Network Working Group
Internet Draft
Intended status: Standards Track
Expires: August 2018

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

February 23, 2018

Segment Routing with MPLS data plane
draft-ietf-spring-segment-routing-mpls-12

Abstract

Segment Routing (SR) leverages the source routing paradigm. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header. In the MPLS dataplane, the SR header is instantiated through a label stack. This document specifies the forwarding behavior to allow instantiating SR over the MPLS dataplane.

Requirements Language

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

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

The Segment Routing architecture [I-D.ietf-spring-segment-routing] can be directly applied to the MPLS architecture with no change in the MPLS forwarding plane. This document specifies the forwarding plane behavior to allow Segment Routing to operate on top of the MPLS data plane. This document does not address the control plane behavior. Control plane behavior is specified in other documents such as [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions], and [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

The Segment Routing problem statement is described in [RFC7855].

Co-existence of SR over MPLS forwarding plane with LDP [RFC5036] is specified in [I-D.ietf-spring-segment-routing-ldp-interop].

Policy routing and traffic engineering using segment routing can be found in [I.D. filsfils-spring-segment-routing-policy]

2. MPLS Instantiation of Segment Routing

MPLS instantiation of Segment Routing fits in the MPLS architecture as defined in [RFC3031] both from a control plane and forwarding plane perspective:

- From a control plane perspective, [RFC3031] does not mandate a single signaling protocol. Segment Routing makes use of various control plane protocols such as link State IGPs [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions]. The flooding mechanisms of link state IGPs fits very well with label stacking on ingress. Future control layer protocol and/or policy/configuration can be used to specify the label stack.

- From a forwarding plane perspective, Segment Routing does not require any change to the forwarding plane because Segment ID (SIDs) are instantiated as MPLS labels and the Segment routing header instantiated as a stack of MPLS labels.

We call "MPLS Control Plane Client (MCC)" any control plane entity installing forwarding entries in the MPLS data plane. IGPs with SR extensions [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions], [I-D.ietf-ospf-ospfv3-segment-routing-extensions] and LDP [RFC5036] are examples of MCCs. Local configuration and policies applied on a router are also examples of MCCs.

2.1. Supporting Multiple Forwarding Behaviors for the Same Prefix

The SR architecture does not prohibit having more than one SID for the same prefix. In fact, by allowing multiple SIDs for the same prefix, it is possible to have different forwarding behaviors (such as different paths, different ECMP/UCMP behaviors,...,etc) for the same destination.

Instantiating Segment routing over the MPLS forwarding plane fits seamlessly with this principle. An operator may assign multiple MPLS labels or indices to the same prefix and assign different forwarding behaviors to each label/SID. The MCC in the network downloads different MPLS labels/SIDs to the FIB for different forwarding behaviors. The MCC at the entry of an SR domain or at any point in the domain can choose to apply a particular forwarding behavior to a particular packet by applying the PUSH action to that packet using the corresponding SID.

2.2. SID Representation in the MPLS Forwarding Plane

When instantiating SR over the MPLS forwarding plane, a SID is represented by an MPLS label or an index [I-D.ietf-spring-segment-routing].

A global segment MUST be a label, or an index which may be mapped to an MPLS label using the SRGB of the node installing the global segment in its FIB/receiving the labeled packet. Section 2.4 specifies the procedure to map a global segment represented by an index to an MPLS label within the SRGB.

The MCC MUST ensure that any label value corresponding to any SID it installs in the forwarding plane follows the following rules:

- The label value MUST be unique within the router on which the MCC is running. i.e. the label MUST only be used to represent the SID.

- The label value MUST NOT be identical to or within the range of any reserved label value or range [reserved-MPLS], respectively.

2.3. Segment Routing Global Block and Local Block

The concepts of Segment Routing Global Block (SRGB) and global SID are explained in [I-D.ietf-spring-segment-routing]. In general, the SRGB need not be a contiguous range of labels.

For the rest of this document, the SRGB is specified by the list of MPLS Label ranges [Ll(1),Lh(1)], [Ll(2),Lh(2)],..., [Ll(k),Lh(k)] where Ll(i) =< Lh(i).

The following rules apply to the list of MPLS ranges representing the SRGB

- The list of ranges comprising the SRGB MUST NOT overlap.

- Every range in the list of ranges specifying the SRGB MUST NOT cover or overlap with a reserved label value or range [reserved-MPLS], respectively.

- If the SRGB of a node does not conform to the structure specified in this section or to the previous two rules, then this SRGB is completely ignored and the node is treated as if it does not have an SRGB.

- The list of label ranges MUST only be used to instantiate global SIDs into the MPLS forwarding plane

Local segments MAY be allocated from the Segment Routing Local Block (SRLB) [I-D.ietf-spring-segment-routing] or from any unused label as long as it does not use a reserved label. The SRLB consists of the range of local labels reserved by the node for certain local segments. In a controller-driven network, some controllers or

applications MAY use the control plane to discover the available set of local SIDs on a particular router [I.D. filsfils-spring-segment-routing-policy]. Just like SRGB, the SRLB need not be a single contiguous range of label, except the SRGB MUST only be used to instantiate global SIDs into the MPLS forwarding plane. Hence it is specified the same way and follows the same rules SRGB is specified above in this sub-section.

2.4. Mapping a SID Index to an MPLS label

This sub-section specifies how the MPLS label value is calculated given the index of a SID. The value of the index is determined by an MCC such as IS-IS [I-D.ietf-isis-segment-routing-extensions] or OSPF [I-D.ietf-ospf-segment-routing-extensions]. This section only specifies how to map the index to an MPLS label. The calculated MPLS label is downloaded to the FIB, sent out with a forwarded packet, or both.

Consider a SID represented by the index "I". Consider an SRGB as specified in Section 2.3. The total size of the SRGB, represented by the variable "Size", is calculated according to the formula:

$$size = Lh(1)- Ll(1) + 1 + Lh(2)- Ll(2) + 1 + ... + Lh(k)- Ll(k) + 1$$

The following rules MUST be applied by the MCC when calculating the MPLS label value corresponding the SID index value "I".

- 0 =< I < size. If the index "I" does not satisfy the previous inequality, then the label cannot be calculated.
- The label value corresponding to the SID index "I" is calculated as follows

```
o j = 1 , temp = 0

o While temp + Lh(j)- Ll(j) < I

     . temp = temp + Lh(j)- Ll(j) + 1

     . j = j+1

o label = I - temp + Ll(j)
```

2.5. Incoming Label Collision

MPLS Architecture [RFC3031] defines Forwarding Equivalence Class (FEC) as the set of packets which are forwarded in the same manner

(e.g.,over the same path, with the same forwarding treatment) and are bound to the same MPLS incoming (local) label. In Segment-Routing MPLS, local label serves as the SID (possibly via an index indirection) for given FEC.

We define Segment Routing (SR) FEC as one of the following [I-D.ietf-spring-segment-routing]:

- o (Prefix, Routing Instance, Topology, Algorithm), where a topology is identified by a set of links with metrics. For the purpose of incoming label collision resolution, the same numerical value SHOULD be used on all routers to identify the same set of links with metrics. For MCCs where the "Topology" and/or "Algorithm" fields are not defined, the numerical value of zero MUST be used for these two fields. For the purpose of incoming label collision resolution, a routing instance is identified by a single incoming label downloader to FIB. Two MCCs running on the same router are considered different routing instances if the only way the two instances can know about the other's incoming labels is through redistribution. The numerical value used to identify a routing instance MAY be derived from other configuration or MAY be explicitly configured. If it is derived from other configuration, then the same numerical value SHOULD be derived from the same configuration as long as the configuration survives router reload. If the derived numerical value varies for the same configuration, then an implementation SHOULD make numerical value used to identify a routing instance configurable.

- o (next-hop, outgoing interface), where the outgoing interface is physical or virtual.

- o (Endpoint, Color) representing an SR policy [I.D. filsfils-spring-segment-routing-policy]

This section covers handling the scenario where, because of an error/misconfiguration, more than one SR FEC as defined in this section, map to the same incoming MPLS label.

An incoming label collision occurs if the SIDs of the set of FECs {FEC1, FEC2,..., FECk} maps to the same incoming SR MPLS label "L1".

The objective of the following steps is to deterministically install in the MPLS Incoming Label MAP, also known as label FIB, a single FEC with the incoming label "L1". Remaining FECs may be installed in the IP FIB without incoming label.

The procedure in this section relies completely on the local FEC and label database within a given router.

The collision resolution procedure is as follows

1. Given the SIDs of the set of FECs, {FEC1, FEC2,..., FECk} map to the same MPLS label "L1".

2. Within an MCC, apply tie-breaking rules to select one FEC only and assign the label to it. The losing FECs are handled as if no labels are attached to them. The losing FECs with a non-zero algo are not installed in FIB.

   a. If the same set of FECs are attached to the same label "L1", then the tie-breaking rules MUST always select the same FEC irrespective of the order in which the FECs and the label "L1" are received. In other words, the tie-breaking rule MUST be deterministic. For example, a first-come-first-serve tie-breaking is not allowed.

3. If there is still collision between the FECs belonging to different MCCs, then re-apply the tie-breaking rules to the remaining FECs to select one FEC only and assign the label to that FEC

4. Install into the IP FIB the selected FEC and its incoming label in the label FIB.

5. The remaining FECs with a zero algorithm are installed in the FIB natively, such as pure IP entries in case of Prefix FEC, without any incoming labels corresponding to their SIDs. The remaining FECs with a non-zero algorithm are not installed in the FIB.

2.5.1. Tie-breaking Rules

The default tie-breaking rules SHOULD be as follows:

1. if FECi has the lowest FEC administrative distance among the competing FEC's as defined in this section below, filter away all the competing FEC's with higher administrative distance.

2. if more than one competing FEC remains after step 1, sort them and select the smallest numerical FEC value

These rules deterministically select the FEC to install in the MPLS forwarding plane for the given incoming label.

This document defines the default tie breaking rules that SHOULD be implemented. An implementation may choose to implement additional tie-breaking rules. All routers in a routing domain SHOULD use the same tie-breaking rules to maximize forwarding consistency.

Each FEC is assigned an administrative distance. The FEC administrative distance is encoded as an 8-bit value. The lower the value, the better the administrative distance.

The default FEC administrative distance order starting from the lowest value SHOULD be

- Explicit SID assignment to a FEC that maps to a label outside the SRGB irrespective of the owner MCC. An explicit SID assignment is a static assignment of a label to a FEC such that the assignment survives router reboot.

  - An example of explicit SID allocation is static assignment of a specific label to an adjacency SID.

  - An implementation of explicit SID assignment MUST guarantee collision freeness on the same router

- Dynamic SID assignment:

  - For all FEC types except for SR policy, use the default administrative distance depending on the implementation

  - Binding SID [I-D.ietf-spring-segment-routing] assigned to SR Policy

A user SHOULD ensure that the same administrative distance preference is used on all routers to maximize forwarding consistency.

The numerical sort across FEC's SHOULD be performed as follows:

- Each FEC is assigned a FEC type encoded in 8 bits. The following are the type code point for each SR FEC defined at the beginning of this Section:

  - 120: (Prefix, Routing Instance, Topology, Algorithm)

  - 130: (next-hop, outgoing interface)

  - 140: (Endpoint, Color) representing an SR policy

- The fields of each FEC are encoded as follows

  - Routing Instance ID represented by 16 bits. For routing instances that are identified by less than 16 bits, encode the Instance ID in the least significant bits while the most significant bits are set to zero

  - Address Family represented by 8 bits, where IPv4 encoded as 100 and IPv6 is encoded as 110

  - All addresses are represented in 128 bits as follows

    - IPv6 address is encoded natively

    - IPv4 address is encoded in the most significant bits and the remaining bits are set to zero

  - All prefixes are represented by 128.

    - A prefix is encoded in the most significant bits and the remaining bits are set to zero.

    - The prefix length is encoded before the prefix

  - Topology ID is represented by 16 bits. For routing instances that identify topologies using less than 16 bits, encode the topology ID in the least significant bits while the most significant bits are set to zero

  - Algorithm is encoded in a 16 bits field.

  - The Color ID is encoded using 16 bits

- Choose the set of FECs of the smallest FEC type code point

- Out of these FECs, choose the FECs with the smallest address family code point

- Encode the remaining set of FECs as follows

  - Prefix, Routing Instance, Topology, Algorithm: (Prefix Length, Prefix, SR Algorithm, routing_instance_id, Topology)

  - (next-hop, outgoing interface): (next-hop, outgoing_interface_id)

  - (Endpoint, Color): (Endpoint_address, Color_id)

- Select the FEC with the smallest numerical value

2.5.2. Redistribution between Routing Protocol Instances

The following rule SHOULD be applied when redistributing SIDs with prefixes between routing protocol instances:

- o If the receiving instance's SRGB is the same as the SRGB of origin instance, THEN
    - o the index is redistributed with the route
- o Else
    - o the index is not redistributed and if needed it is the duty of the receiving instance to allocate a fresh index relative to its own SRGB

It is outside the scope of this document to define local node behaviors that would allow to map the original index into a new index in the receiving instance via the addition of an offset or other policy means.

2.5.2.1. Illustration

```
        A----IS-IS----B---OSPF----C-1.1.1.1/32 (20001)
```

Consider the simple topology above.

- o A and B are in the IS-IS domain with SRGB [16000-17000]
- o B and C are in OSPF domain with SRGB [20000-21000]
- o B redistributes 1.1.1.1/32 into IS-IS domain
- o In that case A learns 1.1.1.1/32 as an IP leaf connected to B as usual for IP prefix redistribution
- o However, according to the redistribution rule above rule, B does not advertise any index with 1.1.1.1/32 into IS-IS because the SRGB is not the same.

2.6. Outgoing Label Collision

For the determination of the outgoing label to use, the ingress node pushing new segments, and hence a stack of MPLS labels, MUST use, for a given FEC, the same label that has been selected by the node receiving the packet with that label exposed as top label. So in case of incoming label collision on this receiving node, the ingress node MUST resolve this collision using this same "Incoming Label Collision resolution procedure", using the data of the receiving node.

In the general case, the ingress node may not have exactly have the same data of the receiving node, so the result may be different. This is under the responsibility of the network operator. But in typical case, e.g. where a centralized node or a distributed link state IGP is used, all nodes would have the same database. However to minimize the chance of misforwarding, a FEC that loses its incoming label to the tie-breaking rules specified in Section 2.5 MUST NOT be installed in FIB with an outgoing segment routing label based on the SID corresponding to the lost incoming label.

2.7. PUSH, CONTINUE, and NEXT

PUSH, NEXT, and CONTINUE are operations applied by the forwarding plan. The specifications of these operations can be found in [I-D.ietf-spring-segment-routing]. This sub-section specifies how to implement each of these operations in the MPLS forwarding plane.

2.7.1. PUSH

PUSH corresponds to pushing one or more labels on top of an incoming packet then sending it out of a particular physical interface or virtual interface, such as UDP tunnel [RFC7510] or L2TPv3 tunnel [RFC4817], towards a particular next-hop. Sections 2.10 and 2.11 specify additional details about forwarding behavior.

2.7.2. CONTINUE

In the MPLS forwarding plane, the CONTINUE operation corresponds to swapping the incoming label with an outgoing label. The value of the outgoing label is calculated as specified in Sections 2.10 and 2.11.

2.7.3. NEXT

In the MPLS forwarding plane, NEXT corresponds to popping the topmost label. The action before and/or after the popping depends on the instruction associated with the active SID on the received packet prior to the popping. For example suppose the active SID in the received packet was an Adj-SID [I-D.ietf-spring-segment-routing], then on receiving the packet, the node applies NEXT operation, which corresponds to popping the top most label, and then sends the packet out of the physical or virtual interface (e.g. UDP tunnel [RFC7510]

or L2TPv3 tunnel [RFC4817]) towards the next-hop corresponding to the adj-SID.

2.8. MPLS Label downloaded to FIB corresponding to Global and Local SIDs

The label corresponding to the global SID "Si" represented by the global index "I" downloaded to FIB is used to match packets whose active segment (and hence topmost label) is "Si". The value of this label is calculated as specified in Section 2.4.

For Local SIDs, the MCC is responsible for downloading the correct label value to FIB. For example, an IGP with SR extensions I-D.ietf-isis-segment-routing-extensions, I-D.ietf-ospf-segment-routing-extensions] allocates and downloads the MPLS label corresponding to an IGP-adjacency-SID [I-D.ietf-spring-segment-routing].

2.9. Active Segment

When instantiated in the MPLS domain, the active segment on a packet corresponds to the topmost label on the packet that is calculated according to the procedure specified in Sections 2.10 and 2.11. When arriving at a node, the topmost label corresponding to the active SID matches the MPLS label downloaded to FIB as specified in Section 2.8.

2.10. Forwarding behavior for Global SIDs

This section specifies forwarding behavior, including the outgoing label(s) calculations corresponding to a global SID when applying PUSH, CONTINUE, and NEXT operations in the MPLS forwarding plane.

This document covers the calculation of outgoing label for the top label only. The case where outgoing label is not the top label and is part of a stack of labels that instantiates a routing policy or a traffic engineering tunnel is covered in other documents such as [I.D.filsfils-spring-segment-routing-policy].

2.10.1. Forwarding Behavior for PUSH and CONTINUE Operation for Global SIDs

Suppose an MCC on a router "R0" determines that PUSH or CONTINUE operation is to be applied to an incoming packet whose active SID is the global SID "Si" represented by the global index "I" and owned by the router Ri before sending the packet towards a neighbor "N" directly connected to "R0" through a physical or virtual interface such as UDP tunnel [RFC7510] or L2TPv3 tunnel [RFC4817].

The method by which the MCC on router "R0" determines that PUSH or CONTINUE operation must be applied using the SID "Si" is beyond the scope of this document. An example of a method to determine the SID "Si" for PUSH operation is the case where IS-IS [I-D.ietf-isis-segment-routing-extensions] receives the prefix-SID "Si" sub-TLV advertised with prefix "P/m" in TLV 135 and the destination address of the incoming IPv4 packet is covered by the prefix "P/m".

For CONTINUE operation, an example of a method to determine the SID "Si" is the case where IS-IS [I-D.ietf-isis-segment-routing-extensions] receives the prefix-SID "Si" sub-TLV advertised with prefix "P" in TLV 135 and the top label of the incoming packet matches the MPLS label in FIB corresponding to the SID "Si" on the router "R0".

The forwarding behavior for PUSH and CONTINUE corresponding to the SID "Si"

- If the neighbor "N" does not support SR or "I" does not satisfy the inequality specified in Section 2.4 for the SRGB of the neighbor "N"

  - If it is possible to send the packet towards the neighbor "N" using standard MPLS forwarding behavior as specified in {RFC3031] and [RFC3032], then forward the packet. The method by which a router decides whether it is possible to send the packet to "N" or not is beyond the scope of this document. For example, the router "R0" can use the downstream label determined by another MCC, such as LDP [RFC5036], to send the packet.

  - Else if there are other useable next-hops, then use other next-hops to forward the incoming packet. The method by which the router "R0" decides on the possibility of using other next-hops is beyond the scope of this document. For example, the MCC on "R0" may chose the send an IPv4 packet without pushing any label to another next-hop.

  - Otherwise drop the packet.

- Else

  - Calculate the outgoing label as specified in Section 2.4 using the SRGB of the neighbor "N"

  - If the operation is PUSH

. Push the calculated label according the MPLS label pushing rules specified in [RFC3032]

- Else

. swap the incoming label with the calculated label according to the label swapping rules in [RFC3032]

- Send the packet towards the neighbor "N"

2.10.2. Forwarding Behavior for NEXT Operation for Global SIDs

As specified in Section 2.7.3 NEXT operation corresponds to popping the top most label. The forwarding behavior is as follows

- Pop the topmost label

- Apply the instruction associated with the incoming label prior to popping

The action on the packet after popping the topmost label depends on the instruction associated with the incoming label as well as the contents of the packet right underneath the top label that got popped. Examples of NEXT operation are described in Section 3.

2.11. Forwarding Behavior for Local SIDs

This section specifies the forwarding behavior for local SIDs when SR is instantiated over the MPLS forwarding plane.

2.11.1. Forwarding Behavior Corresponding to PUSH Operation on Local SIDs

Suppose an MCC on a router "R0" determines that PUSH operation is to be applied to an incoming packet using the local SID "Si" before sending the packet towards a neighbor "N" directly connected to R0 through a physical or virtual interface such as UDP tunnel [RFC7510] or L2TPv3 tunnel [RFC4817].

An example of such local SID is an IGP-Adj-SID allocated and advertised by IS-IS [I-D.ietf-isis-segment-routing-extensions]. The method by which the MCC on "R0" determines that PUSH operation is to be applied to the incoming packet is beyond the scope of this document. An example of such method is backup path used to protect

against a failure using Ti-LFA [I.D.bashandy-rtgwg-segment-routing-ti-lfa].

As mentioned in [I-D.ietf-spring-segment-routing], a local SID is specified by an MPLS label. Hence the PUSH operation for a local SID is identical to label push operation [RFC3032] using any MPLS label. The forwarding action after pushing the MPLS label corresponding to the local SID is also determined by the MCC. For example, if the PUSH operation was done to forward a packet over a backup path calculated using Ti-LFA, then the forwarding action may be sending the packet to a certain neighbor that will in turn continue to forward the packet along the backup path

2.11.2. Forwarding Behavior Corresponding to CONTINUE Operation for Local SIDs

A local SID on a router "R0" corresponds to a local label such as an IGP adj-SID. In such scenario, the outgoing label towards a next-hop "N" is determined by the MCC running on the router "R0"and the forwarding behavior for CONTINUE operation is identical to swap operation [RFC3032] on an MPLS label.

2.11.3. Outgoing label for NEXT Operation for Local SIDs

NEXT operation for Local SIDs is identical to NEXT operation for global SIDs specified in Section 2.10.2.

3. IGP Segments Examples

Consider the network diagram of Figure 1 and the IP address and IGP Segment allocation of Figure 2. Assume that the network is running IS-IS with SR extensions [I-D.ietf-isis-segment-routing-extensions]. The following examples can be constructed.

```
                   +--------+
                  /          \
   R0-----R1-----R2----------R3-----R8
                 | \        / |
                 |  +--R4--+  |
                 |            |
                 +-----R5-----+
```

Figure 1: IGP Segments - Illustration

```
       +-------------------------------------------------------------+
       | IP address allocated by the operator:                       |
       |                     192.0.2.1/32 as a loopback of R1        |
       |                     192.0.2.2/32 as a loopback of R2        |
       |                     192.0.2.3/32 as a loopback of R3        |
       |                     192.0.2.4/32 as a loopback of R4        |
       |                     192.0.2.5/32 as a loopback of R5        |
       |                     192.0.2.8/32 as a loopback of R8        |
       |              198.51.100.9/32 as an anycast loopback of R4   |
       |              198.51.100.9/32 as an anycast loopback of R5   |
       |                                                             |
       | SRGB defined by the operator as 1000-5000                   |
       |                                                             |
       | Global IGP SID indices allocated by the operator:           |
       |                     1 allocated to 192.0.2.1/32             |
       |                     2 allocated to 192.0.2.2/32             |
       |                     3 allocated to 192.0.2.3/32             |
       |                     4 allocated to 192.0.2.4/32             |
       |                     8 allocated to 192.0.2.8/32             |
       |                  1009 allocated to 198.51.100.9/32          |
       |                                                             |
       | Local IGP SID allocated dynamically by R2                   |
       |                    for its "north" adjacency to R3: 9001    |
       |                    for its "north" adjacency to R3: 9003    |
       |                    for its "south" adjacency to R3: 9002    |
       |                    for its "south" adjacency to R3: 9003    |
       +-------------------------------------------------------------+
```

Figure 2: IGP Address and Segment Allocation - Illustration

3.1. Example 1

Suppose R1 wants to send an IPv4 packet P1 to R8. In this case, R1 needs to apply PUSH operation to the IPv4 packet.

Remember that the SID index "8" is a global IGP segment attached to the IP prefix 192.0.2.8/32. Its semantic is global within the IGP domain: any router forwards a packet received with active segment 8 to the next-hop along the ECMP-aware shortest-path to the related prefix.

R2 is the next-hop along the shortest path towards R8. By applying the steps in Section 2.8 the local label downloaded to R1's FIB corresponding to the global SID index 8 is 1008 because the SRGB of R2 is [1000,5000] as shown in Figure 2.

Because the packet is IPv4, R1 applies the PUSH operation using the label value 1008 as specified in 2.10.1. The resulting MPLS header will have the "S" bit [RFC3032] set because it is followed directly by an IPv4 packet.

The packet arrives at router R2. Because the top label 1008 corresponds to the IGP SID "8", which is the prefix-SID attached to the prefix 192.0.2.8/32 owned by the R8, then the instruction associated with the SID is "forward the packet using all ECMP/UCMP interfaces and all ECMP/UCMP next-hop(s) along the shortest path towards R8". Because R2 is not the penultimate hop, R2 applies the CONTINUE operation to the packet and sends it to R3 using one of the two links connected to R3 with top label 1008 as specified in Section 2.10.1.

R3 receives the packet with top label 1008. Because the top label 1008 corresponds to the IGP SID "8", which is the prefix-SID attached to the prefix 192.0.2.8/32 owned by the R8, then the instruction associated with the SID is "send the packet using all ECMP interfaces and all next-hop(s) along the shortest path towards R8". Because R3 is the penultimate hop, R3 applies NEXT operation then sends the packet to R8. The NEXT operation results in popping the outer label and sending the packet as a pure IPv4 packet to R8. The

In conclusion, the path followed by P1 is R1-R2--R3-R8. The ECMP-awareness ensures that the traffic be load-shared between any ECMP path, in this case the two north and south links between R2 and R3.

3.2. Example 2

Suppose the right most router R0 wants to send a packet P2 to R8 over the path <R2, (north link between R2 and R3)>. In that case, the router R0 needs to use the SID list <2, 9001, 8>. Using the calculation techniques specified in Section 2.10 and 2.11 the resulting label stack starting from the topmost label is <1002, 9001, 1008>.

The MPLS label 1002 is the MPLS instantiation of the global IGP segment index 2 attached to the IP prefix 192.0.2.2/32. Its semantic is global within the IGP domain: any router forwards a packet received with active segment 1002 to the next-hop along the shortest-path to the related prefix.

The MPLS label 9001 is a local IGP segment attached by node R2 to its north link to R3. Its semantic is local to node R2: R2 applies NEXT operation, which corresponding to popping the outer label, then

switches a packet received with active segment 9001 towards the north link to R3.

In conclusion, the path followed by P2 is R0-R1-R2-north-link-R3-R8.

3.3. Example 3

R0 may send a packet P3 along the same exact path as P2 using a different segment list <2,9003,8> which corresponds to the label stack <1002, 9003, 1008>.

9003 is a local IGP segment attached by node R2 to both its north and south links to R3. Its semantic is local to node R2: R2 applies NEXT operation, which corresponds to popping the top label, then switches a packet received with active segment 9003 towards either the north or south links to R3 (e.g. per-flow loadbalancing decision).

In conclusion, the path followed by P3 is R0-R1-R2-any-link-R3-R8.

3.4. Example 4

R0 may send a packet P4 to R8 while avoiding the links between R2 and R3 by pushing the SID list <4,8>, which corresponds to the label stack <1004, 1008>.

1004 is a global IGP segment attached to the IP prefix 192.0.2.4/32. Its semantic is global within the IGP domain: any router forwards a packet received with active segment 1004 to the next-hop along the shortest-path to the related prefix.

In conclusion, the path followed by P4 is R0-R1-R2-R4-R3-R8.

3.5. Example 5

R0 may send a packet P5 to R8 while avoiding the links between R2 and R3 and still benefiting from all the remaining shortest paths (via R4 and R5) by pushing the SID list <1009, 8> which corresponds to the label stack <2009, 1008> using the steps specified in Sections 2.10 and 2.11.

1009 is a global anycast-SID [I-D.ietf-spring-segment-routing] attached to the anycast IP prefix 198.51.100.9/32. Its semantic is global within the IGP domain: any router forwards a packet received with top label 2009 (corresponding to the active segment 1009) to the next-hop along the shortest-path to the related prefix.

In conclusion, the path followed by P5 is either R0-R1-R2-R4-R3-R8 or R0-R1-R2-R5-R3-R8.

4. IANA Considerations

This document does not make any request to IANA.

5. Manageability Considerations

This document describes the applicability of Segment Routing over the MPLS data plane. Segment Routing does not introduce any change in the MPLS data plane. Manageability considerations described in [I-D.ietf-spring-segment-routing] applies to the MPLS data plane when used with Segment Routing.

6. Security Considerations

This document does not introduce additional security requirements and mechanisms other than the ones described in [I-D.ietf-spring-segment-routing].

7. Contributors

The following contributors have substantially helped the definition and editing of the content of this document:

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
Email: saku@ytti.fi

8. Acknowledgements

The authors would like to thank Les Ginsberg and Shah Himanshu for their comments on this document.

This document was prepared using 2-Word-v2.0.template.dot.

9. References

9.1. Normative References

[I-D.ietf-spring-segment-routing] Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-12 (work in progress), June 2017.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 0.17487/RFC2119, March 1997, <http://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, <http://www.rfc-editor.org/info/rfc3031>.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <http://www.rfc-editor.org/info/rfc3032>.

[reserved-MPLS] "Special-Purpose Multiprotocol Label Switching (MPLS) Label Values", <https://www.iana.org/assignments/mpls-label-value>

9.2. Informative References

[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-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-09 (work in progress), March 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-16 (work in progress), May 2017.

[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-08 (work in progress), June 2017.

[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>.

[RFC5036] Andersson, L., Acreo, AB, Minei, I., Thomas, B., " LDP Specification", RFC5036, DOI 10.17487/RFC5036, October 2007, <https://www.rfc-editor.org/info/rfc5036>

[RFC7510] Xu, X., Sheth, N., Yong, L., Callon, R., and D. Black, "Encapsulating MPLS in UDP", RFC 7510, DOI 10.17487/RFC7510, April 2015, <https://www.rfc-editor.org/info/rfc7510>.

[RFC4817] Townsley, M., Pignataro, C., Wainner, S., Seely, T., Young, T., "Encapsulation of MPLS over Layer 2 Tunneling Protocol Version 3", RFC4817, DOI 10.17487/RFC4817, March 2007, <https://www.rfc-editor.org/info/rfc4817>

[I-D.ietf-idr-segment-routing-te-policy] Previdi, S., Filsfils, C., Mattes, P., Rosen, E., Lin, S., " Advertising Segment Routing Policies in BGP", draft- ietf-idr-segment-routing-te-policy-00 (work in progress), July 2017

[I.D. filsfils-spring-segment-routing-policy] Filsfils, C., Sivabalan, S., Raza, K., Liste, J. , Clad, F., Voyer, D., Lin, S., Bogdanov, A., Horneffer, M., Steinberg, D., Decraene, B. , Litkowski, S., " Segment Routing Policy for Traffic Engineering", draft-filsfils-spring-segment-routing-policy-01 (work in progress), July 2017

Authors' Addresses

Ahmed Bashandy
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA 95134
US

Email: bashandy@cisco.com

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

Email: cfilsfil@cisco.com

Stefano Previdi (editor)
Cisco Systems, Inc.
Italy

Email: stefano@previdi.net

Bruno Decraene
Orange
FR

Email: bruno.decraene@orange.com

Stephane Litkowski
Orange
FR

Email: stephane.litkowski@orange.com

Rob Shakir
Google
US

Email: robjs@google.com