ietfjournal.org

Cutting Through the Hype and Finding the IETF's Innovative Nugget of Gold – IETF Journal

16-21 minutes

Segment Routing (SR) is a new traffic-engineering technology being developed by the IETF's SPRING Working Group. Two forwarding plane encapsulations are being defined for SR: Multiprotocol Label Switching (MPLS) and IPv6 with a Segment Routing Extension Header. This article provides some historical context by describing the MPLS forwarding plane and control plane protocols, explains how Segment Routing works, introduces the MPLS-SR forwarding plane, and shows how the SR control plane is used. Finally, the article compares SR with legacy MPLS systems, and identifies its unique merits.

MPLS Forwarding

MPLS is a nearly 20-year-old technology. An MPLS domain is a contiguous set of Label Switching Routers (LSRs). Packets enter the MPLS domain through an ingress LSR and exit the MPLS domain through an egress LSR. A single LSR can serve as ingress for some packets and egress for others.

A Label Switched Path (LSP) provides connectivity between an ingress LSR and an egress LSR. An LSP can traverse the least-

cost path or it can traverse a traffic-engineered path.

When an ingress LSR receives a packet, it assigns the packet to Forwarding Equivalence Class (FEC) and encapsulates the packet with an MPLS label stack. It then forwards the packet to the next-hop associated with the FEC.

The MPLS label stack contains one or more label stack entries. Each label stack entry contains a label, a time-to-live (TTL) indicator, a Traffic Class (TC) indicator, and a bottom of stack indicator. These data items determine how a transit LSR will process the packet. In that respect, each label stack entry is an instruction to an LSR.

When an LSR receives a packet it examines the top entry in the label stack and decrements the TTL. If the TTL has not expired, the LSR searches its Forwarding Information Base (FIB) looking for an entry that matches the incoming label.

If the LSR finds a FIB entry that matches the incoming label, the FIB entry will contain the following information:

- Label action
- Next-hop interface

Label actions are the following:

- Push one or more new entries onto the label stack.
- Pop the top entry from the label stack.
- Swap the label in the top entry.

Having found a matching FIB entry, the LSR executes the label action and forwards the packet through the next-hop interface. The next-hop interface can be an internal interface or an external

interface. If the next-hop interface is an internal interface, the LSR forwards the packet to itself and processes the packet as it had just been received, examining outermost protocol header. If the next-hop interface is an external interface, the LSR forwards the packet appropriately.

When a packet reaches the penultimate hop on an LSP, the LSR may pop the final label stack entry and forward the payload packet without any encapsulation.

MPLS Control Plane

Routing Protocols

An MPLS network makes heavy use of the Interior Gateway Protocols (IGPs)—Open Shortest Path First (OSPF) or Intermediate System-to-Intermediate System (IS-IS)—to learn the network topology, establish the least cost paths, and provide information for computing traffic engineering paths. Normal IGP advertisements are used to distribute the connectivity and metrics for the network links, and those messages are enhanced with additional information describing the links (such as bandwidth).

Label Distribution Protocol (LDP)

LDP is a TCP-based protocol that can be run between adjacent LSRs in an MPLS network. Each LSR uses the protocol to advertise the label to use when MPLS encapsulated packets are sent to it for final delivery to an IP prefix. As each LSR receives advertisements from other LSRs it is able to install entries in its FIB showing how to map from the label in a packet it receives (a label it has advertised) to a label in a packet it forwards (a label it has

received in an advertisement).

LDP results in traffic being forwarded along the least cost path and does not support traffic engineering.

Resource Reservation Protocol with TE Extensions (RSVP-TE)

In RSVP-TE, network operators administratively assign TE attributes to interfaces. TE attributes include, but are not limited to, available bandwidth, reserved bandwidth and administrative color. These TE attributes are flooded by the IGP so that each node within the IGP domain maintains an identical copy of a Link State Database (LSDB) and a Traffic Engineering Database (TED). The LSDB describes the IGP topology, while the TED augments the LSDB with TE link attributes.

Network operators request LSPs that meet specific constraints. For example, a network operator could request an LSP that originates at Node A, terminates at Node Z, reserves 100 megabits per second, and traverses blue interfaces only. A path computation module, located on a central controller—such as the Path Computation Element (PCE)—or on the ingress LSR, computes a path that satisfies all of the constraints. In order to construct this SR-path, the path computation function consults the LSDB and TED.

RSVP-TE is a signaling protocol that runs directly over IP. It uses a Path message to signal out along the path of the LSP, and a Resv message is returned to reserve network resources and confirm the establishment of the LSP. The Path message contains details of the requested LSP (bandwidth, etc.) as well as an Explicit Route Object (ERO) that lists the nodes and links that the LSP should traverse.

The Resv message reports the resources that have been reserved (bandwidth, etc.) and a Record Route Object (RRO) that confirms the path of the LSP.

Each LSR selects a label that it will use to receive traffic on the LSP. It includes this label in the Resv message it sends. Thus, each LSR can build a FIB entry for the LSP mapping the label it has advertised to the label it has received.

RSVP-TE requires that state is maintained in the network for each LSP, and the protocol is a "soft state protocol", meaning that Path and Resv messages must be exchanged periodically to keep the LSP active.

Segment Routing

Terminology

An SR domain is a contiguous set of SR-capable routers. An SR-Path (i.e., an SR-signaled LSP) provides connectivity through the SR domain. An SR-path can traverse the IGP least cost path between its endpoints. It can also traverse a traffic-engineered path.

An SR-path contains one or more segments and a segment contains one or more router hops. The SPRING WG has proposed many segment types. However, the following segment types are most common:

- Adjacency
- Prefix
- Anycast

- Binding

  Adjacency segments represent an IGP adjacency between two routers. They typically contain one router hop, but can contain more. Prefix segments represent the IGP least cost path between any router and a specified prefix. Prefix segments contain one or more router hops. Anycast segments are like prefix segments in that they represent the IGP least cost path between any router and a specified prefix. However, the specified prefix can be advertised from multiple points in the network. Binding prefixes represent tunnels through the SR domain. The tunnel can be another SR-Path, an LDP-signaled LSP, an RSVP-TE signaled LSP, or any other encapsulation.

  A Segment Identifier (SID) identifies each segment. SIDs that represent prefix and anycast segments have domain-wide significance. Therefore, network operators allocate them using procedures that are similar to those used to allocate private IP (i.e., RFC 1918) addresses. Conversely, SIDs that represent adjacency and binding segments have local significance only. SR-capable routers allocate these SIDs automatically, without concern for domain-wide coordination.

  Every SID maps to an MPLS label. As stated above, MPLS labels have local significant only. Therefore, SIDs that have local significance only can map directly to MPLS labels. However, SIDs that have domain-wide significance require special treatment.

  Each SR-capable router reserves a range of MPLS labels, called the SR Global Block (SRGB). For example, Router A might reserve labels 10,000 through 20,000, while Router B reserves labels 20,000 through 40,000. Both routers map SIDs to MPLS labels by

adding the SID to the lowest SRGB value. Therefore, Router A maps SID 1 to MPLS label 10,001, while Router B maps the same SID to MPLS label 20,001.

SR Forwarding

When an SR ingress router receives a packet, it assigns the packet to FEC and encapsulates it in an MPLS label stack. Finally, it forwards the packet to the next-hop associated with the FEC.

The MPLS label stack represents an SR-path that is associated with the FEC. Each entry in the label stack represents a segment in the SR-path.

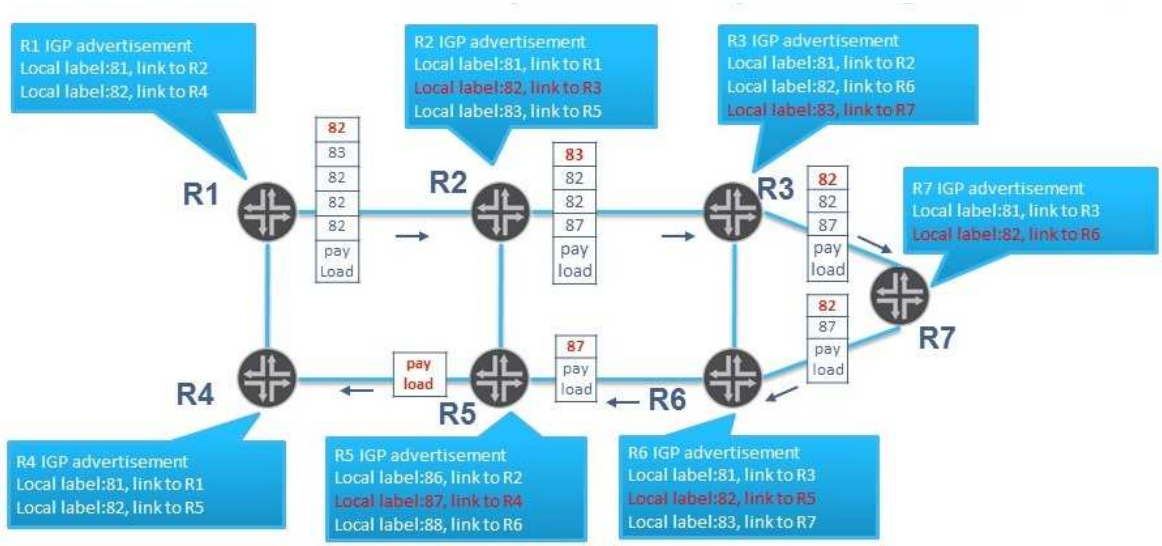


Figure 1. Adjacency Segments

In Figure 1, R1 maintains an SR-path to R4. The SR-path contains five adjacency segments, originating at R2, R3, R7, R6, and R5. The ingress LSR (R1) imposes a label stack with one entry for each adjacency segment. Finally, R1 forwards the packet to R2, where the first adjacency segment begins. R2 processes the outer label stack entry, popping it and forwarding the packet to R3. Each downstream LSR repeats the process until the packet arrives at R4.

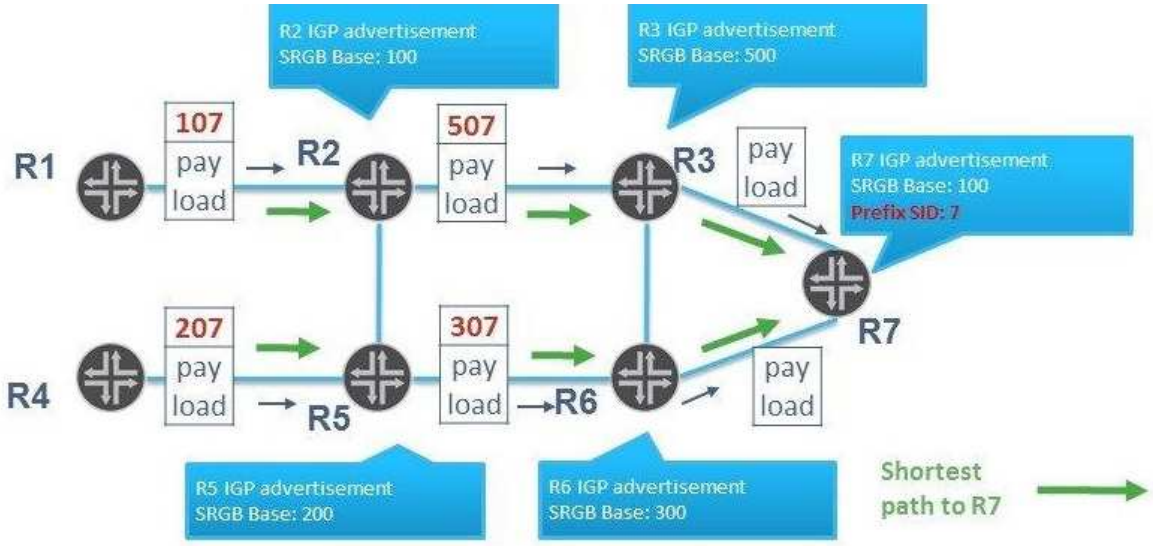
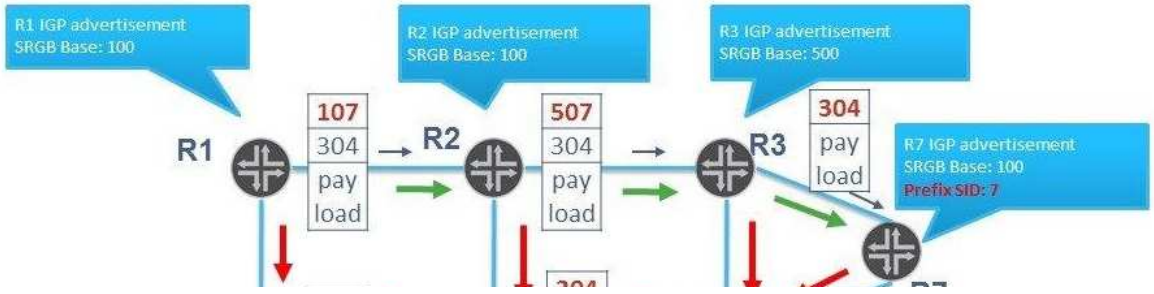


Figure 2. Single Prefix Segment

In Figure 2, R1 through R6 all maintain an SR-path to R7. The SR-path contains a single prefix segment, represented by SID 7. We will examine the path from R4 to R7.

The ingress router (R4) imposes a label stack that contains exactly one entry, representing the prefix segment (i.e., the IGP least cost path) between R4 and R7. This label stack entry carries a label that corresponds to SID 7. In order to calculate that label, R4 adds the SID (7) to the SRGB base advertised by the next-hop, R5 (i.e., 200). The result is 207. Finally, R4 forwards the packet to R5.

R5 processes the label. In order to do so, it identifies the router on the IGP least cost path to R7 (i.e., R6). Then R5 swaps the label, replacing it with the value that R6 maps to SID 7 (i.e., 307). Finally, it forwards the packet to R6. R6 repeats this process and the packet arrives at R7.



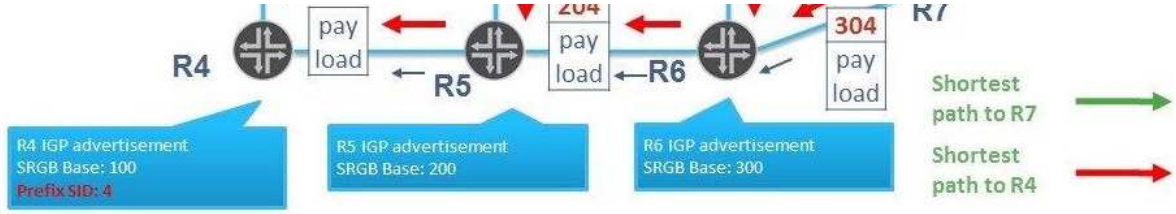


Figure 3. Traffic-Engineering Using Prefix Segments

In Figure 3, R1 maintains a traffic-engineered SR-path to R4 via R7. The SR-path contains two prefix segments. One prefix segment represents the IGP least cost path from R1 to R7, while the other represents the IGP least cost path from R7 to R4

The ingress LSR (R1) imposes a label stack with one entry representing each prefix segment. It calculates the inner label value by adding R4's SID (4) to R7's SRGB base (300). It calculates the outer label by adding R7's SID (7) to R2's SRGB base. Finally, R1 forwards the packet to R2. All downstream routers process the packet as described in the previous example and the packet arrives at R4.

IGP Extensions for Segment Routing

Each SR-capable router allocates a SID and a label for the following:

- Each prefix or anycast segment that it terminates
- Each adjacency or binding segment that it originates

Having done so, it creates a RIB entry for each of the above and installs the RIB entries into the FIB.

Next, the SR-capable router advertises the following into its IGP:

- Its SRGB characteristics
- Each prefix or anycast segment that it terminates

- Each adjacency or binding segment that it originates

  The IGP floods this data, in addition to the previously mentioned TE link attributes, throughout the IGP domain. Therefore, each node within the IGP domain maintains an identical copy of a Link State Database (LSDB) and a Traffic Engineering Database (TED). The LSDB describes the IGP topology, including SIDs and SRGB data, while the TED augments the LSDB with TE link attributes.

  When flooding is complete, every node within the IGP domain constructs two RIB entries for each prefix or anycast segment that it does not terminate. The first RIB entry instructs the local device to process all incoming IP traffic bound for the prefix as follows:

- Push an MPLS label stack entry whose label maps to the SID.
- Forward the packet to the next-hop on the IGP least cost path to the segment endpoint.

  The second RIB entry instructs the local device to process all incoming MPLS traffic whose outermost label maps to the segment as follows:

- Swap the outermost label, accounting for the next-hop's SRGB.
- Forward the packet to the next-hop on the IGP least cost path to the segment endpoint.

Path Computation

A path computation function calculates SR-paths. Given a set of TE constraints, the path computation function yields an MPLS label stack representing an SR-path that satisfies the constraints. In order to construct this SR-path, the path computation function consults the LSDB and TED.

The path computation function can reside on a central controller. Conversely, the path computation function can be distributed among ingress LSRs.

Analysis

LDP and RSVP-TE are end-to-end signaling protocols that establish per-LSP forwarding state in LSRs. Because LDP and RSVP-TE maintain all required forwarding state in LSRs, an LDP, or RSVP-TE signaled LSP can be represented by a single MPLS label stack entry.

By contrast, SR moves some, but not all, forwarding state from the network to the packet. An SR-path is represented by a label stack, with one label stack entry representing each segment in the SR-path. Therefore, the network maintains enough state to route the packet form segment ingress to segment egress, while the packet maintains enough state to route the packet from segment to segment.

By moving state from the network to the packet, SR reduces the amount of memory that LSRs require and the amount of processing needed to maintain state. Recent increases in CPU and memory within routers, and improvements to the RSVP-TE protocol and to implementations have reduced this issue, but it remains an important concern.

A more-significant benefit of moving state from the network to the packet is that it eliminates the need for an end-to-end signaling protocol. While SR requires an IGP and a path computation module, it does not require a signaling protocol like LDP or RSVP-TE.

However, some advanced functions offered by RSVP-TE rely on end-to-end signaling and per-LSP state in the network. Among these are bandwidth reservation, failure detection, and fast-reroute.

In RSVP-TE, the path computation function can be distributed among ingress LSRs, even when TE constraints include bandwidth reservations. This is possible because in RSVP-TE, each LSR maintains state for each LSP that it supports. Having this state, it can compute the remaining bandwidth on each RSVP-enabled interface and flood that information into the IGP. Therefore, every node in the IGP maintains an LSDB and TED with sufficient information to support the path computation function.

In SR, no such mechanism exists. So, when TE constraints include bandwidth reservations, the path computation function must be centralized in a controller where a global view of bandwidth allocation is available.

In RSVP-TE, the end-to-end signaling mechanisms also provides OAM functionality. When an RSVP-TE neighboring session fails, the LSR upstream of the failure signals the ingress LSR, causing it to invoke head-end restoration procedures. If configured to do so, the LSR upstream of the failure can also invoke local restoration procedures.

In SR, restoration is more complex. If the failure occurs at a segment ingress, some OAM mechanism outside of SR detects the failure and informs the path computation module. The path computation module invokes head-end restoration procedures, recalculating the SR-path between the SR ingress and the SR egress. While local restoration procedures for SR are conceivable, none have been standardized to date.

If a failure occurs at some point other that the segment endpoint, SR relies on external recovery mechanisms. For example, if a failure occurs in the middle of a prefix segment, SR relies on an IGP to detect the failure, flood topology changes, and compute the new IGP least-cost path to the segment endpoint. In this example, TI-LFA can be deployed to reduce dependence upon IGP convergence.

Conclusion

SR supports traffic engineering while reducing the amount of state maintained by the network. In many cases, SR eliminates the need for MPLS signaling protocols (i.e., LDP and RSVP-TE). For these reasons, the IETF should continue to develop SR capabilities.

Specifically, IETF should continue to develop IGP extensions for SR, as well as BGP extensions that may be required to extend SR across IGP boundaries. Additional work is essential to develop key networking functions such as OAM and ways to carry entropy to resolve ECMP choices. Furthermore, network equipment vendors and network operators should work together to prototype and experiment with SR to provide operational feedback to the IETF, so that SR can be improved and made ready for wide-scale deployment.

It is likely that network operators will deploy SR incrementally over the next several years. As deployments proceed, the SR community will gain operational experience, SR standards will be refined to address unforeseen problems, and implementations will improve accordingly. Furthermore, network operators will identify use-cases for which SR is well suited, as well as use-cases for

which LDP and RSVP-TE may be better suited.

For these reasons, as well as to support a massive installed base, the IETF and network equipment vendors should continue to refine and support LDP and RSVP-TE with the same intensity that they progress SR.