Internet Engineering Task Force (IETF)
Request for Comments: 8402
Category: Standards Track
ISSN: 2070-1721

C. Filsfils, Ed.
S. Previdi, Ed.
L. Ginsberg
Cisco Systems, Inc.
B. Decraene
S. Litkowski
Orange
R. Shakir
Google, Inc.

July 2018

Segment Routing Architecture

Abstract

Segment Routing (SR) leverages the source routing paradigm. A node steers a packet through an ordered list of instructions, called "segments". A segment can represent any instruction, topological or service based. A segment can have a semantic local to an SR node or global within an SR domain. SR provides a mechanism that allows a flow to be restricted to a specific topological path, while maintaining per-flow state only at the ingress node(s) to the SR domain.

SR can be directly applied to the MPLS architecture with no change to the forwarding plane. A segment is encoded as an MPLS label. An ordered list of segments is encoded as a stack of labels. The segment to process is on the top of the stack. Upon completion of a segment, the related label is popped from the stack.

SR can be applied to the IPv6 architecture, with a new type of routing header. A segment is encoded as an IPv6 address. An ordered list of segments is encoded as an ordered list of IPv6 addresses in the routing header. The active segment is indicated by the Destination Address (DA) of the packet. The next active segment is indicated by a pointer in the new routing header.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc8402.

Copyright Notice

Table of Contents

1. Introduction

Segment Routing (SR) leverages the source routing paradigm. A node steers a packet through an SR Policy instantiated as an ordered list of instructions called "segments". A segment can represent any instruction, topological or service based. A segment can have a semantic local to an SR node or global within an SR domain. SR supports per-flow explicit routing while maintaining per-flow state only at the ingress nodes to the SR domain.

A segment is often referred to by its Segment Identifier (SID).

A segment may be associated with a topological instruction. A topological local segment may instruct a node to forward the packet via a specific outgoing interface. A topological global segment may instruct an SR domain to forward the packet via a specific path to a destination. Different segments may exist for the same destination, each with different path objectives (e.g., which metric is minimized, what constraints are specified).

A segment may be associated with a service instruction (e.g., the packet should be processed by a container or Virtual Machine (VM) associated with the segment). A segment may be associated with a QoS treatment (e.g., shape the packets received with this segment at x Mbps).

The SR architecture supports any type of instruction associated with a segment.

The SR architecture supports any type of control plane: distributed, centralized, or hybrid.

In a distributed scenario, the segments are allocated and signaled by IS-IS or OSPF or BGP. A node individually decides to steer packets on an SR Policy (e.g., pre-computed local protection [RFC8355]). A node individually computes the SR Policy.

In a centralized scenario, the segments are allocated and instantiated by an SR controller. The SR controller decides which nodes need to steer which packets on which source-routed policies. The SR controller computes the source-routed policies. The SR architecture does not restrict how the controller programs the network. Likely options are Network Configuration Protocol (NETCONF), Path Computation Element Communication Protocol (PCEP), and BGP. The SR architecture does not restrict the number of SR controllers. Specifically, multiple SR controllers may program the same SR domain. The SR architecture allows these SR controllers to discover which SIDs are instantiated at which nodes and which sets of local (SRLB) and global (SRGB) labels are available at which node.

A hybrid scenario complements a base distributed control plane with a centralized controller. For example, when the destination is outside the IGP domain, the SR controller may compute an SR Policy on behalf of an IGP node. The SR architecture does not restrict how the nodes that are part of the distributed control plane interact with the SR controller. Likely options are PCEP and BGP.

Hosts MAY be part of an SR domain. A centralized controller can inform hosts about policies either by pushing these policies to hosts or by responding to requests from hosts.

The SR architecture can be instantiated on various data planes. This document introduces two data-plane instantiations of SR: SR over MPLS (SR-MPLS) and SR over IPv6 (SRv6).

SR can be directly applied to the MPLS architecture with no change to the forwarding plane [SR-MPLS]. A segment is encoded as an MPLS label. An SR Policy is instantiated as a stack of labels. The segment to process (the active segment) is on the top of the stack. Upon completion of a segment, the related label is popped from the stack.

SR can be applied to the IPv6 architecture with a new type of routing header called the SR Header (SRH) [IPv6-SRH]. An instruction is associated with a segment and encoded as an IPv6 address. An SRv6 segment is also called an SRv6 SID. An SR Policy is instantiated as an ordered list of SRv6 SIDs in the routing header. The active segment is indicated by the Destination Address (DA) of the packet. The next active segment is indicated by the SegmentsLeft (SL) pointer in the SRH. When an SRv6 SID is completed, the SL is decremented and the next segment is copied to the DA. When a packet is steered on an SR Policy, the related SRH is added to the packet.

In the context of an IGP-based distributed control plane, two topological segments are defined: the IGP-Adjacency segment and the IGP-Prefix segment.

In the context of a BGP-based distributed control plane, two topological segments are defined: the BGP peering segment and the BGP-Prefix segment.

The headend of an SR Policy binds a SID (called a Binding segment or BSID) to its policy. When the headend receives a packet with active segment matching the BSID of a local SR Policy, the headend steers the packet into the associated SR Policy.

This document defines the IGP, BGP, and Binding segments for the SR-MPLS and SRv6 data planes.

Note: This document defines the architecture for Segment Routing, including definitions of basic objects and functions and a description of the overall design. It does NOT define the means of implementing the architecture -- that is contained in numerous referenced documents, some of which are mentioned in this document as a convenience to the reader.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

SR-MPLS: the instantiation of SR on the MPLS data plane.

SRv6: the instantiation of SR on the IPv6 data plane.

Segment: an instruction a node executes on the incoming packet (e.g., forward packet according to shortest path to destination, or, forward packet through a specific interface, or, deliver the packet to a given application/service instance).

SID: a segment identifier. Note that the term SID is commonly used in place of the term "Segment", though this is technically imprecise as it overlooks any necessary translation.

SR-MPLS SID: an MPLS label or an index value into an MPLS label space explicitly associated with the segment.

SRv6 SID: an IPv6 address explicitly associated with the segment.

Segment Routing domain (SR domain): the set of nodes participating in the source-based routing model. These nodes may be connected to the same physical infrastructure (e.g., a Service Provider’s network). They may as well be remotely connected to each other (e.g., an enterprise VPN or an overlay). If multiple protocol instances are deployed, the SR domain most commonly includes all of the protocol instances in a network. However, some deployments may wish to subdivide the network into multiple SR domains, each of which includes one or more protocol instances. It is expected that all nodes in an SR domain are managed by the same administrative entity.

Active Segment: the segment that is used by the receiving router to process the packet. In the MPLS data plane, it is the top label. In the IPv6 data plane, it is the destination address [IPv6-SRH].

PUSH: the operation consisting of the insertion of a segment at the top of the segment list. In SR-MPLS, the top of the segment list is the topmost (outer) label of the label stack. In SRv6, the top of the segment list is represented by the first segment in the Segment Routing Header as defined in [IPv6-SRH].

NEXT: when the active segment is completed, NEXT is the operation consisting of the inspection of the next segment. The next segment becomes active. In SR-MPLS, NEXT is implemented as a POP of the top label. In SRv6, NEXT is implemented as the copy of the next segment from the SRH to the destination address of the IPv6 header.

CONTINUE: the active segment is not completed; hence, it remains active. In SR-MPLS, the CONTINUE operation is implemented as a SWAP of the top label [RFC3031]. In SRv6, this is the plain IPv6 forwarding action of a regular IPv6 packet according to its destination address.

SR Global Block (SRGB): the set of global segments in the SR domain. If a node participates in multiple SR domains, there is one SRGB for each SR domain. In SR-MPLS, SRGB is a local property of a node and identifies the set of local labels reserved for global segments. In SR-MPLS, using identical SRGBs on all nodes within the SR domain is strongly recommended. Doing so eases operations and troubleshooting as the same label represents the same global segment at each node. In SRv6, the SRGB is the set of global SRv6 SIDs in the SR domain.

SR Local Block (SRLB): local property of an SR node. If a node participates in multiple SR domains, there is one SRLB for each SR domain. In SR-MPLS, SRLB is a set of local labels reserved for local segments. In SRv6, SRLB is a set of local IPv6 addresses reserved for local SRv6 SIDs. In a controller-driven network, some controllers or applications may use the control plane to discover the available set of local segments.

Global Segment: a segment that is part of the SRGB of the domain. The instruction associated with the segment is defined at the SR domain level. A topological shortest-path segment to a given destination within an SR domain is a typical example of a global segment.

Local Segment: In SR-MPLS, this is a local label outside the SRGB. It may be part of the explicitly advertised SRLB. In SRv6, this can be any IPv6 address, i.e., the address may be part of the SRGB, but used such that it has local significance. The instruction associated with the segment is defined at the node level.

IGP Segment: the generic name for a segment attached to a piece of information advertised by a link-state IGP, e.g., an IGP prefix or an IGP adjacency.

IGP-Prefix Segment: an IGP-Prefix segment is an IGP segment representing an IGP prefix. When an IGP-Prefix segment is global within the SR IGP instance/topology, it identifies an instruction to

forward the packet along the path computed using the routing algorithm specified in the algorithm field, in the topology, and in the IGP instance where it is advertised. Also referred to as "prefix segment".

Prefix-SID: the SID of the IGP-Prefix segment.

IGP-Anycast Segment: an IGP-Anycast segment is an IGP-Prefix segment that identifies an anycast prefix advertised by a set of routers.

Anycast-SID: the SID of the IGP-Anycast segment.

IGP-Adjacency Segment: an IGP-Adjacency segment is an IGP segment attached to a unidirectional adjacency or a set of unidirectional adjacencies. By default, an IGP-Adjacency segment is local (unless explicitly advertised otherwise) to the node that advertises it. Also referred to as "Adj-SID".

Adj-SID: the SID of the IGP-Adjacency segment.

IGP-Node Segment: an IGP-Node segment is an IGP-Prefix segment that identifies a specific router (e.g., a loopback). Also referred to as "Node Segment".

Node-SID: the SID of the IGP-Node segment.

SR Policy: an ordered list of segments. The headend of an SR Policy steers packets onto the SR Policy. The list of segments can be specified explicitly in SR-MPLS as a stack of labels and in SRv6 as an ordered list of SRv6 SIDs. Alternatively, the list of segments is computed based on a destination and a set of optimization objective and constraints (e.g., latency, affinity, SRLG, etc.). The computation can be local or delegated to a PCE server. An SR Policy can be configured by the operator, provisioned via NETCONF [RFC6241] or provisioned via PCEP [RFC5440]. An SR Policy can be used for Traffic Engineering (TE), Operations, Administration, and Maintenance (OAM), or Fast Reroute (FRR) reasons.

Segment List Depth: the number of segments of an SR Policy. The entity instantiating an SR Policy at a node N should be able to discover the depth-insertion capability of the node N. For example, the PCEP SR capability advertisement described in [PCEP-SR-EXT] is one means of discovering this capability.

Forwarding Information Base (FIB): the forwarding table of a node

3. Link-State IGP Segments

Within an SR domain, an SR-capable IGP node advertises segments for its attached prefixes and adjacencies. These segments are called "IGP segments" or "IGP SIDs". They play a key role in Segment Routing and use cases as they enable the expression of any path throughout the SR domain. Such a path is either expressed as a single IGP segment or a list of multiple IGP segments.

Advertisement of IGP segments requires extensions in link-state IGP protocols. These extensions are defined in [ISIS-SR-EXT], [OSPF-SR-EXT], and [OSPFv3-SR-EXT].

3.1. IGP-Prefix Segment (Prefix-SID)

An IGP-Prefix segment is an IGP segment attached to an IGP prefix. An IGP-Prefix segment is global (unless explicitly advertised otherwise) within the SR domain. The context for an IGP-Prefix segment includes the prefix, topology, and algorithm. Multiple SIDs MAY be allocated to the same prefix so long as the tuple <prefix, topology, algorithm> is unique.

Multiple instances and topologies are defined in IS-IS and OSPF in: [RFC5120], [RFC8202], [RFC6549], and [RFC4915].

3.1.1. Prefix-SID Algorithm

Segment Routing supports the use of multiple routing algorithms, i.e, different constraint-based shortest-path calculations can be supported. An algorithm identifier is included as part of a Prefix-SID advertisement. Specification of how an algorithm-specific path calculation is done is required in the document defining the algorithm.

This document defines two algorithms:

- Shortest Path First: this algorithm is the default behavior. The packet is forwarded along the well known ECMP-aware Shortest Path First (SPF) algorithm employed by the IGPs. However, it is explicitly allowed for a midpoint to implement another forwarding based on local policy. The Shortest Path First algorithm is, in fact, the default and current behavior of most of the networks where local policies may override the SPF decision.

- Strict Shortest Path First (Strict-SPF): This algorithm mandates that the packet be forwarded according to the ECMP-aware SPF algorithm and instructs any router in the path to ignore any possible local policy overriding the SPF decision. The SID

advertised with the Strict-SPF algorithm ensures that the path the packet is going to take is the expected, and not altered, SPF path. Note that Fast Reroute (FRR) [RFC5714] mechanisms are still compliant with the Strict Shortest Path First algorithm. In other words, a packet received with a Strict-SPF SID may be rerouted through an FRR mechanism. Strict-SPF uses the same topology used by the Shortest Path First algorithm. Obviously, nodes that do not support Strict-SPF will not install forwarding entries for this algorithm. Restricting the topology only to those nodes that support this algorithm will not produce the desired forwarding paths since the desired behavior is to follow the path calculated by the Shortest Path First algorithm. Therefore, a source SR node MUST NOT use an SR Policy containing a strict SPF segment if the path crosses a node not supporting the Strict-SPF algorithm.

An IGP-Prefix segment identifies the path, to the related prefix, computed as per the associated algorithm. A packet injected anywhere within the SR domain with an active Prefix-SID is expected to be forwarded along a path computed using the specified algorithm. For this to be possible, a fully connected topology of routers supporting the specified algorithm is required.

3.1.2. SR-MPLS

When SR is used over the MPLS data plane, SIDs are an MPLS label or an index into an MPLS label space (either SRGB or SRLB).

Where possible, it is recommended that identical SRGBs be configured on all nodes in an SR domain. This simplifies troubleshooting as the same label will be associated with the same prefix on all nodes. In addition, it simplifies support for anycast as detailed in Section 3.3.

The following behaviors are associated with SR operating over the MPLS data plane:

- The IGP signaling extension for IGP-Prefix segment includes a flag to indicate whether directly connected neighbors of the node on which the prefix is attached should perform the NEXT operation or the CONTINUE operation when processing the SID. This behavior is equivalent to Penultimate Hop Popping (NEXT) or Ultimate Hop Popping (CONTINUE) in MPLS.

- A Prefix-SID is allocated in the form of an MPLS label (or an index in the SRGB) according to a process similar to IP address allocation. Typically, the Prefix-SID is allocated by policy by the operator (or Network Management System (NMS)), and the SID very rarely changes.

- While SR allows a local segment to be attached to an IGP prefix, where the terminology "IGP-Prefix segment" or "Prefix-SID" is used, the segment is assumed to be global (i.e., the SID is defined from the advertised SRGB). This is consistent with all the described use cases that require global segments attached to IGP prefixes.

- The allocation process MUST NOT allocate the same Prefix-SID to different prefixes.

- If a node learns of a Prefix-SID that has a value that falls outside the locally configured SRGB range, then the node MUST NOT use the Prefix-SID and SHOULD issue an error log reporting a misconfiguration.

- If a node N advertises Prefix-SID SID-R for a prefix R that is attached to N and specifies CONTINUE as the operation to be performed by directly connected neighbors, then N MUST maintain the following FIB entry:

  Incoming Active Segment: SID-R
  Ingress Operation: NEXT
  Egress interface: NULL

- A remote node M MUST maintain the following FIB entry for any learned Prefix-SID SID-R attached to prefix R:

  Incoming Active Segment: SID-R
  Ingress Operation:
    If the next-hop of R is the originator of R
    and M has been instructed to remove the active segment: NEXT
    Else: CONTINUE
  Egress interface: the interface(s) towards the next-hop along the
                    path computed using the algorithm advertised with
                    the SID toward prefix R.

As Prefix-SIDs are specific to a given algorithm, if traffic associated with an algorithm arrives at a node that does not support that algorithm, the traffic will be dropped as there will be no forwarding entry matching the incoming label.

3.1.3. SRv6

When SR is used over the IPv6 data plane:

- A Prefix-SID is an IPv6 address.
- An operator MUST explicitly instantiate an SRv6 SID. IPv6 node addresses are not SRv6 SIDs by default.

A node N advertising an IPv6 address R usable as a segment identifier MUST maintain the following FIB entry:

```
Incoming Active Segment: R
Ingress Operation: NEXT
Egress interface: NULL
```

Note that forwarding to R does not require an entry in the FIBs of all other routers for R. Forwarding can be, and most often will be, achieved by a shorter mask prefix that covers R.

Independent of SR support, any remote IPv6 node will maintain a plain IPv6 FIB entry for any prefix, no matter if the prefix represents a segment or not. This allows forwarding of packets to the node that owns the SID even by nodes that do not support SR.

Support of multiple algorithms applies to SRv6. Since algorithm-specific SIDs are simply IPv6 addresses, algorithm-specific forwarding entries can be achieved by assigning algorithm-specific subnets to the (set of) algorithm specific SIDs that a node allocates.

Nodes that do not support a given algorithm may still have a FIB entry covering an algorithm-specific address even though an algorithm-specific path has not been calculated by that node. This is mitigated by the fact that nodes that do not support a given algorithm will not be included in the topology associated with that algorithm-specific SPF; therefore, traffic using the algorithm-specific destination will normally not flow via the excluded node. If such traffic were to arrive and be forwarded by such a node, it will still progress towards the destination node. The next-hop will be either a node that supports the algorithm -- in which case, the packet will be forwarded along algorithm-specific paths (or be dropped if none are available) -- or a node that does NOT support the algorithm -- in which case, the packet will continue to be forwarded along Algorithm 0 paths towards the destination node.

3.2. IGP-Node Segment (Node-SID)

An IGP Node-SID MUST NOT be associated with a prefix that is owned by more than one router within the same routing domain.

3.3. IGP-Anycast Segment (Anycast-SID)

An Anycast segment or Anycast-SID enforces the ECMP-aware shortest-path forwarding towards the closest node of the anycast set. This is useful to express macro-engineering policies or protection mechanisms.

An IGP-Anycast segment MUST NOT reference a particular node.

Within an anycast group, all routers in an SR domain MUST advertise the same prefix with the same SID value.

3.3.1. Anycast-SID in SR-MPLS

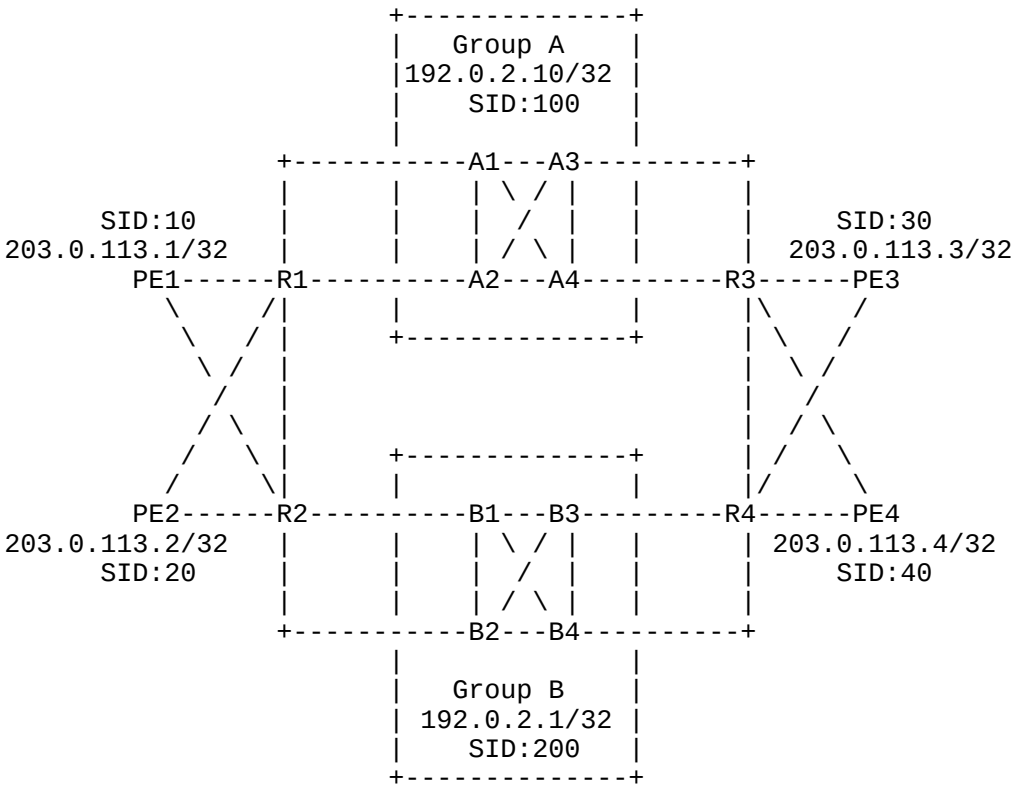

Figure 1: Transit Device Groups

The Figure 1 illustrates a network example with two groups of transit devices. Group A consists of devices {A1, A2, A3, and A4}. They are all provisioned with the anycast address 192.0.2.10/32 and the Anycast-SID 100.

Similarly, Group B consists of devices {B1, B2, B3, and B4}, and they are all provisioned with the anycast address 192.0.2.1/32 and the Anycast-SID 200. In the above network topology, each Provide Edge (PE) device has a path to each of the groups: A and B.

PE1 can choose a particular transit device group when sending traffic to PE3 or PE4. This will be done by pushing the Anycast-SID of the group in the stack.

Processing the anycast, and subsequent segments, requires special care.

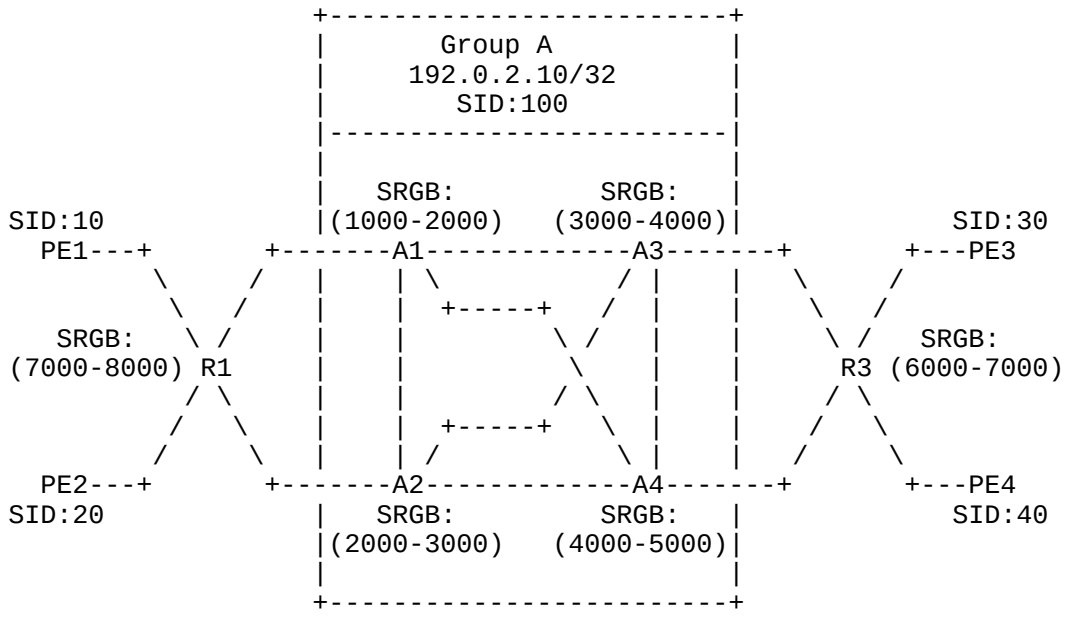

Figure 2: Transit Paths via Anycast Group A

Considering an MPLS deployment, in the above topology, if device PE1 (or PE2) requires the sending of a packet to the device PE3 (or PE4), it needs to encapsulate the packet in an MPLS payload with the following stack of labels.

- Label allocated by R1 for Anycast-SID 100 (outer label).
- Label allocated by the nearest router in Group A for SID 30 (for destination PE3).

In this case, the first label is easy to compute. However, because there is more than one device that is topologically nearest (A1 and A2), determining the second label is impossible unless A1 and A2 allocated the same label value to the same prefix. Devices A1 and A2 may be devices from different hardware vendors. If both don't allocate the same label value for SID 30, it is impossible to use the anycast Group A as a transit anycast group towards PE3. Hence, PE1 (or PE2) cannot compute an appropriate label stack to steer the packet exclusively through the Group A devices. Same holds true for devices PE3 and PE4 when trying to send a packet to PE1 or PE2.

To ease the use of an anycast segment, it is recommended to configure identical SRGBs on all nodes of a particular anycast group. Using this method, as mentioned above, computation of the label following the anycast segment is straightforward.

Using an anycast segment without configuring identical SRGBs on all nodes belonging to the same anycast group may lead to misrouting (in an MPLS VPN deployment, some traffic may leak between VPNs).

3.4. IGP-Adjacency Segment (Adj-SID)

The adjacency is formed by the local node (i.e., the node advertising the adjacency in the IGP) and the remote node (i.e., the other end of the adjacency). The local node MUST be an IGP node. The remote node may be an adjacent IGP neighbor or a non-adjacent neighbor (e.g., a forwarding adjacency, [RFC4206]).

A packet injected anywhere within the SR domain with a segment list {SN, SNL} where SN is the Node-SID of node N and SNL is an Adj-SID attached by node N to its adjacency over link L will be forwarded along the shortest path to N and then be switched by N, without any IP shortest-path consideration, towards link L. If the Adj-SID identifies a set of adjacencies, then the node N load-balances the traffic among the various members of the set.

Similarly, when using a global Adj-SID, a packet injected anywhere within the SR domain with a segment list {SNL}, where SNL is a global Adj-SID attached by node N to its adjacency over link L, will be forwarded along the shortest path to N and then be switched by N, without any IP shortest-path consideration, towards link L. If the Adj-SID identifies a set of adjacencies, then the node N does load-balance the traffic among the various members of the set. The use of global Adj-SID allows to reduce the size of the segment list when expressing a path at the cost of additional state (i.e., the global Adj-SID will be inserted by all routers within the area in their forwarding table).

An "IGP-Adjacency segment" or "Adj-SID" enforces the switching of the packet from a node towards a defined interface or set of interfaces. This is key to theoretically prove that any path can be expressed as a list of segments.

The encodings of the Adj-SID include a set of flags supporting the following functionalities:

- Eligible for Protection (e.g., using IPFRR or MPLS-FRR). Protection allows that in the event the interface(s) associated with the Adj-SID are down, that the packet can still be forwarded via an alternate path. The use of protection is clearly a policy-based decision; that is, for a given policy protection may or may not be desirable.

- Indication whether the Adj-SID has local or global scope. Default scope SHOULD be local.

- Indication whether the Adj-SID is persistent across control plane restarts. Persistence is a key attribute in ensuring that an SR Policy does not temporarily result in misforwarding due to reassignment of an Adj-SID.

A weight (as described below) is also associated with the Adj-SID advertisement.

A node SHOULD allocate one Adj-SID for each of its adjacencies.

A node MAY allocate multiple Adj-SIDs for the same adjacency. An example is to support an Adj-SID that is eligible for protection and an Adj-SID that is NOT eligible for protection.

A node MAY associate the same Adj-SID to multiple adjacencies.

In order to be able to advertise in the IGP all the Adj-SIDs representing the IGP adjacencies between two nodes, parallel adjacency suppression MUST NOT be performed by the IGP.

When a node binds an Adj-SID V to a local data-link L, the node MUST install the following FIB entry:

```
Incoming Active Segment: V
Ingress Operation: NEXT
Egress Interface: L
```

The Adj-SID implies, from the router advertising it, the forwarding of the packet through the adjacency or adjacencies identified by the Adj-SID, regardless of its IGP/SPF cost. In other words, the use of adjacency segments overrides the routing decision made by the SPF algorithm.

3.4.1. Parallel Adjacencies

Adj-SIDs can be used in order to represent a set of parallel interfaces between two adjacent routers.

A node MUST install a FIB entry for any locally originated Adj-SID of value W attached to a set of links B with:

```
Incoming Active Segment: W
Ingress Operation: NEXT
Egress interfaces: load-balance between any data-link within set B
```

When parallel adjacencies are used and associated with the same Adj-SID, and, in order to optimize the load-balancing function, a "weight" factor can be associated with the Adj-SID advertised with each adjacency. The weight tells the ingress (or an SDN/ orchestration system) about the load-balancing factor over the parallel adjacencies. As shown in Figure 3, A and B are connected through two parallel adjacencies

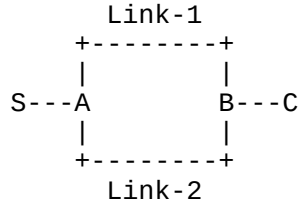

Figure 3: Parallel Links and Adj-SIDs

Node A advertises following Adj-SIDs and weights:

- Link-1: Adj-SID 1000, weight: 1
- Link-2: Adj-SID 1000, weight: 2

Node S receives the advertisements of the parallel adjacencies and understands that by using Adj-SID 1000 node A will load-balance the traffic across the parallel links (Link-1 and Link-2) according to a 1:2 ratio i.e., twice as many packets will flow over Link-2 as compared to Link-1.

3.4.2. LAN Adjacency Segments

In LAN subnetworks, link-state protocols define the concept of Designated Router (DR, in OSPF) or Designated Intermediate System (DIS, in IS-IS) that conduct flooding in broadcast subnetworks and that describe the LAN topology in a special routing update (OSPF Type2 LSA or IS-IS Pseudonode LSP).

The difficulty with LANs is that each router only advertises its connectivity to the DR/DIS and not to each of the individual nodes in the LAN. Therefore, additional protocol mechanisms (IS-IS and OSPF) are necessary in order for each router in the LAN to advertise an Adj-SID associated with each neighbor in the LAN.

3.5. Inter-Area Considerations

In the following example diagram, it is assumed that the all areas are part of a single SR domain.

The Figure 4 assumes the IPv6 control plane with the MPLS data plane.

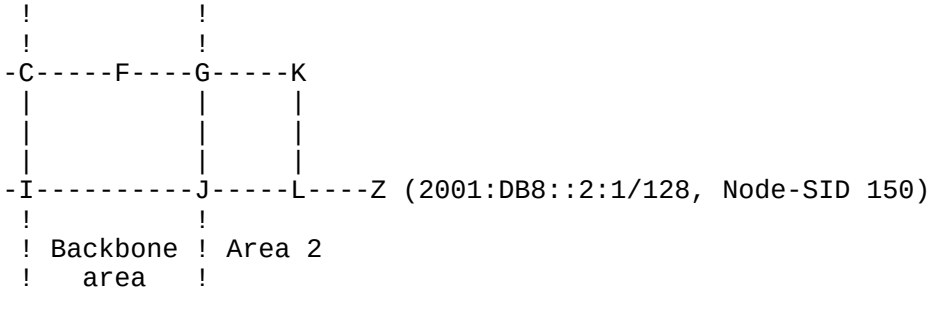

Figure 4: Inter-Area Topology Example

In Area 2, node Z allocates Node-SID 150 to his local IPv6 prefix 2001:DB8::2:1/128.

Area Border Routers (ABRs) G and J will propagate the prefix and its SIDs into the backbone area by creating a new instance of the prefix according to normal inter-area/level IGP propagation rules.

Nodes C and I will apply the same behavior when leaking prefixes from the backbone area down to area 1. Therefore, node S will see prefix 2001:DB8::2:1/128 with Prefix-SID 150 and advertised by nodes C and I.

Therefore, the result is that a Prefix-SID remains attached to its related IGP prefix through the inter-area process, which is the expected behavior in a single SR domain.

When node S sends traffic to 2001:DB8::2:1/128, it pushes Node-SID(150) as an active segment and forwards it to A.

When a packet arrives at ABR I (or C), the ABR forwards the packet according to the active segment (Node-SID(150)). Forwarding continues across area borders, using the same Node-SID(150) until the packet reaches its destination.

4. BGP Segments

BGP segments may be allocated and distributed by BGP.

4.1. BGP-Prefix Segment

A BGP-Prefix segment is a BGP segment attached to a BGP prefix.

A BGP-Prefix segment is global (unless explicitly advertised otherwise) within the SR domain.

The BGP-Prefix segment is the BGP equivalent to the IGP-Prefix segment.

A likely use case for the BGP-Prefix segment is an IGP-free hyper-scale spine-leaf topology where connectivity is learned solely via BGP [RFC7938]

4.2. BGP Peering Segments

In the context of BGP Egress Peer Engineering (EPE), as described in [SR-CENTRAL-EPE], an EPE-enabled egress node MAY advertise segments corresponding to its attached peers. These segments are called BGP peering segments or BGP peering SIDs. They enable the expression of source-routed inter-domain paths.

An ingress border router of an Autonomous System (AS) may compose a list of segments to steer a flow along a selected path within the AS towards a selected egress border router C of the AS and through a specific peer. At a minimum, a BGP peering engineering policy applied at an ingress node involves two segments: the Node-SID of the chosen egress node and the BGP peering segment for the chosen egress node peer or peering interface.

Three types of BGP peering segments/SIDs are defined: PeerNode SID, PeerAdj SID, and PeerSet SID.

- PeerNode SID: a BGP PeerNode segment/SID is a local segment. At the BGP node advertising it, its semantics are:
  - SR operation: NEXT.
  - Next-Hop: the connected peering node to which the segment is related.
- PeerAdj SID: a BGP PeerAdj segment/SID is a local segment. At the BGP node advertising it, the semantics are:
  - SR operation: NEXT.
  - Next-Hop: the peer connected through the interface to which the segment is related.
- PeerSet SID: a BGP PeerSet segment/SID is a local segment. At the BGP node advertising it, the semantics are:
  - SR operation: NEXT.
  - Next-Hop: load-balance across any connected interface to any peer in the related group.

  A peer set could be all the connected peers from the same AS or a subset of these. A group could also span across AS. The group definition is a policy set by the operator.

The BGP extensions necessary in order to signal these BGP peering segments are defined in [BGPLS-SR-EPE].

5. Binding Segment

In order to provide greater scalability, network opacity, and service independence, SR utilizes a Binding SID (BSID). The BSID is bound to an SR Policy, instantiation of which may involve a list of SIDs. Any packets received with an active segment equal to BSID are steered onto the bound SR Policy.

A BSID may be either a local or a global SID. If local, a BSID SHOULD be allocated from the SRLB. If global, a BSID MUST be allocated from the SRGB.

Use of a BSID allows the instantiation of the policy (the SID list) to be stored only on the node or nodes that need to impose the policy. Direction of traffic to a node supporting the policy then only requires imposition of the BSID. If the policy changes, this also means that only the nodes imposing the policy need to be updated. Users of the policy are not impacted.

5.1. IGP Mirroring Context Segment

One use case for a Binding segment is to provide support for an IGP node to advertise its ability to process traffic originally destined to another IGP node, called the "mirrored node" and identified by an IP address or a Node-SID, provided that a Mirroring Context segment is inserted in the segment list prior to any service segment local to the mirrored node.

When a given node B wants to provide egress node A protection, it advertises a segment identifying node’s A context. Such a segment is called "Mirroring Context segment" and is identified by the Mirror SID.

The Mirror SID is advertised using the Binding segment defined in SR IGP protocol extensions [ISIS-SR-EXT].

In the event of a failure, a Point of Local Repair (PLR) diverting traffic from A to B does a PUSH of the Mirror SID on the protected traffic. When receiving the traffic with the Mirror SID as the active segment, B uses that segment and processes underlying segments in the context of A.

6. Multicast

Segment Routing is defined for unicast. The application of the source-route concept to Multicast is not in the scope of this document.

7. IANA Considerations

This document has no IANA actions.

8. Security Considerations

Segment Routing is applicable to both MPLS and IPv6 data planes.

SR adds some metadata (instructions) to the packet, with the list of forwarding path elements (e.g., nodes, links, services, etc.) that the packet must traverse. It has to be noted that the complete source-routed path may be represented by a single segment. This is the case of the Binding SID.

By default, SR operates within a trusted domain. Traffic MUST be filtered at the domain boundaries.

The use of best practices to reduce the risk of tampering within the trusted domain is important. Such practices are discussed in [RFC4381] and are applicable to both SR-MPLS and SRv6.

8.1. SR-MPLS

When applied to the MPLS data plane, SR does not introduce any new behavior or any change in the way the MPLS data plane works. Therefore, from a security standpoint, this document does not define any additional mechanism in the MPLS data plane.

SR allows the expression of a source-routed path using a single segment (the Binding SID). Compared to RSVP-TE, which also provides explicit routing capability, there are no fundamental differences in terms of information provided. Both RSVP-TE and Segment Routing may express a source-routed path using a single segment.

When a path is expressed using a single label, the syntax of the metadata is equivalent between RSVP-TE [RFC3209] and SR.

When a source-routed path is expressed with a list of segments, additional metadata is added to the packet consisting of the source-routed path the packet must follow expressed as a segment list.

When a path is expressed using a label stack, if one has access to the meaning (i.e., the Forwarding Equivalence Class) of the labels, one has the knowledge of the explicit path. For the MPLS data plane, as no data-plane modification is required, there is no fundamental change of capability. Yet, the occurrence of label stacking will increase.

SR domain boundary routers MUST filter any external traffic destined to a label associated with a segment within the trusted domain. This includes labels within the SRGB of the trusted domain, labels within the SRLB of the specific boundary router, and labels outside either of these blocks. External traffic is any traffic received from an interface connected to a node outside the domain of trust.

From a network protection standpoint, there is an assumed trust model such that any node imposing a label stack on a packet is assumed to be allowed to do so. This is a significant change compared to plain IP offering shortest path routing, but it is not fundamentally different compared to existing techniques providing explicit routing capability such as RSVP-TE. By default, the explicit routing information MUST NOT be leaked through the boundaries of the administered domain. Segment Routing extensions that have been defined in various protocols, leverage the security mechanisms of these protocols such as encryption, authentication, filtering, etc.

In the general case, a segment-routing-capable router accepts and installs labels only if the labels have been previously advertised by a trusted source. The received information is validated using existing control-plane protocols providing authentication and security mechanisms. Segment Routing does not define any additional security mechanism in existing control-plane protocols.

SR does not introduce signaling between the source and the midpoints of a source-routed path. With SR, the source-routed path is computed using SIDs previously advertised in the IP control plane. Therefore, in addition to filtering and controlled advertisement of SIDs at the boundaries of the SR domain, filtering in the data plane is also required. Filtering MUST be performed on the forwarding plane at the boundaries of the SR domain and may require looking at multiple labels/instructions.

For the MPLS data plane, there are no new requirements as the existing MPLS architecture already allows such source routing by stacking multiple labels. And, for security protection, [RFC4381] and [RFC5920] already call for the filtering of MPLS packets on trust boundaries.

8.2. SRv6

When applied to the IPv6 data plane, Segment Routing does introduce the Segment Routing Header (SRH, [IPv6-SRH]) which is a type of Routing Extension header as defined in [RFC8200].

The SRH adds some metadata to the IPv6 packet, with the list of forwarding path elements (e.g., nodes, links, services, etc.) that the packet must traverse and that are represented by IPv6 addresses. A complete source-routed path may be encoded in the packet using a single segment (single IPv6 address).

SR domain boundary routers MUST filter any external traffic destined to an address within the SRGB of the trusted domain or the SRLB of the specific boundary router. External traffic is any traffic received from an interface connected to a node outside the domain of trust.

From a network-protection standpoint, there is an assumed trust model such that any node adding an SRH to the packet is assumed to be allowed to do so. Therefore, by default, the explicit routing information MUST NOT be leaked through the boundaries of the administered domain. Segment Routing extensions that have been defined in various protocols, leverage the security mechanisms of these protocols such as encryption, authentication, filtering, etc.

In the general case, an SRv6 router accepts and install segments identifiers (in the form of IPv6 addresses), only if these SIDs are advertised by a trusted source. The received information is validated using existing control-plane protocols providing authentication and security mechanisms. Segment Routing does not define any additional security mechanism in existing control-plane protocols.

Problems that may arise when the above behaviors are not implemented or when the assumed trust model is violated (e.g., through a security breach) include:

- Malicious looping
- Evasion of access controls
- Hiding the source of DoS attacks

Security concerns with SR at the IPv6 data plane are more completely discussed in [RFC5095]. The new IPv6-based Segment Routing Header is defined in [IPv6-SRH]. This document also discusses the above security concerns.

8.3. Congestion Control

SR does not introduce new requirements for congestion control. By default, traffic delivery is assumed to be best effort. Congestion control may be implemented at endpoints. Where SR policies are in use, bandwidth allocation may be managed by monitoring incoming traffic associated with the binding SID identifying the SR Policy. Other solutions such as presented in [RFC8084] may be applicable.

9. Manageability Considerations

In SR-enabled networks, the path the packet takes is encoded in the header. As the path is not signaled through a protocol, OAM mechanisms are necessary in order for the network operator to validate the effectiveness of a path as well as to check and monitor its liveness and performance. However, it has to be noted that SR allows to reduce substantially the number of states in transit nodes; hence, the number of elements that a transit node has to manage is smaller.

SR OAM use cases for the MPLS data plane are defined in [RFC8403]. SR OAM procedures for the MPLS data plane are defined in [RFC8287].

SR routers receive advertisements of SIDs (index, label, or IPv6 address) from the different routing protocols being extended for SR. Each of these protocols have monitoring and troubleshooting mechanisms to provide operation and management functions for IP addresses that must be extended in order to include troubleshooting and monitoring functions of the SID.

SR architecture introduces the usage of global segments. Each global segment MUST be bound to a unique index or address within an SR domain. The management of the allocation of such an index or address by the operator is critical for the network behavior to avoid situations like misrouting. In addition to the allocation policy/ tooling that the operator will have in place, an implementation SHOULD protect the network in case of conflict detection by providing a deterministic resolution approach.

When a path is expressed using a label stack, the occurrence of label stacking will increase. A node may want to signal, in the control plane, its ability in terms of size of the label stack it can support.

A YANG data model [RFC6020] for SR configuration and operations has been defined in [SR-YANG].

When SR is applied to the IPv6 data plane, segments are identified through IPv6 addresses. The allocation, management, and troubleshooting of segment identifiers is no different than the existing mechanisms applied to the allocation and management of IPv6 addresses.

The DA of the packet gives the active segment address. The segment list in the SRH gives the entire path of the packet. The validation of the source-routed path is done through inspection of DA and SRH present in the packet header matched to the equivalent routing table entries.

In the context of the SRv6 data plane, the source-routed path is encoded in the SRH as described in [IPv6-SRH]. The SRv6 source-routed path is instantiated into the SRH as a list of IPv6 addresses where the active segment is in the DA field of the IPv6 packet header. Typically, by inspecting, in any node, the packet header, it is possible to derive the source-routed path to which it belongs. Similar to the context of the SR-MPLS data plane, an implementation may originate path control and monitoring packets where the source-routed path is inserted in the SRH and where each segment of the path inserts in the packet the relevant data in order to measure the end-to-end path and performance.

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

10.2. Informative References

[BGPLS-SR-EPE]
Previdi, S., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", Work in Progress, draft-ietf-idr-bgpls-segment-routing-epe-15, March 2018.

[IPv6-SRH]
Filsfils, C., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, Ed., "IPv6 Segment Routing Header (SRH)", Work in Progress, draft-ietf-6man-segment-routing-header-14, June 2018.

[ISIS-SR-EXT]
Previdi, S., Ed., Ginsberg, L., Ed., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and J. Tantsura, "IS-IS Extensions for Segment Routing", Work in Progress, draft-ietf-isis-segment-routing-extensions-19, July 2018.

[OSPF-SR-EXT]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", Work in Progress, draft-ietf-ospf-segment-routing-extensions-25, April 2018.

[OSPFv3-SR-EXT]
Psenak, P., Ed., Filsfils, C., Previdi, S., Ed., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPFv3 Extensions for Segment Routing", Work in Progress, draft-ietf-ospf-ospfv3-segment-routing-extensions-13, May 2018.

[PCEP-SR-EXT]
Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "PCEP Extensions for Segment Routing", Work in Progress, draft-ietf-pce-segment-routing-12, June 2018.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, DOI 10.17487/RFC4206, October 2005, <https://www.rfc-editor.org/info/rfc4206>.

[RFC4381] Behringer, M., "Analysis of the Security of BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4381, DOI 10.17487/RFC4381, February 2006, <https://www.rfc-editor.org/info/rfc4381>.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, DOI 10.17487/RFC4915, June 2007, <https://www.rfc-editor.org/info/rfc4915>.

[RFC5095] Abley, J., Savola, P., and G. Neville-Neil, "Deprecation of Type 0 Routing Headers in IPv6", RFC 5095, DOI 10.17487/RFC5095, December 2007, <https://www.rfc-editor.org/info/rfc5095>.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <https://www.rfc-editor.org/info/rfc5120>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, DOI 10.17487/RFC5714, January 2010, <https://www.rfc-editor.org/info/rfc5714>.

[RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, DOI 10.17487/RFC5920, July 2010, <https://www.rfc-editor.org/info/rfc5920>.

[RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, DOI 10.17487/RFC6020, October 2010, <https://www.rfc-editor.org/info/rfc6020>.

[RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011, <https://www.rfc-editor.org/info/rfc6241>.

[RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, DOI 10.17487/RFC6549, March 2012, <https://www.rfc-editor.org/info/rfc6549>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[RFC8084] Fairhurst, G., "Network Transport Circuit Breakers", BCP 208, RFC 8084, DOI 10.17487/RFC8084, March 2017, <https://www.rfc-editor.org/info/rfc8084>.

[RFC8202] Ginsberg, L., Previdi, S., and W. Henderickx, "IS-IS Multi-Instance", RFC 8202, DOI 10.17487/RFC8202, June 2017, <https://www.rfc-editor.org/info/rfc8202>.

[RFC8287] Kumar, N., Ed., Pignataro, C., Ed., Swallow, G., Akiya, N., Kini, S., and M. Chen, "Label Switched Path (LSP) Ping/Traceroute for Segment Routing (SR) IGP-Prefix and IGP-Adjacency Segment Identifiers (SIDs) with MPLS Data Planes", RFC 8287, DOI 10.17487/RFC8287, December 2017, <https://www.rfc-editor.org/info/rfc8287>.

[RFC8355] Filsfils, C., Ed., Previdi, S., Ed., Decraene, B., and R. Shakir, "Resiliency Use Cases in Source Packet Routing in Networking (SPRING) Networks", RFC 8355, DOI 10.17487/RFC8355, March 2018, <https://www.rfc-editor.org/info/rfc8355>.

[RFC8403] Geib, R., Ed., Filsfils, C., Pignataro, C., Ed., and N. Kumar, "A Scalable and Topology-Aware MPLS Data-Plane Monitoring System", RFC 8403, DOI 10.17487/RFC8403, July 2018, <http://www.rfc-editor.org/info/rfc8403>.

[SR-CENTRAL-EPE] Filsfils, C., Previdi, S., Dawra, G., Aries, E., and D. Afanasiev, "Segment Routing Centralized BGP Egress Peer Engineering", Work in Progress, draft-ietf-spring-segment-routing-central-epe-10, December 2017.

[SR-MPLS] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with MPLS data plane", Work in Progress, draft-ietf-spring-segment-routing-mpls-14, June 2018.

[SR-YANG] Litkowski, S., Qu, Y., Sarkar, P., and J. Tantsura, "YANG Data Model for Segment Routing", Work in Progress, draft-ietf-spring-sr-yang-09, June 2018.

Acknowledgements

We would like to thank Dave Ward, Peter Psenak, Dan Frost, Stewart Bryant, Pierre Francois, Thomas Telkamp, Ruediger Geib, Hannes Gredler, Pushpasis Sarkar, Eric Rosen, Chris Bowers, and Alvaro Retana for their comments and review of this document.

Contributors

The following people have substantially contributed to the definition of the Segment Routing architecture and to the editing of this document:

Ahmed Bashandy
Cisco Systems, Inc.
Email: bashandy@cisco.com

Martin Horneffer
Deutsche Telekom
Email: Martin.Horneffer@telekom.de

Wim Henderickx
Nokia
Email: wim.henderickx@nokia.com

Jeff Tantsura
Email: jefftant@gmail.com

Edward Crabbe
Email: edward.crabbe@gmail.com

Igor Milojevic
Email: milojevicigor@gmail.com

Saku Ytti
TDC
Email: saku@ytti.fi

Authors' Addresses

Clarence Filsfils (editor)
Cisco Systems, Inc.
Brussels
Belgium

Email: cfilsfil@cisco.com

Stefano Previdi (editor)
Cisco Systems, Inc.
Italy

Email: stefano@previdi.net

Les Ginsberg
Cisco Systems, Inc.

Email: ginsberg@cisco.com

Bruno Decraene
Orange
FR

Email: bruno.decraene@orange.com

Stephane Litkowski
Orange
France

Email: stephane.litkowski@orange.com

Rob Shakir
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
United States of America

Email: robjs@google.com