Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 1, 2018

C. Filsfils, Ed.
S. Previdi, Ed.
Cisco Systems, Inc.
L. Ginsberg
Cisco Systems, Inc
B. Decraene
S. Litkowski
Orange
R. Shakir
Google, Inc.
October 28, 2017

Segment Routing Architecture
draft-ietf-spring-segment-routing-13

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 allows to enforce a flow through any topological path while maintaining per-flow state only at the ingress nodes to the SR domain.

Segment Routing can be directly applied to the MPLS architecture with no change on 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.

Segment Routing 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 of the packet. The next active segment is indicated by a pointer in the new routing header.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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 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 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 a source-routed policy (e.g., pre-computed local protection [I-D.ietf-spring-resiliency-use-cases]) . A node individually computes the source-routed 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 NETCONF, 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 SID's are instantiated at which nodes and which sets of local (SRLB) and global labels (SRGB) 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 a source-routed policy on behalf of an IGP node. The SR architecture does not restrict how the nodes which 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 responding to requests from hosts.

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

Segment Routing can be directly applied to the MPLS architecture with no change on the forwarding plane [I-D.ietf-spring-segment-routing-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.

Segment Routing can be applied to the IPv6 architecture with a new type of routing header called the SR header (SRH) [I-D.ietf-6man-segment-routing-header] . 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 SID's 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 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 dataplanes.

2. Terminology

SR-MPLS: the instantiation of SR on the MPLS dataplane

SRv6: the instantiation of SR on the IPv6 dataplane.

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 single SR domain. However, some deployments may wish

to sub-divide 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 MUST be used by the receiving router to process the packet. In the MPLS dataplane it is the top label. In the IPv6 dataplane it is the destination address. [I-D.ietf-6man-segment-routing-header].

PUSH: the instruction 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 [I-D.ietf-6man-segment-routing-header].

NEXT: when the active segment is completed, NEXT is the instruction 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 and hence remains active. In SR-MPLS, CONTINUE instruction 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 the same SRGB 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 SID's. 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 which is part of the SRGB of the domain. The instruction associated to 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 to 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 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 which identify 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 Adjacency Segment.

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

IGP-Node Segment: an IGP-Node Segment is an IGP-Prefix Segment which 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 SID's. Alternatively, the list of segments is computed based on a destination and a set of optimization objective and constraints (e.g., latency, affinity, SRLG, ...). The computation can be local or delegated to a PCE server. An SR policy can be configured by the operator, provisioned via NETCONF or

provisioned via PCEP [RFC5440] . An SR policy can be used for traffic-engineering, OAM or 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 [I-D.ietf-pce-segment-routing] 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
[I-D.ietf-isis-segment-routing-extensions]
[I-D.ietf-ospf-segment-routing-extensions]
[I-D.ietf-ospf-ospfv3-segment-routing-extensions]

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.

This document defines two algorithms:

- "Shortest Path": this algorithm is the default behavior. The packet is forwarded along the well known ECMP-aware 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" 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": This algorithm mandates that the packet is forwarded according to 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 "Strict Shortest Path" 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. In other words, a packet received with a Strict-SPF SID may be rerouted through a FRR mechanism.

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. Clearly, if not all SR capable nodes in an SR Domain support a given algorithm it is not possible to guarantee that the packet will follow a path consistent with the associated algorithm.

A router MUST drop any SR traffic associated with an SR algorithm, if the nexthop router has not advertised support for the SR algorithm.

The ingress node of an SR domain SHOULD validate that the path to a prefix, advertised with a given algorithm, includes nodes all supporting the advertised algorithm. If this constraint cannot be met the packet SHOULD be dropped by the ingress node. Note that in the special case of "Shortest Path", all nodes (SR Capable or not) are assumed to support this algorithm.

3.1.2. SR-MPLS

When SR is used over the MPLS dataplane SIDs are an MPLS label or an index into an MPLS label space (either SRGB or SRLB). An SRGB/SRLB is advertised as an ordered set of ranges which has the following properties:

- Each range specifies a starting label and the number of labels in that range

- The set of ranges advertised by a given node MUST be disjoint

When the SID is an index, the mapping of the index to a label is computed using the following algorithm:

```
r = # of advertised ranges
L(Ri) is the starting label of Range #i
N(Ri) is the number of labels in Range #i
X is the 0 based index

i = 1
while ((i <= r) && (X >= N(Ri)) {
  X = X - N(Ri)
  i = i + 1
}
if (i <= r)
 LABEL = L(Ri) + X
else
  no valid label exists for this index
```

Where possible, it is recommended that a consistent SRGB 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 dataplane:

- 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 NMS) and the SID very rarely changes.

- While SR allows to attach a local segment to an IGP prefix, it is specifically assumed that when the terms "IGP-Prefix Segment" and "Prefix-SID" are used, the segment is global (the SID is allocated from the SRGB or as an index into 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 IP prefixes.

- If a node learns a Prefix-SID having 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, if N specifies CONTINUE as the operation to be performed by directly connected neighbors, 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 IP prefix R:

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

3.1.3. SRv6

When SR is used over the IPv6 dataplane:

- 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

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

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

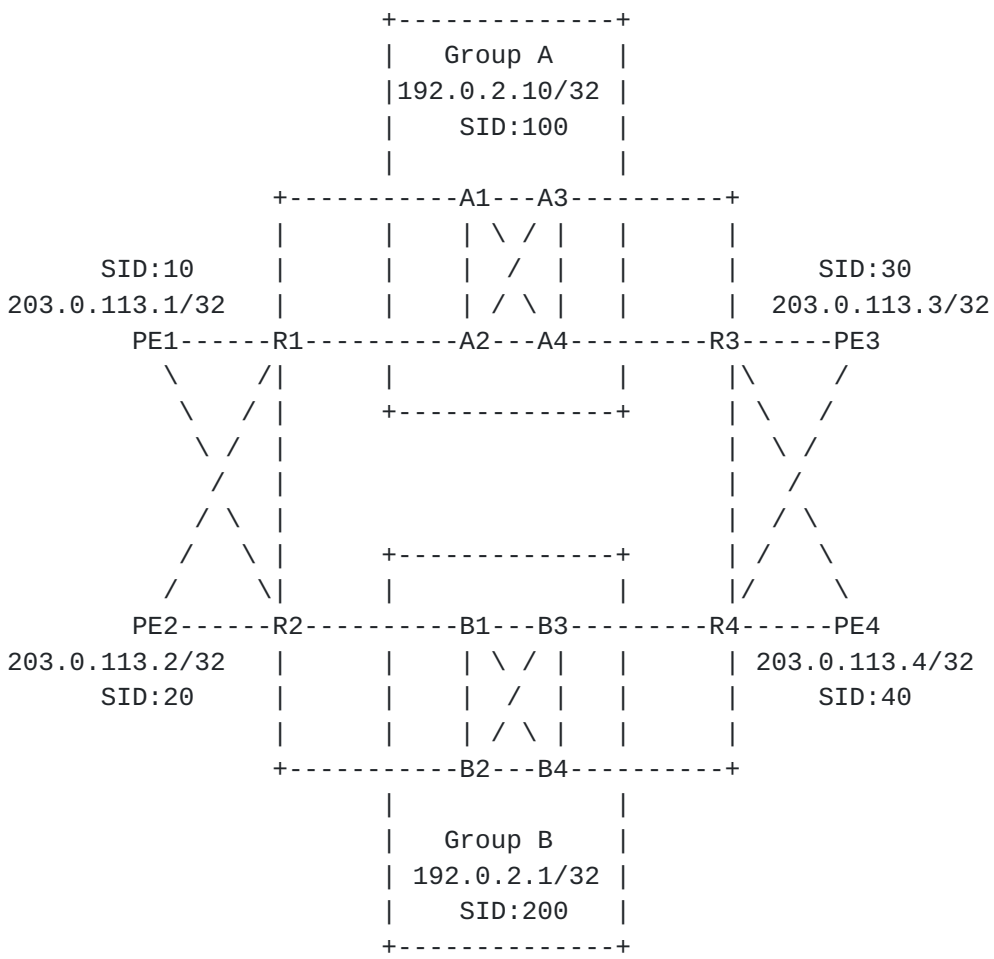

Figure 1: Transit device groups

The figure above describes 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 are all provisioned with the anycast address 192.0.2.1/32, anycast SID 200. In the above network topology, each 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.

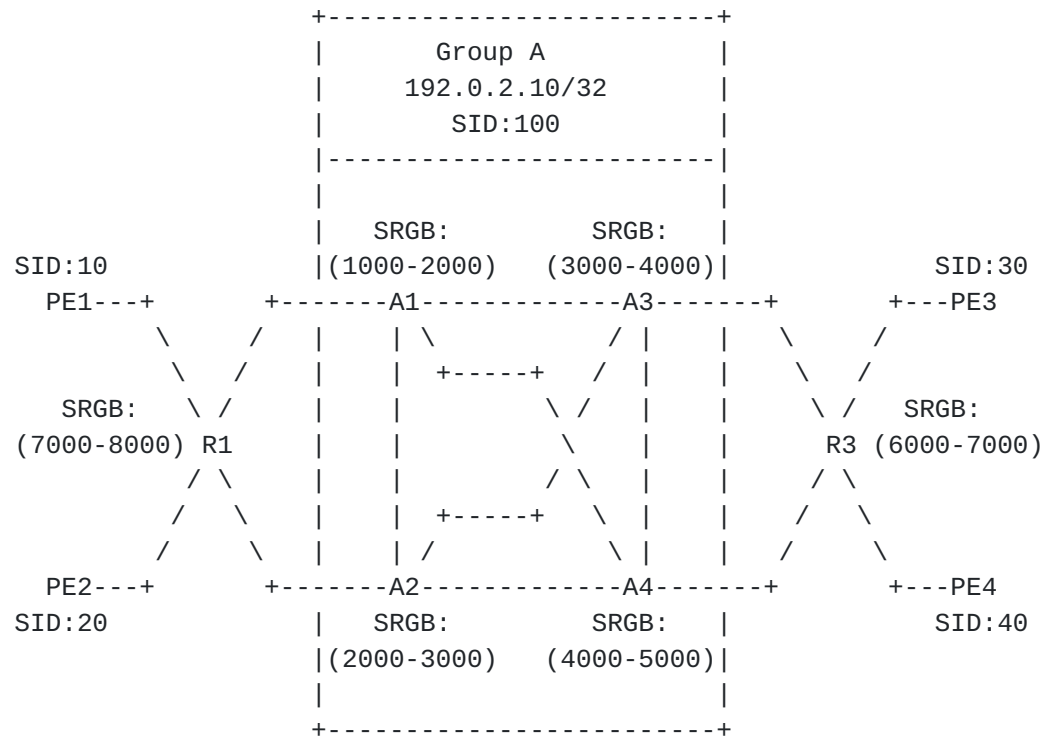

Figure 2: Transit paths via anycast group A

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

o 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).

While the first label is easy to compute, in this case since there are more than one topologically nearest devices (A1 and A2), unless A1 and A2 allocated the same label value to the same prefix, determining the second label is impossible. 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 anycast segment in a short term, it is recommended to configure the same SRGB 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 anycast segment without configuring the same SRGB on nodes belonging to the same device 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)

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

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 which is eligible for protection and an Adj-SID which 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 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(ies) 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 adjacency segment (Adj-SID) of value W attached to a set of links B with:

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

When parallel adjacencies are used and associated to the same Adj-SID, and in order to optimize the load balancing function, a "weight" factor can be associated to 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

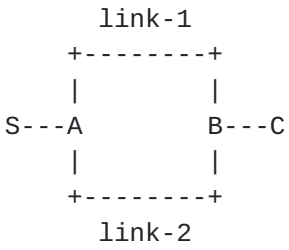

Figure 3: Parallel Links and Adj-SIDs

Node A advertises following Adj-SIDs and weights:

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

o 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 to each neighbor in the LAN. These extensions are defined in IGP SR extensions documents.

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 example here below assumes the IPv6 control plane with the MPLS dataplane.

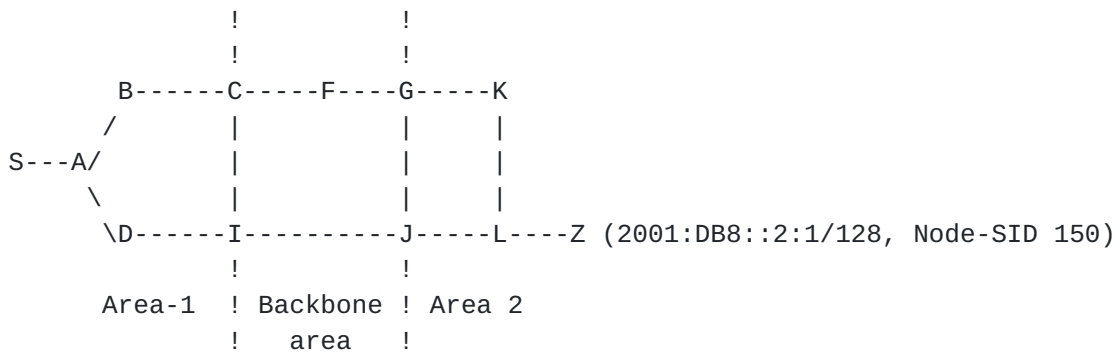

Figure 4: Inter-Area Topology Example

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

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.

It therefore results that a Prefix-SID remains attached to its related IGP Prefix through the inter-area process.

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

When 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 Peering 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 SID 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 [I-D.ietf-spring-segment-routing-central-epe], an EPE enabled Egress PE 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 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 minimum, a BGP peering Engineering policy applied at an ingress PE involves two segments: the Node SID of the chosen egress PE and then the BGP peering segment for the chosen egress PE 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 is:
  - SR header 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 semantic is:
  - SR header operation: NEXT.

* Next-Hop: the peer connected through the interface to which the segment is related.

o PeerSet SID. a BGP PeerSet segment/SID is a local segment. At the BGP node advertising it, the semantic is:

* SR header 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 [I-D.ietf-idr-bgpls-segment-routing-epe]

5. Binding Segment

An SR Policy is bound to a so-called Binding SID (BSID). Any packets received with active segment = BSID are steered onto the bound SR Policy.

A BSID may either be 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.

One of the possible use cases for a BSID is to overcome a Segment List Depth limitation on a given node by requiring that node only to impose a BSID which could be SWAPPED on downstream nodes with a set of SIDs associated with an SR policy.

5.1. IGP Mirroring Context Segment

Another 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 be 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 segment is called "Mirror Context Segment" and identified by the Mirror SID.

The Mirror SID is advertised using the binding segment defined in SR IGP protocol extensions [I-D.ietf-isis-segment-routing-extensions] .

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. B, when receiving the traffic with the Mirror SID as the active segment, 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 does not require any action from IANA.

8. Security Considerations

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

Segment Routing adds some meta-data (instructions) on 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.

8.1. SR-MPLS

When applied to the MPLS data plane, Segment Routing does not introduce any new behavior or any change in the way 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 term 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 meta-data is equivalent between RSVP-TE [RFC3209] and SR.

When a source routed path is expressed with a list of segments additional meta-data 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.

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 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 install labels, only if these 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.

Segment Routing does not introduce signaling between the source and the mid points 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/instruction.

For the MPLS data plane, there are no new requirement 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, [I-D.ietf-6man-segment-routing-header]) which is a type of Routing Extension header as defined in [RFC8200].

The SRH adds some meta-data on 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).

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 SR IPv6 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.

In addition, SR domain boundary routers, by default, MUST apply data plane filters so to only accept packets whose DA and SRH (if any) contain addresses previously advertised as SIDs.

There are a number of security concerns with source routing at the IPv6 data plane [RFC5095]. The new IPv6-based segment routing header is defined in [I-D.ietf-6man-segment-routing-header]. This document also discusses the above security concerns.

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 and hence the number of elements that a transit node has to manage is smaller.

SR OAM use cases and requirements for the MPLS data plane are defined in [I-D.ietf-spring-oam-usecase] and [I-D.ietf-spring-sr-oam-requirement]. SR OAM procedures for the MPLS data plane are defined in [I-D.ietf-mpls-spring-lsp-ping].

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 index or address by the operator is critical for the network behavior to avoid situations like mis-routing. 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 segment routing configuration and operations has been defined in [I-D.ietf-spring-sr-yang].

When Segment Routing 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 SR over the IPv6 data plane, the source routed path is encoded in the SRH as described in [I-D.ietf-6man-segment-routing-header]. The SR IPv6 source routed path is instantiated into the SRH as a list of IPv6 address where the active segment is in the Destination Address (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 it belongs to. Similar to the context of SR over 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. 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

11. 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.

12. References

12.1. 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>.

[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>.

12.2. Informative References

[I-D.ietf-6man-segment-routing-header]
Previdi, S., Filsfils, C., Raza, K., Leddy, J., Field, B., daniel.voyer@bell.ca, d., daniel.bernier@bell.ca, d., Matsushima, S., Leung, I., Linkova, J., Aries, E., Kosugi, T., Vyncke, E., Lebrun, D., Steinberg, D., and R. Raszuk, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-07 (work in progress), July 2017.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", draft-ietf-idr-bgpls-segment-routing-epe-13 (work in progress), June 2017.

[I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and j. jefftant@gmail.com, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-13 (work in progress), June 2017.

[I-D.ietf-mpls-spring-lsp-ping]
Kumar, N., Pignataro, C., Swallow, G., Akiya, N., Kini, S., and M. Chen, "Label Switched Path (LSP) Ping/ Traceroute for Segment Routing IGP Prefix and Adjacency SIDs with MPLS Data-plane", draft-ietf-mpls-spring-lsp-ping-13 (work in progress), October 2017.

[I-D.ietf-ospf-ospfv3-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPFv3 Extensions for Segment Routing", draft-ietf-ospf-ospfv3-segment-routing-extensions-10 (work in progress), September 2017.

[I-D.ietf-ospf-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", draft-ietf-ospf-segment-routing-extensions-21 (work in progress), October 2017.

[I-D.ietf-pce-segment-routing]
Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "PCEP Extensions for Segment Routing", draft-ietf-pce-segment-routing-10 (work in progress), October 2017.

[I-D.ietf-spring-oam-usecase]
Geib, R., Filsfils, C., Pignataro, C., and N. Kumar, "A Scalable and Topology-Aware MPLS Dataplane Monitoring System", draft-ietf-spring-oam-usecase-09 (work in progress), July 2017.

[I-D.ietf-spring-resiliency-use-cases]
Filsfils, C., Previdi, S., Decraene, B., and R. Shakir, "Resiliency use cases in SPRING networks", draft-ietf-spring-resiliency-use-cases-11 (work in progress), May 2017.

[I-D.ietf-spring-segment-routing-central-epe]
Filsfils, C., Previdi, S., Aries, E., and D. Afanasiev, "Segment Routing Centralized BGP Egress Peer Engineering", draft-ietf-spring-segment-routing-central-epe-06 (work in progress), June 2017.

[I-D.ietf-spring-segment-routing-mpls]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with MPLS data plane", draft-ietf-spring-segment-routing-mpls-10 (work in progress), June 2017.

[I-D.ietf-spring-sr-oam-requirement]
Kumar, N., Pignataro, C., Akiya, N., Geib, R., Mirsky, G., and S. Litkowski, "OAM Requirements for Segment Routing Network", draft-ietf-spring-sr-oam-requirement-03 (work in progress), January 2017.

[I-D.ietf-spring-sr-yang]
Litkowski, S., Qu, Y., Sarkar, P., and J. Tantsura, "YANG Data Model for Segment Routing", draft-ietf-spring-sr-yang-07 (work in progress), July 2017.

[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>.

[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>.

[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>.

Authors' Addresses

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

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
FR

Email: stephane.litkowski@orange.com

Rob Shakir
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: robjs@google.com