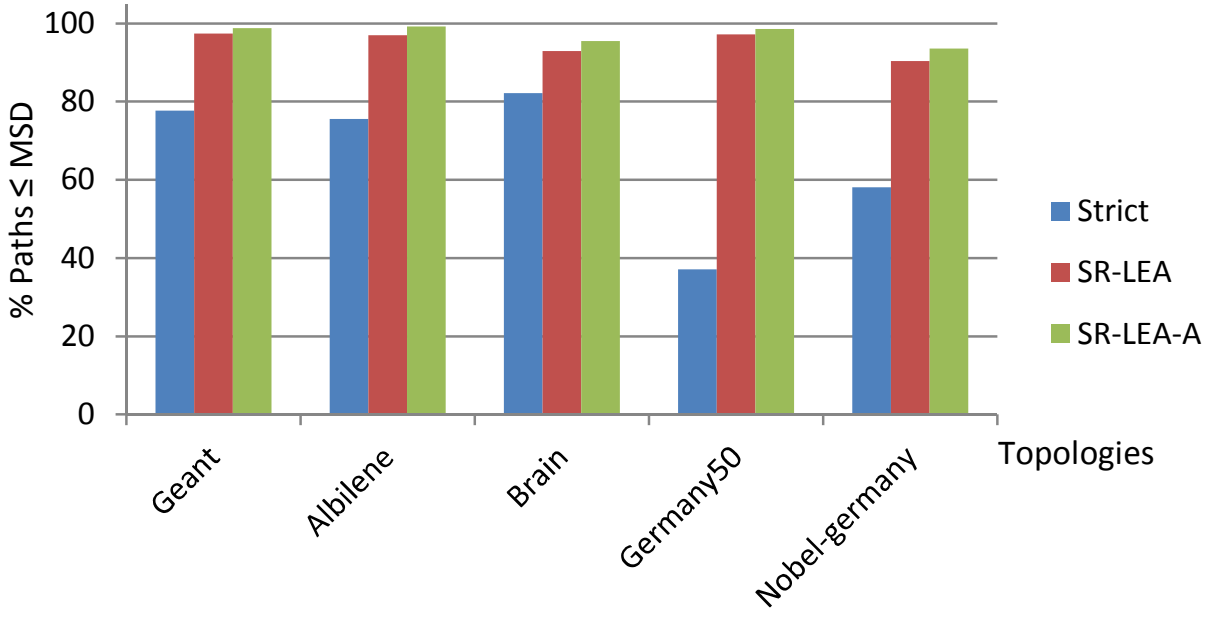
› Average SID stack satisfying service requests



Possible solutions - #1
SID stack compression

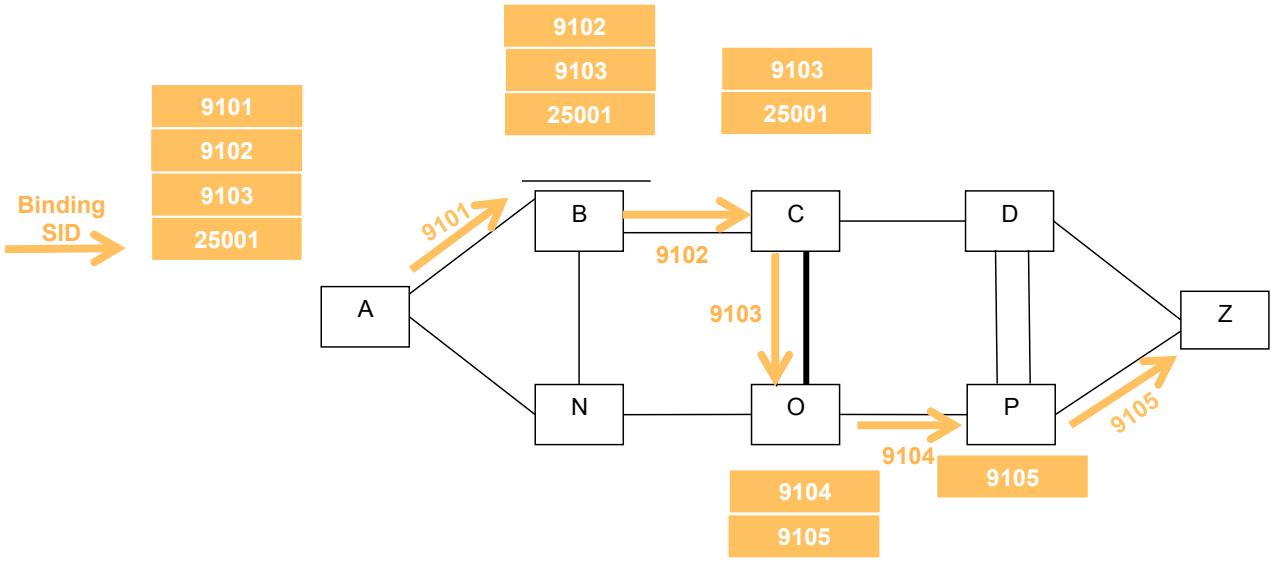
› Analysis based on topologies available from Network Design Library

› % of usable paths satisfying service requests with and without compression



Possible solutions - #2
SID stack expansion, MSD = 4

› Node O (anchor for Binding SID 25001) expands 25001 into new SID stack {9104,9105}



Signaling (standardization)

- IETF MSD drafts (ready for the Working Group Last Call)
  - OSPF
    - draft-ietf-ospf-segment-routing-msd
  - ISIS
    - draft-ietf-isis-segment-routing-msd
  - BGP-LS
    - draft-ietf-idr-bgp-ls-segment-routing-msd

- PCEP
  - Binding SID setup
    - draft-sivabalan-pce-binding-label-sid

- Common IANA registry for MSD types for OSPF/ISIS/BGP allows simplified and quick addition/ adoption of new types (IPv6, recirculation, entropy, etc)

Conclusions

- HW limitations are impairing service agility
  - TTM for a new ASIC is around 2 years
- Innovation in SW provides tangible results
  - Work in IETF ensures - the solution is technically sound and can interoperate
- Get your vendors to implement it ☺
  - First implementations are there (FRR, Cisco (planned))

Questions

Thank you!