Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 20, 2017

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

February 16, 2017

Segment Routing Architecture
draft-ietf-spring-segment-routing-11

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 and service chain 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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Copyright Notice

Table of Contents

1. Introduction

With Segment Routing (SR), 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 path and service chain while maintaining per-flow state only at the ingress node of the SR domain.

Segment Routing can be directly applied to the MPLS architecture ([RFC3031]) 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 active segment is on the top of the stack. A completed segment is popped off the stack. The addition of a segment is performed with a push.

In the Segment Routing MPLS instantiation, a segment could be of several types:

- an IGP segment,
- a BGP Peering segment,
- an LDP LSP segment,
- an RSVP-TE LSP segment,
- a BGP LSP segment.

The first two (IGP and BGP peering segments) types of segments are defined in this document. The use of the last three types of segments is illustrated in [I-D.ietf-spring-segment-routing-mpls].

Segment Routing can be applied to the IPv6 architecture ([RFC2460]), 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. Upon completion of a segment, a pointer in the new routing header is incremented and indicates the next segment.

Numerous use-cases illustrate the benefits of source routing either for FRR, OAM or Traffic Engineering reasons ([I-D.ietf-spring-oam-usecase]).

This document defines a set of instructions (called segments) that are required to fulfill the described use-cases. These segments can either be used in isolation (one single segment defines the source route of the packet) or in combination (these segments are part of an ordered list of segments that define the source route of the packet).

1.1. Companion Documents

This document defines the SR architecture, its routing model, the IGP-based segments, the BGP-based segments and the service segments.

The problem statement and requirements are described in [RFC7855].

Use cases are described in .[I-D.ietf-spring-ipv6-use-cases], [I-D.ietf-spring-resiliency-use-cases] and [I-D.ietf-spring-oam-usecase].

2. Terminology

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. Examples of SIDs are: an MPLS label, an index value in an MPLS label space, an IPv6 address. Other types of SIDs can be defined in the future.

Segment List: ordered list of SIDs encoding the ordered set of instructions to be applied to the packet as it traverses the SR domain. For example, the topological and service source route of the packet. The Segment List is instantiated as a stack of labels in the MPLS architecture and as an ordered list of IPv6 addresses in the IPv6 architecture.

Segment Routing Domain (SR Domain): the set of nodes participating into 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). Note that an SR domain may also be confined within an IGP instance, in which case it is named SR-IGP Domain.

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 of a packet having the Segment Routing Header (SRH) as defined in [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 the MPLS dataplane the top of the segment list is the topmost (outer) label of the label stack. In the IPv6 dataplane, 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.

CONTINUE: the active segment is not completed and hence remains active. The CONTINUE instruction is implemented as the SWAP instruction in the MPLS dataplane. In IPv6, this is the plain IPv6 forwarding action of a regular IPv6 packet according to its Destination Address.

SR Global Block (SRGB): local property of an SR node. In the MPLS architecture, SRGB is the set of local labels reserved for global segments. Using the same SRGB on all nodes within the SR domain ease operations and troubleshooting and is expected to be a deployment guideline. In the IPv6 architecture, the equivalent of the SRGB is in fact the set of addresses used as global segments. Since there are no restrictions on which IPv6 address can be used, the concept of the SRGB includes all IPv6 global address space used within the SR domain.

Global Segment: the related instruction is supported by all the SR-capable nodes in the domain. In the MPLS architecture, a global segment is represented by a globally-unique index. The related local label at a given node N is found by adding the globally-unique index to the SRGB of node N. In the IPv6 architecture, a global segment is a globally-unique IPv6 address.

Local Segment: the related instruction is supported only by the node originating it. In the MPLS architecture, this is a local label outside the SRGB. In the IPv6 architecture, this can be any IPv6

address whose reachability is not advertised in any routing protocol (hence, the segment is known only by the local node).

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. An IGP-Prefix Segment is global (unless explicitly advertised otherwise) within the SR IGP instance/topology and 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.

Note that for all of the above, the SID is often used to refer to the Segment itself. For example, Prefix-SID is sometimes used to refer to Prefix Segment.

SR Tunnel: a list of segments to be pushed on the packets directed on the tunnel. The list of segments can be specified explicitly or implicitly via a set of abstract constraints (latency, affinity, SRLG, ...). In the latter case, a constraint-based path computation is used to determine the list of segments associated with the tunnel. The computation can be local or delegated to a PCE server. An SR tunnel can be configured by the operator, provisioned via netconf or

provisioned via PCEP. An SR tunnel can be used for traffic-engineering, OAM or FRR reasons.

Segment List Depth: the number of segments of an SR tunnel. The entity instantiating an SR Tunnel at a node N should be able to discover the depth insertion capability of the node N. The PCEP discovery capability is described in [I-D.ietf-pce-segment-routing].

3. Link-State IGP Segments

Within a link-state IGP 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 IGP domain. Such a path is either expressed as a single IGP segment or a list of multiple IGP segments.

IGP segments require extensions in link-state IGP protocols. IGP extensions are required in order to advertise the IGP segments.

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/IGP domain.

3.1.1. Prefix-SID Algorithm

The IGP protocol extensions for Segment Routing define the Prefix-SID advertisement which includes a set of flags and the algorithm field. The algorithm field has the purpose of associating a given Prefix-SID to a routing algorithm.

In the context of an instance and a topology, multiple Prefix-SID's MAY be allocated to the same IGP Prefix as long as the algorithm value is different in each one.

Multiple instances and topologies are defined in IS-IS and OSPF in: [RFC5120], [RFC6822], [RFC6549] and [RFC4915].

Initially, two "algorithms" have been defined:

- "Shortest Path": this algorithm is the default behavior. The packet is forwarded along the well known ECMP-aware SPF algorithm 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 instruct any router in the path to ignore any possible local policy overriding 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.

An IGP-Prefix Segment identifies the path, to the related prefix, computed as per the algorithm field.

A packet injected anywhere within the SR/IGP domain with an active Prefix-SID will be forwarded along path computed by the algorithm expressed in the algorithm field.

A router MUST drop any SR traffic associated with the SR algorithm to the adjacent router, if the adjacent router has not advertised support for such SR algorithm.

The ingress node of an SR domain validates that the path to a prefix, advertised with a given algorithm, includes nodes all supporting the advertised algorithm. As a consequence, if a node on the path does not support algorithm X, the IGP-Prefix segment will be interrupted and will drop packet on that node. It's the responsibility of the ingress node using a segment to check that all downstream nodes support the algorithm of the segment.

It has to be noted that Fast Reroute (FRR) mechanisms are still compliant with the Strict-SPF. In other words, a packet received with a Strict-SPF SID may be rerouted through a FRR mechanism.

Details of the two defined algorithms are defined in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

3.1.2. MPLS Dataplane

When SR is used over the MPLS dataplane:

- the IGP signaling extension for IGP-Prefix segment includes the P-Flag ([I-D.ietf-isis-segment-routing-extensions]) or the NP-Flag ([I-D.ietf-ospf-segment-routing-extensions]). A Node N advertising a Prefix-SID SID-R for its attached prefix R unsets the P-Flag (or NP-Flag) in order to instruct its connected neighbors to perform the NEXT operation while processing SID-R. This behavior is equivalent to Penultimate Hop Popping in MPLS. When the flag is unset, the neighbors of N MUST perform the NEXT operation while processing SID-R. When the flag is set, the

neighbors of N MUST perform the CONTINUE operation while processing SID-R.

- A Prefix-SID is allocated in the form of an index in the SRGB (or as a local MPLS label) 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, we specifically assume 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). 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 warning for misconfiguration.

- If a node N advertises Prefix-SID SID-R for a prefix R that is attached to N, N MUST either clear the P-Flag in the advertisement of SID-R, or else 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. IPv6 Dataplane

When SR is used over the IPv6 dataplane:

- The Prefix-SID is the prefix itself. No additional identifier is needed for Segment Routing over IPv6.

- Any address belonging to any of the node's prefixes can be used as Prefix-SIDs.

- An operator may want to explicitly indicate which of the node's prefixes can be used as Prefix-SIDs through the setting of a flag (e.g.: using the IGP prefix attribute defined in [RFC7794]) in the routing protocol used for advertising the prefix.

- A global SID is instantiated through any globally advertised IPv6 address.

- A local SID is instantiated through a local IPv6 prefix not being advertised and therefore known only by the local node.

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
```

Regardless Segment Routing, any remote IPv6 node will maintain a plain IPv6 FIB entry for any prefix, no matter if they represent a segment or not.

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 MUST advertise the same prefix with the same SID value.

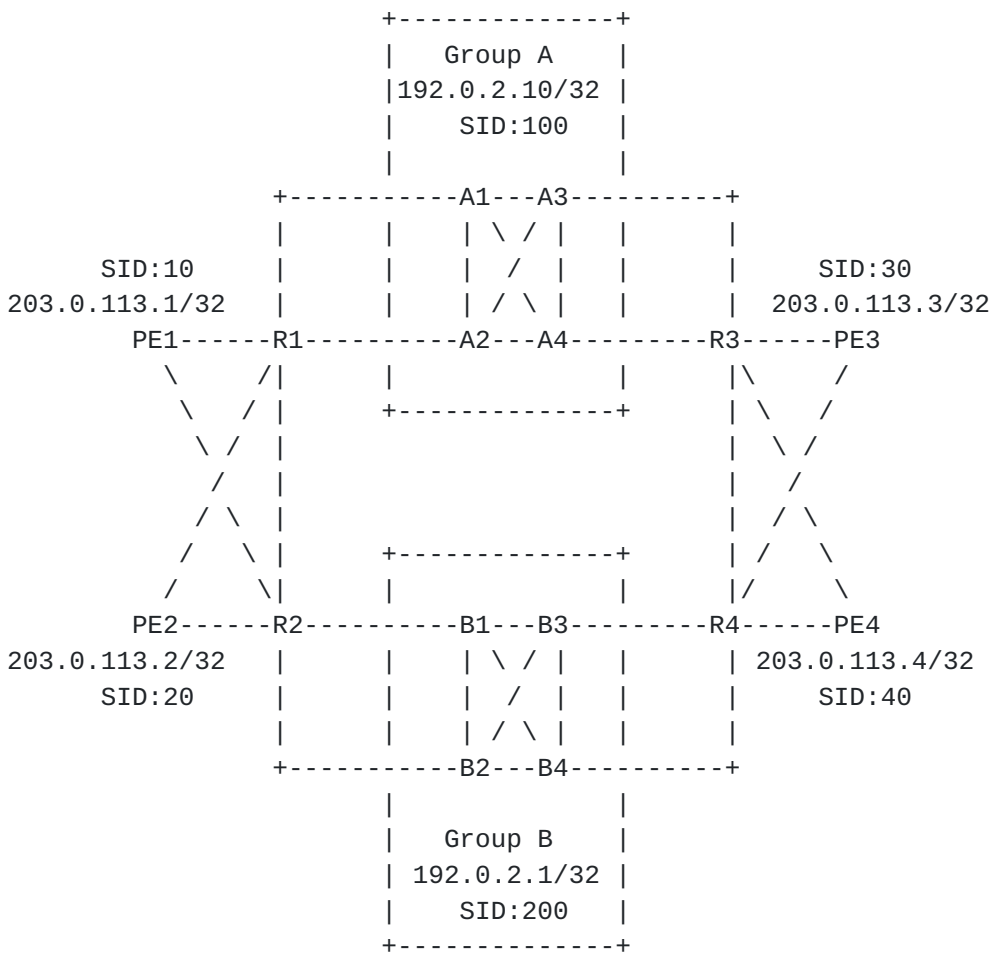

Figure 1: Transit device groups

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 is connected to two routers in 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.

Obviously, the value of the SID following the anycast SID MUST be understood by all nodes advertising the same anycast segment.

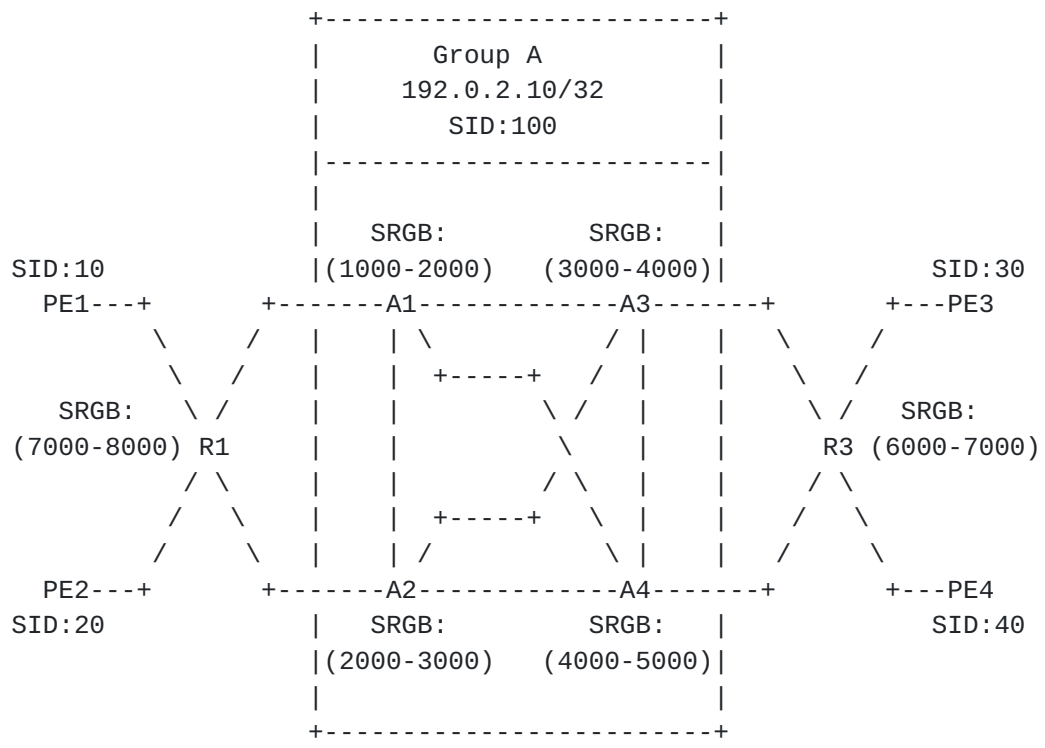

Figure 2: Transit paths via anycast group A

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.

- 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 the a set of flags among which there is the B-flag. When set, the Adj-SID refers to an adjacency that is eligible for protection (e.g.: using IPFRR or MPLS-FRR).

The encodings of the Adj-SID also include the L-flag. When set, the Adj-SID has local significance. By default, the L-flag is set.

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

A node MAY allocate multiple Adj-SIDs to the same adjacency. An example is where the adjacency is established over a bundle interface. Each bundle member MAY have its own Adj-SID.

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

Obviously, 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.

A node MUST install a FIB entry for any Adj-SID of value V attached to data-link L:

```
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 identified by the Adj-SID, regardless 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 link 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 a 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

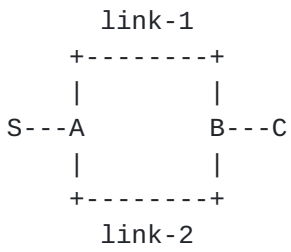

Figure 3: Parallel Links and Adj-SIDs

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.

The weight value is advertised with the Adj-SID as defined in IGP SR extensions documents.

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 other 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. Binding Segment

3.5.1. Mapping Server

A Remote-Binding SID S advertised by the mapping server M for remote prefix R attached to non-SR-capable node N signals the same information as if N had advertised S as a Prefix-SID. Further details are described in the SR/LDP interworking procedures ([I-D.ietf-spring-segment-routing-ldp-interop].

The segment allocation and SRGB Maintenance rules are the same as those defined for Prefix-SID.

3.5.2. Tunnel Head-end

The segment allocation and SRGB Maintenance rules are the same as those defined for Adj-SID. A tunnel attached to a head-end H acts as an adjacency attached to H.

Note: an alternative consists of representing tunnels as forwarding-adjacencies ([RFC4206]). In such case, the tunnel is presented to the routing area as a routing adjacency and is considered as such by all area routers. The Remote-Binding SID is preferred as it allows to advertise the presence of a tunnel without influencing the LSDB and the SPF computation.

3.6. Inter-Area Considerations

In the following example diagram we assume an IGP deployed using areas and where SR has been deployed.

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

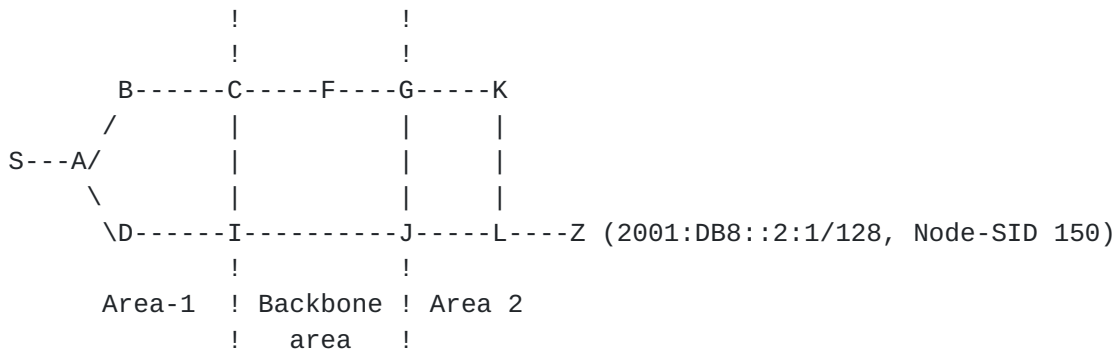

Figure 4: Inter-Area Topology Example

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.

When an ABR propagates a prefix from one area to another it MUST set the R-Flag.

4. 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.

Hereafter, we will define three types of BGP peering segments/SIDs: 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.

- 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 will be defined in a separate document.

5. IGP Mirroring Context Segment

It is beneficial for an IGP node to be able 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], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-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. MPLS Data Plane

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 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] section 2.4 and [RFC5920] section 8.2 already calls for the filtering of MPLS packets on trust boundaries.

8.2. IPv6 Data Plane

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 [RFC2460].

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 globally-unique index or address. 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.

An operator may implement tools in order to audit the network and ensure the good allocation of indexes, SIDs or IP addresses. Conflict detection between SIDs, including Mapping Server binding SIDs, and their resolution are addressed in [I-D.ietf-spring-conflict-resolution].

SR with the MPLS data plane, can be gracefully introduced in an existing LDP [RFC5036] network. This is described in [I-D.ietf-spring-segment-routing-ldp-interop]. SR and LDP may also inter-work. In this case, the introduction of mapping-server may introduce some additional manageability considerations that are discussed in [I-D.ietf-spring-segment-routing-ldp-interop].

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, Les Ginsberg, Ruediger Geib, Hannes Gredler, Pushpasis Sarkar, Eric Rosen and Chris Bowers 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, <http://www.rfc-editor.org/info/rfc2119>.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, DOI 10.17487/RFC2460, December 1998, <http://www.rfc-editor.org/info/rfc2460>.

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

[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, <http://www.rfc-editor.org/info/rfc4206>.

12.2. Informative References

[I-D.ietf-6man-segment-routing-header]
Previdi, S., Filsfils, C., Field, B., Leung, I., Linkova, J., Aries, E., Kosugi, T., Vyncke, E., and D. Lebrun, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-05 (work in progress), February 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-09 (work in progress), October 2016.

[I-D.ietf-mpls-spring-lsp-ping]
Kumar, N., Swallow, G., Pignataro, C., Akiya, N., Kini, S., Gredler, H., and M. Chen, "Label Switched Path (LSP) Ping/Trace for Segment Routing Networks Using MPLS Dataplane", draft-ietf-mpls-spring-lsp-ping-02 (work in progress), December 2016.

[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-07 (work in progress), October 2016.

[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-10 (work in progress), October 2016.

[I-D.ietf-pce-segment-routing]
Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., Raszuk, R., Lopez, V., Tantsura, J., Henderickx, W., and J. Hardwick, "PCEP Extensions for Segment Routing", draft-ietf-pce-segment-routing-08 (work in progress), October 2016.

[I-D.ietf-spring-conflict-resolution]
Ginsberg, L., Psenak, P., Previdi, S., and M. Pilka, "Segment Routing Conflict Resolution", draft-ietf-spring-conflict-resolution-02 (work in progress), October 2016.

[I-D.ietf-spring-ipv6-use-cases]
Brzozowski, J., Leddy, J., Filsfils, C., Maglione, R., and W. Townsley, "IPv6 SPRING Use Cases", draft-ietf-spring-ipv6-use-cases-09 (work in progress), February 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-05 (work in progress), February 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-08 (work in progress), October 2016.

[I-D.ietf-spring-segment-routing-central-epe]
Filsfils, C., Previdi, S., Aries, E., and D. Afanasiev, "Segment Routing Centralized BGP Peer Engineering", draft-ietf-spring-segment-routing-central-epe-03 (work in progress), November 2016.

[I-D.ietf-spring-segment-routing-ldp-interop]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., and S. Litkowski, "Segment Routing interworking with LDP", draft-ietf-spring-segment-routing-ldp-interop-06 (work in progress), February 2017.

[I-D.ietf-spring-segment-routing-mpls]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Shakir, R., jefftant@gmail.com, j., and E. Crabbe, "Segment Routing with MPLS data plane", draft-ietf-spring-segment-routing-mpls-07 (work in progress), February 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-05 (work in progress), October 2016.

[RFC4381] Behringer, M., "Analysis of the Security of BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4381, DOI 10.17487/RFC4381, February 2006, <http://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, <http://www.rfc-editor.org/info/rfc4915>.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <http://www.rfc-editor.org/info/rfc5036>.

[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, <http://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, <http://www.rfc-editor.org/info/rfc5120>.

[RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, DOI 10.17487/RFC5920, July 2010, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc6549>.

[RFC6822] Previdi, S., Ed., Ginsberg, L., Shand, M., Roy, A., and D. Ward, "IS-IS Multi-Instance", RFC 6822, DOI 10.17487/RFC6822, December 2012, <http://www.rfc-editor.org/info/rfc6822>.

[RFC7794] Ginsberg, L., Ed., Decraene, B., Previdi, S., Xu, X., and U. Chunduri, "IS-IS Prefix Attributes for Extended IPv4 and IPv6 Reachability", RFC 7794, DOI 10.17487/RFC7794, March 2016, <http://www.rfc-editor.org/info/rfc7794>.

[RFC7855] Previdi, S., Ed., Filsfils, C., Ed., Decraene, B., Litkowski, S., Horneffer, M., and R. Shakir, "Source Packet Routing in Networking (SPRING) Problem Statement and Requirements", RFC 7855, DOI 10.17487/RFC7855, May 2016, <http://www.rfc-editor.org/info/rfc7855>.

Authors' Addresses

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

Email: cfilsfil@cisco.com

Stefano Previdi (editor)
Cisco Systems, Inc.
Via Del Serafico, 200
Rome 00142
Italy

Email: sprevidi@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