Segment Routing: what marketing doesn’t talk about

Massimo Magnani – Systems Engineer

mmagnani@juniper.net

JUNIPER NETWORKS | Engineering Simplicity

Objective / Disclaimer

Objective:

- Let's start having operations-oriented discussions around segment routing

Disclaimer:

- This is a discussion of some of the details that don't come up when people are waxing poetic about segment routing
- Nothing discussed here is intractable - It's just work
- As an industry we are still working through many of these issues
  - It's going to take time
  - There will be bruises (and probably some scarring)
- This discussion assumes the desire to do something optimal with traffic
- If you're simply replacing LDP, most of this doesn't apply to you

AGENDA

- Segment routing in a FLASH!
- Obvious things
  - label management (space and stacks)
  - RSVP-TE and SR coexistence / migration
- Less obvious things
  - Controller care and feeding
  - SRTE protocols
  - traffic protection
- Summary

SEGMENT TYPES AND LABEL SPACES

BASIC SEGMENT TYPES

Adjacency-SID (single router hop)
- represents an IGP adjacency
- node-local significance

- Prefix-SID (one or more hops)
  - Represents IGP least cost path to a prefix
  - Node-SIDs are a special form of Prefix-SIDs bound to loopback
  - Domain-wide significance

ADVANCED SEGMENT TYPES

Anycast-SID (one or more hops)
- Represents IGP least cost path to a non-uniquely announced prefix

- Binding-SID
  - represents a tunnel

SEGMENT ID (SID) SPACE

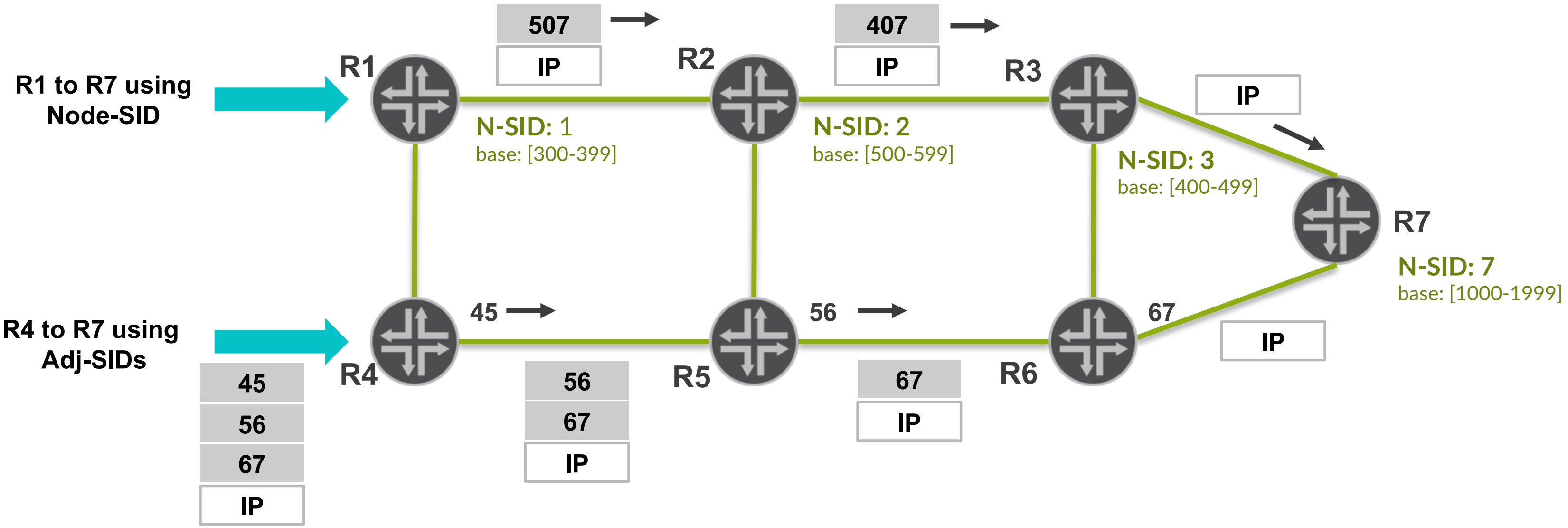
- SIDs are not labels
  - but - SIDs are encoded (carried) in labels
- Domain-wide SIDs coordinated via IGP
- Domain-wide SIDs are allocated in a manner much like RFC1918 addresses
  - Each node reserves a block of labels. this label block is the Segment Routing Global Block (SRGB).
  - Global label = SRGB base value + index

© 2018 Juniper Networks

BASIC SR FORWARDING EXAMPLES

Prefix/Node-SID forwarding (using SRGB)

- R1 shortest path to R7 is via R2.
- R2 expects a label value equal to {R2 label-base + index of destination}
  R1 => R2 label = 507 {500 + 7}

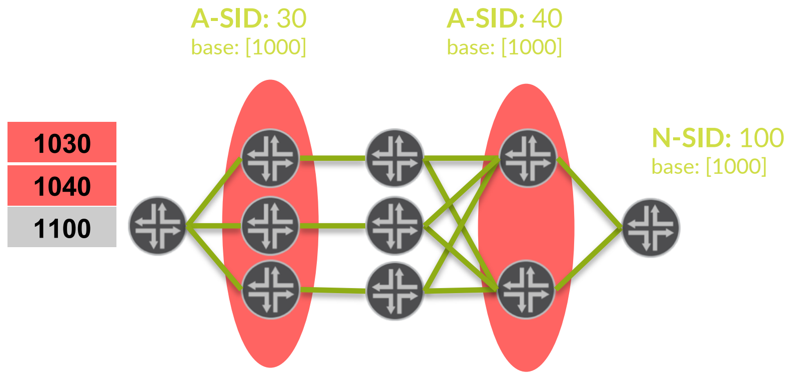


© 2018 Juniper Networks

ANYCAST-SIDs / Binding-SIDs

Anycast-SIDs

- Have domain-wide significance
- Define a set of nodes via a non-uniquely announced prefix
- Forwarding choice is made via IGP SPF
- Can use ECMP for forwarding
- Add redundancy, enable load balancing
- Commonly represent a set of geographically close nodes (e.g.: metro)

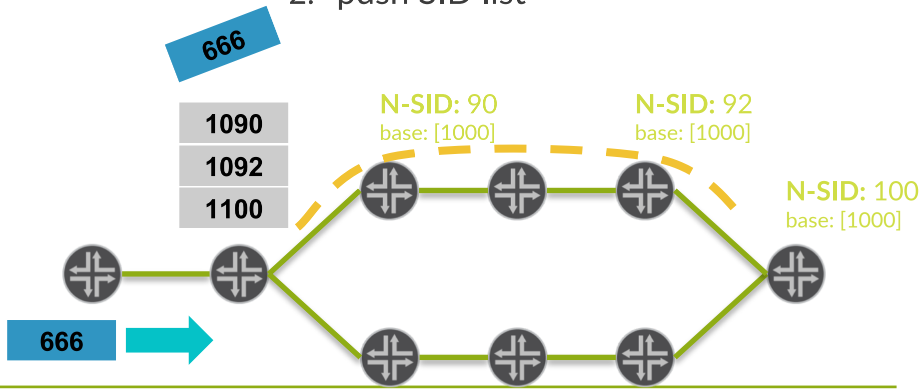


Binding-SIDs

- have node-local significance
- are bound to other SR paths
- enable an SR path to include another SR path by reference
- are useful for scaling the SID stack at ingress

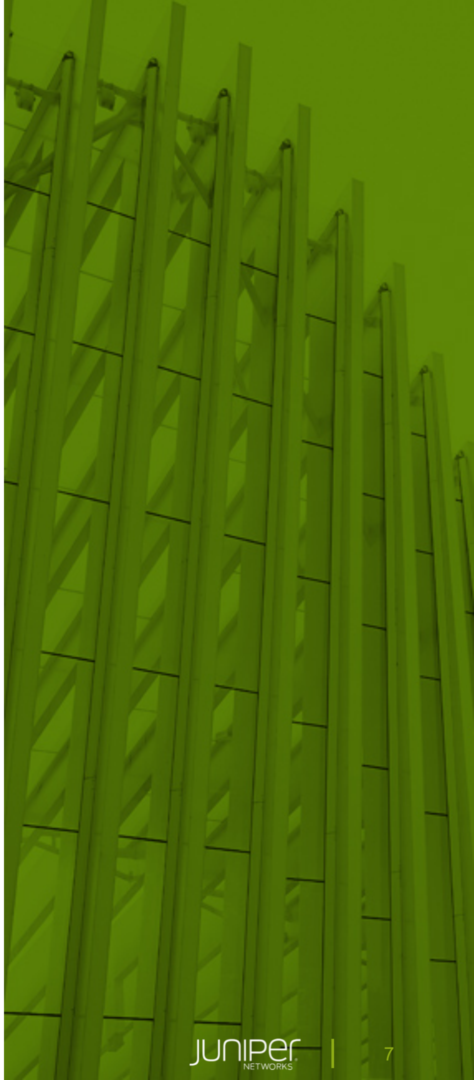
Binding-SID forwarding operation:

1. pop Binding-SID label
2. push SID list



© 2018 Juniper Networks

OBVIOUS STUFF



LABEL SPACE MANAGEMENT - GLOBAL LABELS

Global Label Space - Prefix-SIDs, Node-SIDs, Anycast-SIDs

- Operation of Prefix-SIDs is reasonably well established across implementations
- Anycast-SID operation may have SRGB-specific considerations
  - It is recommended that nodes announcing an Anycast-SID have an identical SRGB, drafts are reasonably explicit on this point
  - Further, labels after Anycast-SID must be resolvable by downstream nodes
- Anycast-SID has had interesting interop considerations
  - Behavior across major vendors has largely been clarified
  - However, there is still opportunity for misconfiguration and blackholing
    - e.g.: discontinuities in the resolution or announcement of Anycast-SIDs
  - Good News: Successful interop-tests already done @ EANTC (March 2018)

© 2018 Juniper Networks

Label Space Management - Local Labels

Local Label Space – Adjacency-SIDs, OAM labels, service-specific labels

There may be implementation subtleties in the operation and allocation of local label space

E.G.: some implementations have the concept of static or local service labels, the migration to SR may require managing through the allocation of these service-specific labels in your environment.

JUNOS supports both static and dynamic allocation models for Adjacency-SID

© 2018 Juniper Networks

LABEL STACK SIZE

SR provides for very granular traffic control, where the controller does explicit path specification with a combination of global and/or interface specific labels on the head of the packet.

Sounds great, doesn't it? But it carries additional considerations…

Hardware Encapsulation Capabilities - some hardware is severely constrained as to the number of labels that can be imposed in a single pass

- Includes some popular chipsets
- If you control one end of the connection you may be able to offload some label imposition processing to your host stack
- If you're a transit/network provider pay careful attention to the ingress (edge) hardware capabilities
- f you need very specific traffic engineering capabilities (read: link-specific placement) this is a notable consideration

© 2018 Juniper Networks

LABEL STACK SIZE

tl;dr - Make sure you understand your hardware capabilities and traffic behaviors.

deep label stacks have additional hardware considerations, beyond encapsulation.

- Transit Node/Link implications
  - Will all transit nodes support / forward deep label stacks?
  - On all line cards in the system?
- Load balancing considerations
  - For nodes that support forwarding deep label stacks what are the entropy sources available or activated?
  - Does use of deep label stacks obscure L3/L4 entropy sources that you really need to achieve load balancing objectives on LAGs?

“No worries! I’m going to use Anycast-SIDs and Prefix-SIDs to define paths and I’ll have a small label stack.” -- We’ll come back to this.

© 2018 Juniper Networks

RSVP/SR COEXISTENCE (AND MIGRATION)

2 parts to this discussion

- Objectives
- Control-plane behaviors and operation

Objectives

- Dominant assumption is that **migration** from RSVP to SR is the objective.
- If there is a long-term need to run both RSVP and SR on the same infrastructure – it's likely preferable to put both domains under a common controller as soon as possible
    - Particularly if P2MP-TE is in the mix

© 2018 Juniper Networks

RSVP/SR COEXISTENCE

Control-plane behaviors and operation

- Placement of SR LSPs in the same domain as RSVP-TE LSPs runs the risk of introducing inaccuracies in the TED that is used by distributed or centralized RSVP path computation engines
- Generic problem associated with management of dark bandwidth pools

draft-ietf-teas-sr-rsvp-coexistence-rec-04 in the work to address RSVP/SR Coexistence

© 2018 Juniper Networks

RSVP/SR COEXISTENCE SOLUTION OPTIONS (1)

Static Bandwidth Partitioning

- Reservable interface bandwidth is statically partitioned between SR and RSVP-TE
- Each operates within respective bandwidth allocation

Downside

Potentially strands bandwidth; protocols cannot use bandwidth left unused by the other protocol

Centralized Capacity Management

Central controller performs path placement for both RSVP-TE and SR LSPs

Downside

Requires the introduction of a central controller managing the RSVP-TE LSPs as a prerequisite to the deployment of any SR LSPs

© 2018 Juniper Networks

RSVP/SR COEXSITENCE SOLUTION OPTION(2)

Flooding SR Utilization in IGP	Running SR over RSVP-TE
SR utilization information can be flooded in IGP-TE and the RSVP-TE path computation engine (CSPF) can be changed to consider this information	Run SR over dedicated RSVP-TE LSPs that carry only SR traffic.
Downside ● Requires changes to the RSVP-TE path computation logic ● Carries upgrade requirement in deployments where distributed path computation is done across the network	**Downside** Requires SR to rely on RSVP-TE for deployment

© 2018 Juniper Networks

RSVP/SR COEXISTENCE SOLUTION OPTIONS (3)

Reflect SR traffic utilization by adjusting Max-Reservable-BW

- Dynamically measure SR traffic utilization on each TE interface and reduce the bandwidth allowed for use by RSVP-TE
- Incurs no change to existing RSVP path calculation procedure
- Assumes the use of Auto-BW w/i RSVP domain
- Controller may operate entirely within the context of the SR traffic domain

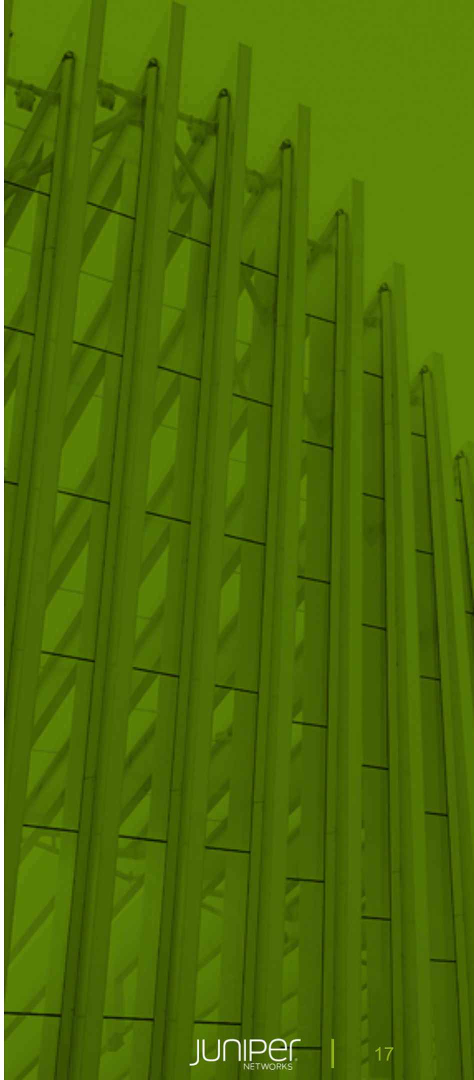
Reflection procedure on each TE node as follows:

- Periodically retrieve SR traffic statistics for each TE interface
- Periodically calculate SR traffic average over a set of collected traffic samples
- If the change in SR traffic average is greater than or equal to SR traffic threshold percentage (configured), adjust Max-Reservable-BW
  - Results in the RSVP-Unreserved-BW-At-Priority-X being adjusted
- RSVP-TE nodes can re-optimize LSPs accordingly

Implementations are shipping, Junos supports is today

© 2018 Juniper Networks

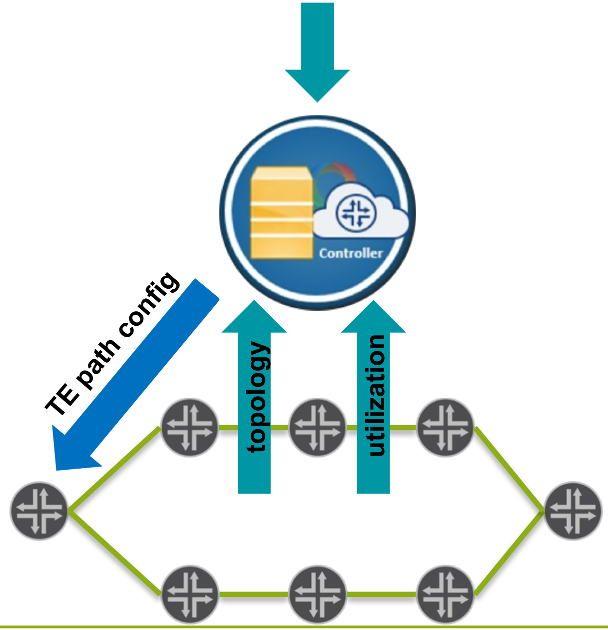
LESS OBVIOUS STUFF



CONTROLLER (+ COLLECTOR)

- Controller acquires LSDB
  - Passive IGP / BGP-LS / telemetry
- Controller understands current network state and utilization via collector
- Calculates traffic demands vs. capacity and availability requirement
  - Understands H/W capabilities
  - Aware of current and projected loads
- Controller sends segment list (path) to ingress router to place traffic
  - Configuration / BGP SRTE / PCEP
  - Other RIB programming mechanisms

- Business logic
- Workload demands
- Availability requirements
- Network capabilities



© 2018 Juniper Networks

CENTRALIZED PATH COMPUTATION

Benefits

- Centralized control has global view of reserved/available bandwidth
  - Not available at any other point in the network
- Facilitates analytics driven policy
  - Controller receives telemetry
  - Based on Telemetry, Controller configures / alters policy

Additional considerations

- Requires developing a controller or purchasing a controller
  - Staffing and ongoing maintenance of controller development
  - New deployment and/or vendor dependencies
- Concentrated point of failure / congestion
  - Risks mitigated by redundant controllers

© 2018 Juniper Networks

SR TRAFFIC ENGINEERING

A Brief Aside

With Segment Routing Traffic Engineering is now primarily Controller driven

- If there are Hardware constraints (on imposition or transit) the controller must calculate longest best paths taking into consideration Anycast/Prefix-SIDs
- Algorithms to compress the label stack are a hot area of optimization
- Some implementations are being extended to support dynamic, distributed computation with SR ingress nodes providing RSVP-like path calculation taking into consideration path constraints (Affinity, SRLGs, etc.)
  - For instance, JUNOS will support this starting with 19.2 release (Q2-2019);

© 2018 Juniper Networks

CONTROLLER CARE AND FEEDING

- To effectively place workloads on the network the controller must have visibility into current network utilization and loading
- A controller must respond to fluctuations in traffic quickly to prevent overloading hot links and gracefully migrate traffic loads
- Implies significantly more aggressive instrumentation cycles than is commonly seen in today’s networks with a complementary feedback loop to move workloads onto less-utilized paths / rebalance traffic
- Reworking instrumentation to utilize streaming telemetry is a practical day-0 requirement
- Per-label traffic statistics - something we’re now talking about

© 2018 Juniper Networks

WE NEED TO TALK ABOUT STATS

Given the Controller's need for stats, what does the hardware do?

- **It depends:** the ideal is per-interface, per-direction, per-label, per-class statistics, ditto for policy stats (draft-ali-spring-sr-traffic-accounting-02)
- Reality is far uglier
  - Outside of FIB and ACL space, counters are the most precious resource on modern ASICs
  - You're more likely to get a subset of the above (wish)list
- Getting stats off of network elements is another consideration
  - Per-interface, per-label statistics requires significant and often new collection infrastructure
- If you get some useful subset of stats info, what does a label counter get you?

© 2018 Juniper Networks

WHAT'S IN A COUNTER?

Anycast/Prefix-SIDs

- Present as a single counter for lots of traffic underneath
- What are the sources for all that traffic?
  - What's been merged underneath these labels?
  - Multiple ingress points in the network?
  - How do you find the right traffic to re-optimize?

SRTE policy counters

- How many policies may resolve to a common segment list?
- How many segment lists collapse to a common set of AnyCast/Prefix-SID destinations at midpoints?
- Will require planning on how to manage and instrument sources and sinks within the network

Punchline: double down on your investment in IPFIX / sFlow collection infra!

© 2018 Juniper Networks

SRTE PROTOCOLS: BGP SRTE

BGP SRTE

- The current draft remains an active area of development
- Provides useful capabilities in ECMP-dense environments
- No tunnel/virtual interface configuration, forwarding is instead tied to policy
    - Think “rules for steering” - not, explicit-path placement

- New considerations re: data-plane programming and validation
    - Q: How do you know the node accepted the list of segment lists you sent it?
    - Q: How do you know what might have been tangled up in policy logic?
    - A: You don’t. You’ll have to ask the node afterwards. You’ll want telemetry for that.
- Q: do you need to specify a protection / bypass path?
    - This might not be the tool you’re looking for

© 2018 Juniper Networks

SRTE PROTOCOLS: PCEP

PCEP (Stateful)

- Provides single protocol for the management of RSVP and SR paths
- Flexible management and delegation models
- Requires additional mechanisms for prefix binding and flow specification
- Has an RSVP-ish operational view
  - Capable of signaling SR paths; traffic / flow-mapping is work-in-progress
  - Protection path placement pending … (resurrect the local protection-draft)
- Provides options for some form of contract with the ingress nodes
  - Can the hardware do what you asked of it?
  - With PCEP the controller can understand node capabilities and act accordingly

© 2018 Juniper Networks

SRTE PROTOCOLS: RPC-BASED PATH PLACEMENT

Emergent RPC-based mechanisms for path placement

Some operators are looking to leverage RIB APIs available from vendors and modeling consortia

- pRPD from Juniper (https://juni.pr/2rtY2fV)
- gRIBI from OpenConfig (https://bit.ly/2HZwN7i)
- EOS APIs from arista (https://bit.ly/2xuHNVp)
- Service layer APIs from Cisco (https://bit.ly/2fRvzhz)

RIB APIs

- Commonly provide mechanisms to define label stacks / paths
- Provide mechanisms to associate RIB entries with these paths
- Enable new controller selection models
- Use modern software development tools
  - Leverage widely available tools & protocols
  - Make your developers happy(-ish)
  - Enables more sophisticated error-handling

Additional considerations:

- Requires internal development expertise
- Commonly leveraging a vendor-specific interfaces
  - associated API management policies
  - new test, cert and deployment packaging considerations

© 2018 Juniper Networks

SRTE TRAFFIC PROTECTION

- It's 1AM, do you know what your protect path is?
- Did you get to specify it? Probably not.
- How much traffic is going to go over that path? Are you sure?
  - **TI-LFA is commonly the reflexive response for SR traffic protection**

Lots to like

- No midpoint state
- True post-convergence path provides optimality - no u-loops!
- Cool sounding acronym

Practical reality

- Computationally intensive
  - Particularly if SRLGs, etc. in the mix
- May not be deterministic
  - Particularly across vendors
- May require label stack compression to stay within protection encapsulation capabilities
- Ref. prior conversation about counters and load placement (or finding big flows)

Deployment considerations

- Protect path placement remains an active area of development
- Operators requiring explicit protection placement and an understanding of protect path capacity will want to understand available TI-LFA behaviors deeply or explore other options

© 2018 Juniper Networks

SUMMARY

- TE didn’t really get easier - It just got different
- Lots of work remains to operationalize segment routing for traffic engineering
- Data Plane simplification and elimination of control plane state network means building new infrastructure to account for lost or shifted functionality
- Vendors are actively developing the tooling to make deployments happen
- In the meantime
    - Expect considerable variability in implementation capabilities and installed footprint
    - Be prepared to roll your own solutions to some of these problems

Look forward to more ITNOG discussion around these topics as we, as an industry, gain operational experience

© 2018 Juniper Networks

Thank you.

Juniper Networks Italy – Massimo Magnani