Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 28, 2017

S. Previdi, Ed.
C. Filsfils
A. Sreekantiah
S. Sivabalan
Cisco Systems, Inc.
P. Mattes
Microsoft
E. Rosen
Juniper Networks
S. Lin
Google

February 24, 2017

Advertising Segment Routing Policies in BGP
draft-previdi-idr-segment-routing-te-policy-05

Abstract

This document defines a new BGP SAFI with a new NLRI in order to advertise an explicit path of a Segment Routing Policy (SR Policy). An SR Policy is a set of dynamic and/or explicit paths each represented by one or more segment lists. The path of the SR Policy is advertised along with the Tunnel Encapsulation Attribute for which this document also defines new sub-TLVs.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 28, 2017.

Table of Contents

1. Introduction

Segment Routing (SR) technology leverages the source routing and tunneling paradigms. [I-D.ietf-spring-segment-routing] describes the SR architecture. [I-D.ietf-spring-segment-routing-mpls] describes its instantiation on the MPLS data plane and [I-D.ietf-6man-segment-routing-header] describes the Segment Routing instantiation over the IPv6 data plane.

This document defines a new BGP SAFI with a new NLRI in order to advertise a Segment Routing Policy (SR Policy) into BGP.

While for commodity we often write that BGP advertises an SR Policy, the reader should remember that BGP advertises a path of an SR policy and that this SR Policy might have several other candidate paths provided via BGP, PCEP, NETCONF or local policy configuration.

The BGP behavior described in this document is only focused on the signaling of a candidate path to a head-end.

The rules to select the best candidate path, to install it in the forwarding plane and to steer traffic on this policy are defined in [I-D.filsfils-spring-segment-routing-policy].

An SR Policy is advertised in the Border Gateway Protocol (BGP) by the BGP speaker being a router or a controller and using extensions defined in this document. Among the information encoded in the BGP message and representing the SR Policy, the steering mechanism is defined in [I-D.filsfils-spring-segment-routing-policy]. This steering mechanism makes also use of the Extended Color Community currently defined in [I-D.ietf-idr-tunnel-encaps].

Typically, a controller defines the set of policies and advertise them to BGP routers (typically ingress routers). The policy advertisement uses BGP extensions defined in this document. The policy advertisement is, in most but not all of the cases, tailored

for the receiver. In other words, a policy advertised to a given BGP speaker has significance only for that particular router and is not intended to be propagated anywhere else. Then, the receiver of the policy instantiate the policy in its routing and forwarding tables and steer traffic into it based on both the policy and destination prefix color and next-hop.

Alternatively, a router (i.e.: an BGP egress router) advertises SR Policies representing paths to itself. These advertisements are sent to SR policy head-end nodes who instantiate these policies and steer traffic into them according to the color and endpoint/BGP next-hop of both the policy and the destination prefix.

An SR Policy intended only for the receiver will, in most cases, not traverse any Route Reflector (RR, [RFC4456]).

However, there are cases where a SR Policy is intended for multiple receivers. Also, in a deployment scenario, a controller may also rely on the standard BGP update propagation scheme which makes use of route reflectors. These cases require mechanisms that:

- Uniquely identify each SR path of a given policy.
- Uniquely identify the intended receiver of a given SR Policy advertisement.

The BGP extensions for the advertisement of SR Policies include following components:

- A new Subsequent Address Family Identifier (SAFI) identifying the content of the BGP message (i.e.: the SR Policy).
- A new NLRI identifying the SR Policy.
- A set of new TLVs to be inserted into the Tunnel Encapsulation Attribute (as defined in [I-D.ietf-idr-tunnel-encaps]) and describing the SR Policy.
- An IPv4 address format route-target extended community ([RFC4360]) attached to the SR Policy advertisement and that indicates the intended receiver of such SR Policy advertisement.

- The Extended Color Community (as defined in [I-D.ietf-idr-tunnel-encaps]) and used in order to steer traffic into an SR Policy. This document (Section 3) modifies the format of the Extended Color Community by using the two leftmost bits of the RESERVED field.

1.1. Requirements Language

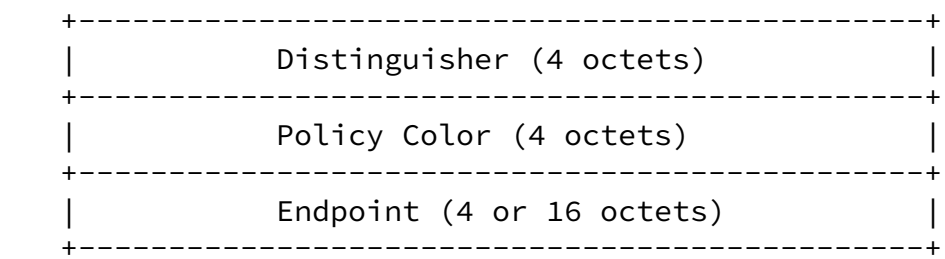
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. SR TE Policy Encoding

2.1. SR TE Policy SAFI and NLRI

A new SAFI is defined: the SR Policy SAFI, (codepoint 73 assigned by IANA (see Section 7) from the "Subsequent Address Family Identifiers (SAFI) Parameters" registry).

The SR Policy SAFI uses a new NLRI defined as follows:



where:

- Distinguisher: 4-octet value uniquely identifying the policy in the context of <color, endpoint> tuple. The distinguisher has no semantic and it's solely used by the SR Policy originator in order to make unique (from a NLRI perspective) multiple occurrences of the same SR Policy.

- Policy Color: 4-octet value identifying (with the endpoint) the policy. The color is used to match the color of the destination prefixes in order to steer traffic into the SR Policy [I-D.filsfils-spring-segment-routing-policy].

- Endpoint: identifies the endpoint of a policy. The Endpoint may represent a single node or a set of nodes (e.g.: an anycast address or a summary address). The Endpoint is an IPv4 (4-octet) address or an IPv6 (16-octet) address according to the AFI of the NLRI.

The NLRI containing the SR Policy is carried in a BGP UPDATE message [RFC4271] using BGP multiprotocol extensions [RFC4760] with an AFI of 1 or 2 (IPv4 or IPv6) and with a SAFI of TBD1 (to be assigned by IANA from the "Subsequent Address Family Identifiers (SAFI) Parameters" registry).

An update message that carries the MP_REACH_NLRI or MP_UNREACH_NLRI attribute with the SR Policy SAFI MUST also carry the BGP mandatory attributes. In addition, the BGP update message MAY also contain any of the BGP optional attributes.

The next-hop of the SR Policy SAFI NLRI is set based on the AFI. For example, if the AFI is set to IPv4 (1), then the next-hop is encoded as a 4-byte IPv4 address. If the AFI is set to IPv6 (2), then the next-hop is encoded as a 16-byte IPv6 address of the router. It is important to note that any BGP speaker receiving a BGP message with an SR Policy NLRI, will process it only if the NLRI is a best path as per the BGP best path selection algorithm.

It has to be noted that if several candidate paths of the same SR Policy (endpoint, color) are signaled via BGP to a head-end, we recommend that each NLRI use a different RD. Doing so, BGP passes all the paths to the SR Policy. The selection among all the candidate paths is best done by the SR Policy (BGP is only a conveyor of path, like PCEP, NETCONF or local CLI).

2.2. SR TE Policy and Tunnel Encapsulation Attribute

The content of the SR Policy is encoded in the Tunnel Encapsulation Attribute originally defined in [I-D.ietf-idr-tunnel-encaps] using a new Tunnel-Type TLV (codepoint is 15, assigned by IANA (see Section 7) from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).

The SR Policy Encoding structure is as follows:

```
SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>
Attributes:
   Tunnel Encaps Attribute (23)
      Tunnel Type: SR Policy
          Binding SID
          Preference
          Segment List
              Weight
              Segment
              Segment
              ...
          ...
```

where:

- SR Policy SAFI NLRI is defined in Section 2.1.
- Tunnel Encapsulation Attribute is defined in [I-D.ietf-idr-tunnel-encaps].
- Tunnel-Type is set to TBD2 (to be assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).
- Preference, Binding SID, Segment-List, Weight and Segment are defined in this document.
- Additional sub-TLVs may be defined in the future.

A single occurrence of "Tunnel Type: SR Policy" MUST be encoded within the same Tunnel Encapsulation Attribute.

Multiple occurrences of "Segment List" MAY be encoded within the same SR Policy.

Multiple occurrences of "Segment" MAY be encoded within the same Segment List.

2.3. Remote Endpoint and Color

The Remote Endpoint and Color sub-TLVs, as defined in [I-D.ietf-idr-tunnel-encaps], MAY also be present in the SR Policy encodings.

If present, the Remote Endpoint sub-TLV MUST match the Endpoint of the SR Policy SAFI NLRI.

If present, the Color sub-TLV MUST match the Policy Color of the SR Policy SAFI NLRI.

2.4. SR TE Policy Sub-TLVs

This section defines the SR Policy sub-TLVs.

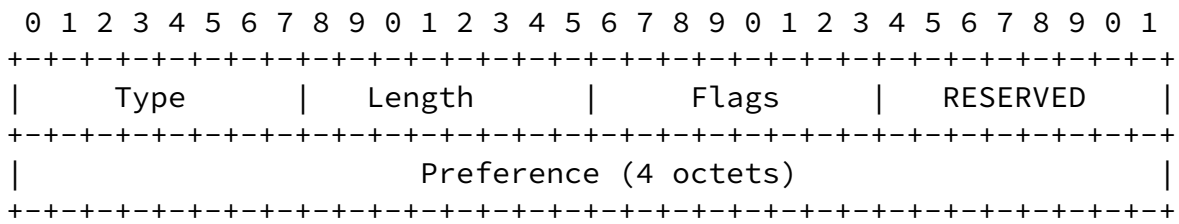
Preference, Binding SID, Segment-List are assigned from the "BGP Tunnel Encapsulation Attribute sub-TLVs" registry.

Weight and Segment Sub-TLVs are assigned from a new registry defined in this document and called: "SR Policy List Sub-TLVs". See Section 7 for the details of the registry.

2.4.1. Preference sub-TLV

The Preference sub-TLV is used in order to select the best path among a given SR Policy. This selection of the best path among the candidate paths of the SR policy is not done by BGP. BGP is only a conveyor of paths to the SR Policy. Other paths can be provided via NETCONF, PCEP or local CLI. The selection of the best path of an SR policy among its candidate paths is defined in [I-D.filsfils-spring-segment-routing-policy].

The Preference sub-TLV is optional, MAY appear only once in the SR Policy and has following format:



where:

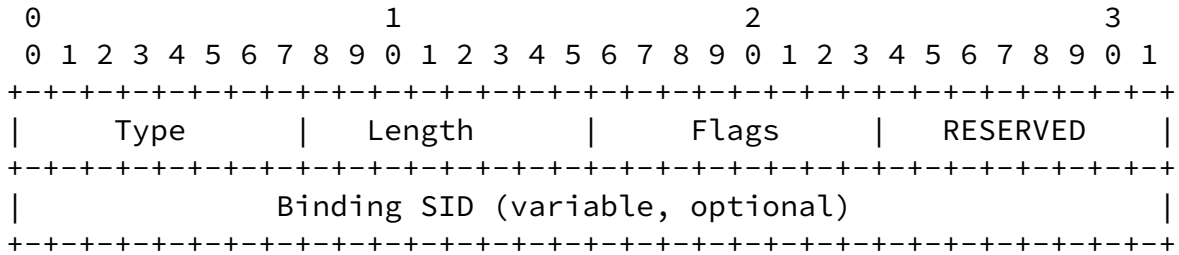
- Type: TBD3 (to be assigned by IANA from the "BGP Tunnel Encapsulation Attribute sub-TLVs" registry).
- Length: 6.
- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- Preference: a 4-octet value. The highest value is preferred.

The Preference is used by the the receiver in order to apply a selection rule among different SR paths of the same SR Policy. SR Policies may be originated and advertised through multiple means and protocols (not limited to BGP) therefore, the preference value is opaque to BGP and MUST NOT influence in any way the selection or the propagation of the BGP update. [I-D.filsfils-spring-segment-routing-policy] defines the use of the Preference value.

2.4.2. SR TE Binding SID Sub-TLV

The Binding SID sub-TLV specifies the BSID of the path.

The Binding SID sub-TLV is optional, MAY appear only once in the SR Policy and has the following format:



where:

- Type: TBD4 (to be assigned by IANA from the "BGP Tunnel Encapsulation Attribute sub-TLVs" registry).

- Length: specifies the length of the value field not including Type and Length fields. Can be 2 or 6 or 18.

- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.

- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

- Binding SID: if length is 2, then no Binding SID is present. If length is 6 then the Binding SID contains a 4-octet SID. If length is 18 then the Binding SID contains a 16-octet IPv6 SID.

The Binding SID sub-TLV specifies the BSID of the path.

When a controller is used in order to define and advertise SR Policies and when the Binding SID is assigned by the receiver, such Binding SID SHOULD be reported to the controller. The mechanisms and/or APIs used for the reporting of the Binding SID are outside the scope of this document.

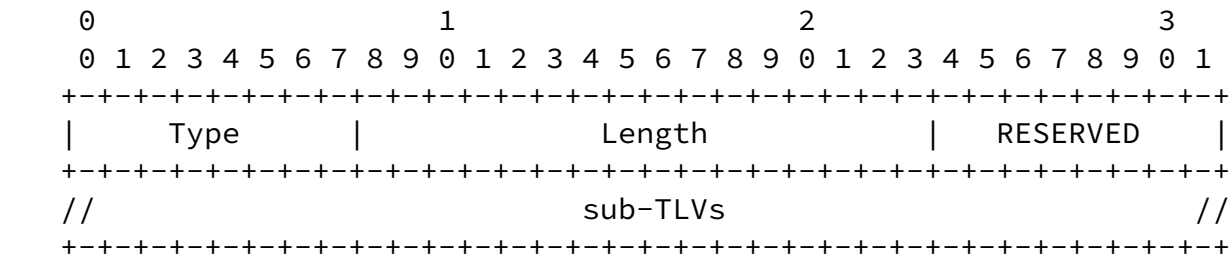
The Binding SID concept is defined in [I-D.ietf-spring-segment-routing] and its use in the context of SR Policies is defined in [I-D.filsfils-spring-segment-routing-policy].

2.4.3. Segment List Sub-TLV

The Segment List sub-TLV is used in order to encode a single explicit path towards the endpoint. The Segment List sub-TLV includes the elements of the paths (i.e.: segments) as well as an optional Weight TLV.

The Segment List sub-TLV may exceed 255 bytes length due to large number of segments. Therefore a 2-octet length is required. According to [I-D.ietf-idr-tunnel-encaps], the first bit of the sub-TLV codepoint defines the size of the length field. Therefore, for the Segment List sub-TLV a code point of 128 (or higher) is used. See Section 7 section for details of codepoints allocation.

The Segment List sub-TLV is mandatory, MAY appear multiple times in the SR Policy and has the following format:



where:

- Type: TBD5 (to be assigned by IANA from the "BGP Tunnel Encapsulation Attribute sub-TLVs" registry).
- Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.
- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- sub-TLVs:
  - An optional single Weight sub-TLV.
  - One or more Segment sub-TLVs.

The Segment List sub-TLV is mandatory.

Multiple occurrences of the Segment List sub-TLV MAY appear in the SR Policy.

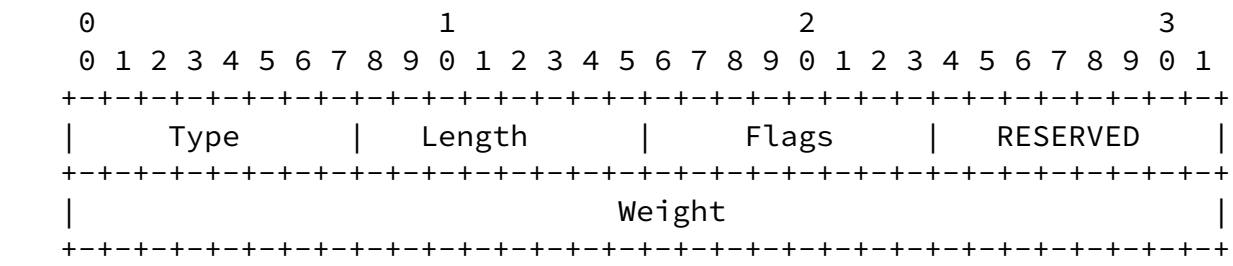
When multiple occurrences of the Segment List sub-TLV appear in the SR Policy, the traffic is load-balanced across them either through an ECMP scheme (if no Weight sub-TLV is present) or through a weighted ECMP scheme according to Section 2.4.3.1.

The Segment-List Sub-TLV MUST contain at least one Segment Sub-TLV and MAY contain a Weight Sub-TLV.

2.4.3.1. Weight Sub-TLV

The Weight sub-TLV specifies the weight associated to a given path (i.e.: a given segment list). The weight is used in order to apply weighted ECMP mechanism when steering traffic into a policy that includes multiple Segment Lists sub-TLVs (i.e.: multiple explicit paths). The use of the weight for ECMP purposes is described in [I-D.filsfils-spring-segment-routing-policy].

The Weight sub-TLV is optional, MAY only appear once inside the Segment List sub-TLV, and has the following format:



where:

Type: 9 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).

Length: 6.

Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.

RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

The use of the Weight sub-TLV is specified in [I-D.filsfils-spring-segment-routing-policy]. It is important to note that the Weight has no meaning for the BGP speaker and MUST be considered as an opaque information.

2.4.3.2. Segment Sub-TLV

The Segment sub-TLV describes a single segment in a segment list

(i.e.: a single element of the explicit path). Multiple Segment sub-TLVs constitute an explicit path of the SR Policy.

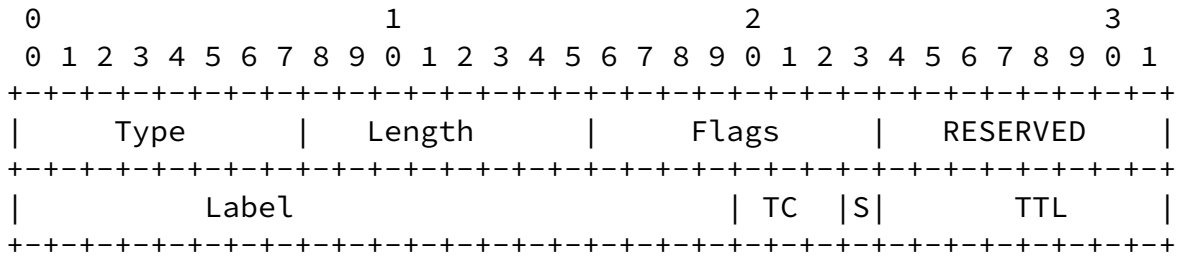
The Segment sub-TLV is mandatory and MAY appear multiple times in the Segment List sub-TLV.

[I-D.filsfils-spring-segment-routing-policy] defines several types of Segment Sub-TLVs:

Type 1: SID only, in the form of MPLS Label
Type 2: SID only, in the form of IPv6 address
Type 3: IPv4 Node Address with optional SID
Type 4: IPv6 Node Address with optional SID
Type 5: IPv4 Address + index with optional SID
Type 6: IPv4 Local and Remote addresses with optional SID
Type 7: IPv6 Address + index with optional SID
Type 8: IPv6 Local and Remote addresses with optional SID

2.4.3.2.1. Type 1: SID only, in the form of MPLS Label

The Type-1 Segment Sub-TLV encodes a single SID in the form of an MPLS label. The format is as follows:



where:

o Type: 1 (to be assigned by IANA from the registry "SR Policy List

Sub-TLVs" defined in this document).

- Length is 6.
- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- Label: 20 bits of label value.
- TC: 3 bits of traffic class.
- S: 1 bit of bottom-of-stack.
- TTL: 1 octet of TTL.

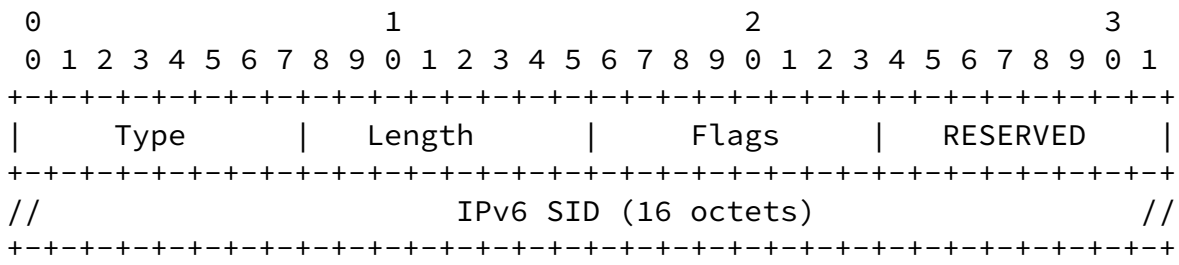
The following applies to the Type-1 Segment sub-TLV:

- The S bit SHOULD be zero upon transmission, and MUST be ignored upon reception.

- If the originator wants the receiver to choose the TC value, it sets the TC field to zero.
- If the originator wants the receiver to choose the TTL value, it sets the TTL field to 255.
- If the originator wants to recommend a value for these fields, it puts those values in the TC and/or TTL fields.
- The receiver MAY override the originator's values for these fields. This would be determined by local policy at the receiver. One possible policy would be to override the fields only if the fields have the default values specified above.

2.4.3.2.2. Type 2: SID only, in the form of IPv6 address

The Type-2 Segment Sub-TLV encodes a single SID in the form of an IPv6 SID. The format is as follows:



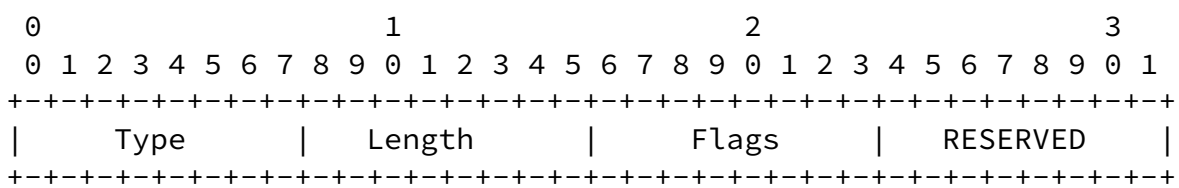
where:

- Type: 2 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- Length is 18.
- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- IPv6 SID: 16 octets of IPv6 address.

The IPv6 Segment Identifier (IPv6 SID) is defined in [I-D.ietf-6man-segment-routing-header].

2.4.3.2.3. Type 3: IPv4 Node Address with optional SID

The Type-3 Segment Sub-TLV encodes an IPv4 node address and an optional SID in the form of either an MPLS label or an IPv6 address. The format is as follows:



```
|                  IPv4 Node Address (4 octets)                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//                  SID (optional, 4 or 16 octets)               //
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

where:

- Type: 3 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- Length is 6 or 10 or 22.
- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- IPv4 Node Address: a 4 octet IPv4 address representing a node.
- SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

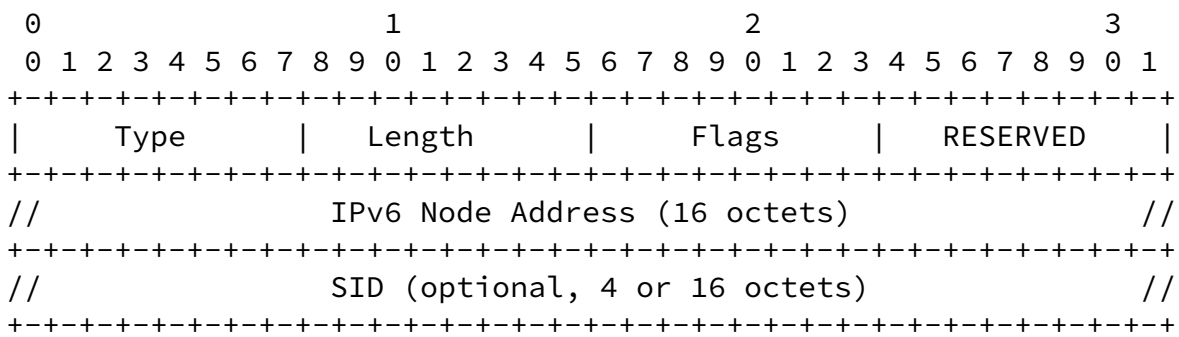
The following applies to the Type-3 Segment sub-TLV:

- The IPv4 Node Address MUST be present.
- The SID is optional and MAY be of one of the following formats:
  - MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.
  - IPV6 SID: a 16 octet IPv6 address.
- If length is 6, then only the IPv4 Node Address is present.
- If length is 10, then the IPv4 Node Address and the MPLS SID are present.

- If length is 22, then the IPv4 Node Address and the IPv6 SID are present.

2.4.3.2.4. Type 4: IPv6 Node Address with optional SID

The Type-4 Segment Sub-TLV encodes an IPv6 node address and an optional SID in the form of either an MPLS label or an IPv6 address. The format is as follows:



where:

- Type: 4 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).

- Length is 18 or 22 or 34.

- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.

- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

- IPv6 Node Address: a 16 octet IPv6 address representing a node.

- SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

The following applies to the Type-4 Segment sub-TLV:

- The IPv6 Node Address MUST be present.

- The SID is optional and MAY be of one of the following formats:

  * MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.

  * IPV6 SID: a 16 octet IPv6 address.

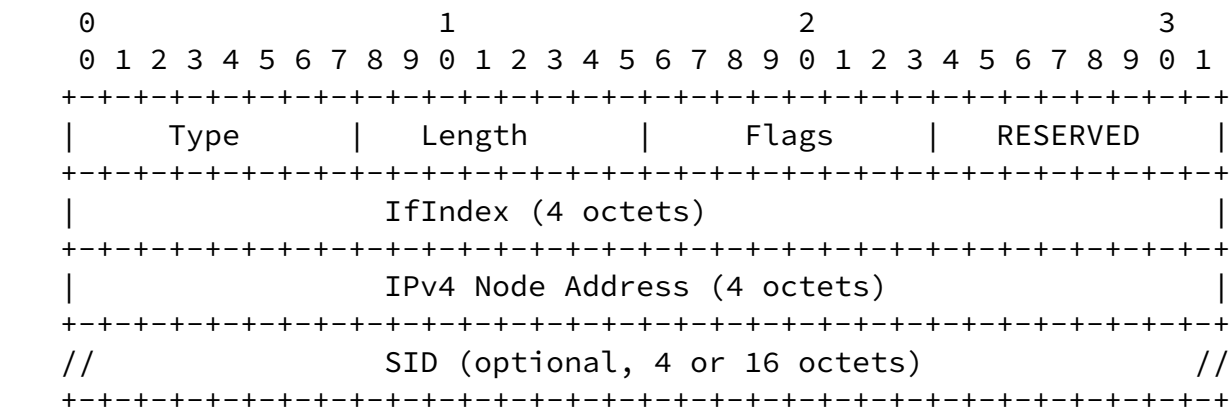
- If length is 18, then only the IPv6 Node Address is present.

- If length is 22, then the IPv6 Node Address and the MPLS SID are present.

- If length is 34, then the IPv6 Node Address and the IPv6 SID are present.

2.4.3.2.5. Type 5: IPv4 Address + index with optional SID

The Type-5 Segment Sub-TLV encodes an IPv4 node address, an interface index (IfIndex) and an optional SID in the form of either an MPLS label or an IPv6 address. The format is as follows:



where:

- Type: 5 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).

- Length is 10 or 14 or 26.

- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.

- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

- IfIndex: 4 octets of interface index.

- IPv4 Node Address: a 4 octet IPv4 address representing a node.

- SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

The following applies to the Type-5 Segment sub-TLV:

- The IPv4 Node Address MUST be present.

o The SID is optional and MAY be of one of the following formats:

  * MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.

  * IPV6 SID: a 16 octet IPv6 SID.

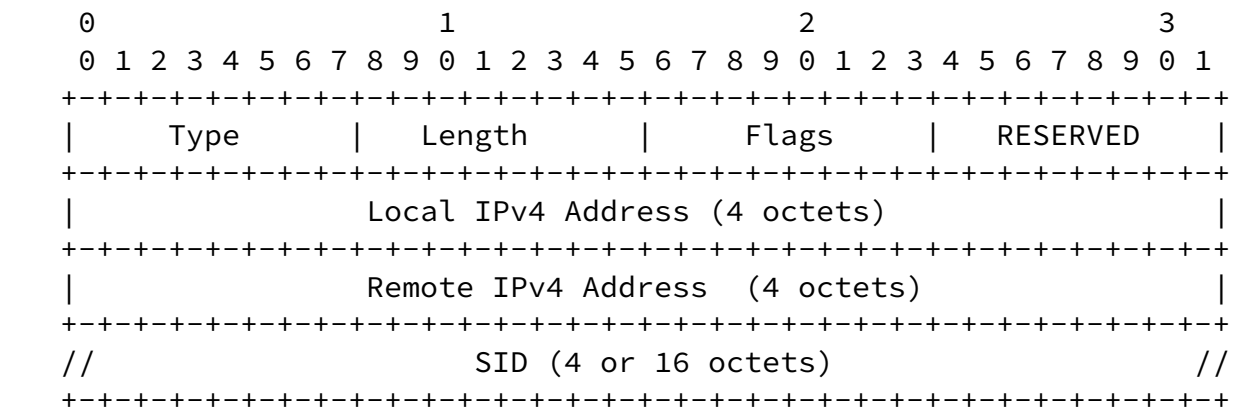
o If length is 10, then the IPv4 Node Address and IfIndex are present.

o If length is 14, then the IPv4 Node Address, the IfIndex and the MPLS SID are present.

o If length is 26, then the IPv4 Node Address, the IfIndex and the IPv6 SID are present.

2.4.3.2.6. Type 6: IPv4 Local and Remote addresses with optional SID

The Type-6 Segment Sub-TLV encodes an IPv4 node address, an adjacency local address, an adjacency remote address and an optional SID in the form of either an MPLS label or an IPv6 address. The format is as follows:



where:

o Type: 6 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).

- Length is 10 or 14 or 26.

- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.

- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

- Local IPv4 Address: a 4 octet IPv4 address.

- Remote IPv4 Address: a 4 octet IPv4 address.

- SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

The following applies to the Type-6 Segment sub-TLV:

- The Local IPv4 Address MUST be present and represents an adjacency local address.

- The Remote IPv4 Address MUST be present and represents the remote end of the adjacency.

- The SID is optional and MAY be of one of the following formats:

  * MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.

  * IPV6 SID: a 16 octet IPv6 address.

- If length is 10, then only the IPv4 Local and Remote addresses are present.

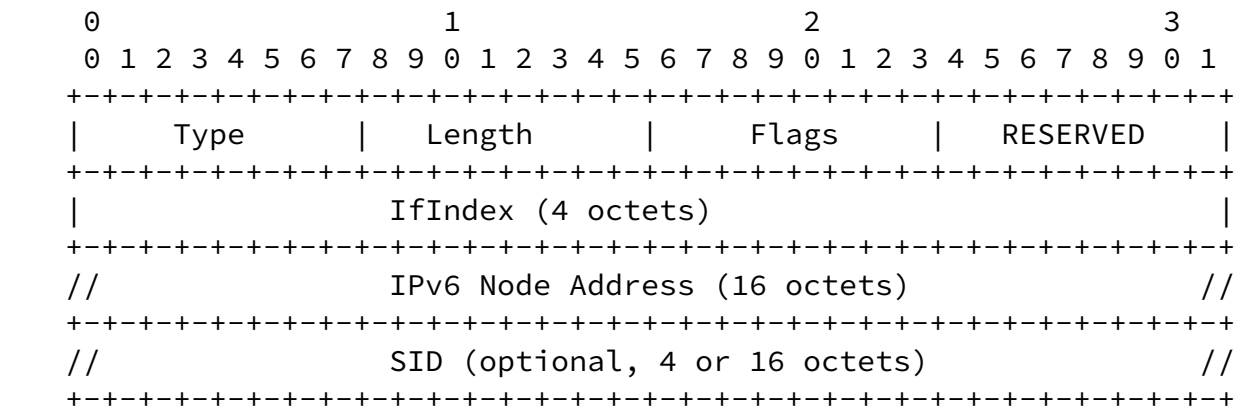
- If length is 14, then the IPv4 Local address, IPv4 Remote address and the MPLS SID are present.

- If length is 26, then the IPv4 Local address, IPv4 Remote address and the IPv6 SID are present.

2.4.3.2.7. Type 7: IPv6 Address + index with optional SID

The Type-7 Segment Sub-TLV encodes an IPv6 node address, an interface

index and an optional SID in the form of either an MPLS label or an IPv6 address. The format is as follows:



where:

- Type: 7 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).

- Length is 22 or 26 or 38.

- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.

- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

- IfIndex: 4 octets of interface index.

- IPv6 Node Address: a 16 octet IPv6 address representing a node.
- SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

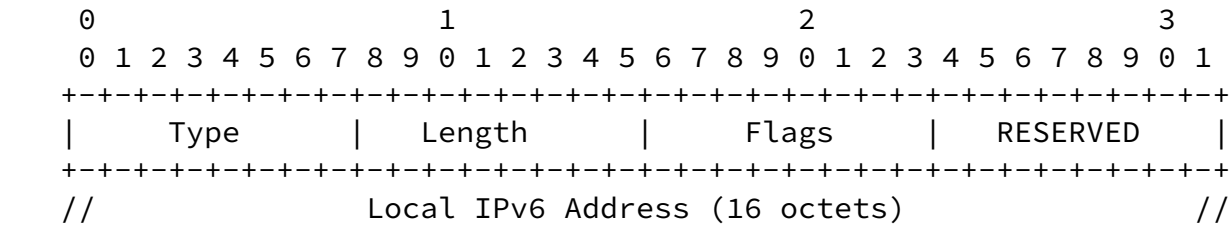
The following applies to the Type-7 Segment sub-TLV:

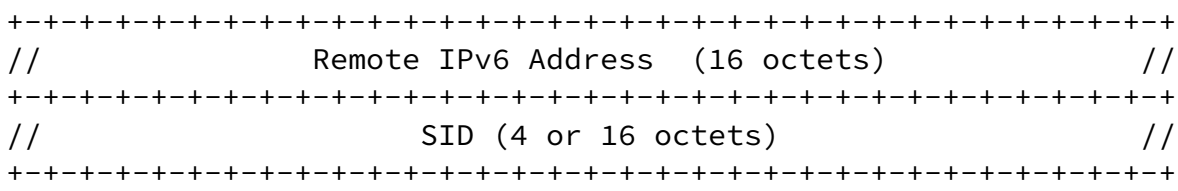
- The IPv6 Node Address MUST be present.
- The Interface Index MUST be present.
- The SID is optional and MAY be of one of the following formats:
  - MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.
  - IPV6 SID: a 16 octet IPv6 address.
- If length is 22, then the IPv6 Node Address and IfIndex are present.

- If length is 26, then the IPv6 Node Address, the IfIndex and the MPLS SID are present.
- If length is 38, then the IPv6 Node Address, the IfIndex and the IPv6 SID are present.

2.4.3.2.8. Type 8: IPv6 Local and Remote addresses with optional SID

The Type-8 Segment Sub-TLV encodes an IPv6 node address, an adjacency local address, an adjacency remote address and an optional SID in the form of either an MPLS label or an IPv6 address. The format is as follows:





where:

- Type: 8 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- Length is 34 or 38 or 50.
- Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- Local IPv6 Address: a 16 octet IPv6 address.
- Remote IPv6 Address: a 16 octet IPv6 address.
- SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

The following applies to the Type-8 Segment sub-TLV:

- The Local IPv6 Address MUST be present and represents an adjacency local address.
- The Remote IPv6 Address MUST be present and represents the remote end of the adjacency.
- The SID is optional and MAY be of one of the following formats:
  - MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.
  - IPV6 SID: a 16 octet IPv6 address.

- If length is 34, then only the IPv6 Local and Remote addresses are present.

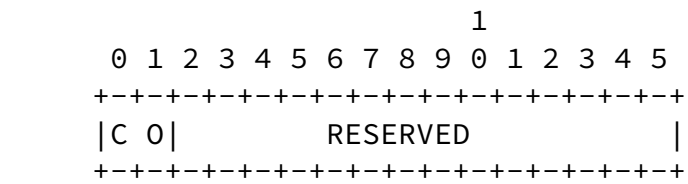
- If length is 38, then the IPv6 Local address, IPv4 Remote address and the MPLS SID are present.

- If length is 50, then the IPv6 Local address, IPv4 Remote address and the IPv6 SID are present.

3. Extended Color Community

The Extended Color Community as defined in [I-D.ietf-idr-tunnel-encaps] is used in order to steer traffic into a policy. This document applies the following changes to the Extended Color Community attribute.

The RESERVED field is changed as follows:



where CO bits are defined as the "Color-Only" bits. The settings and use of these bits are defined in Section 4.3.

4. SR Policy Operations

4.1. Configuration and Advertisement of SR TE Policies

Typically, but not limited to, a SR Policy is configured into a controller and on the base of each receiver. In other words, each SR Policy configured is related to the intended receiver. It is therefore normal for a given <color,endpoint> SR Policy to have

multiple SR paths with different content where each of these SR paths (of the same policy) is intended to be sent to different receivers.

When advertised in BGP, each SR path of the same SR Policy will have a different Distinguisher in order to prevent BGP selection among

these SR paths along the distribution of BGP updates.

Moreover, a Route-Target extended community SHOULD be attached to the SR Policy and that identifies the intended receiver of the advertisement.

If no route-target is attached to the SR Policy NLRI, then it is assumed that the originator sends the SR Policy update directly (e.g.: through iBGP multihop) to the intended receiver. In such case, the NO_ADVERTISE community MUST be attached to the SR Policy update.

If no route-target is attached to the SR Policy NLRI, then it is assumed that the originator sends the SR Policy update directly (e.g.: through iBGP multihop) to the intended receiver. In such case, the NO_ADVERTISE community MUST be attached to the SR Policy update.

4.2. Reception of an SR Policy

On reception of a SR Policy, a BGP speaker MUST determine if the SR Policy is first acceptable, then usable.

While only usable SR Policies are instantiated, acceptable SR Policies (i.e.: also the non-usable ones) MAY be propagated.

Any SR Policy update that has been determined acceptable is kept in the BGP database. This includes non-usable SR Policies.

4.2.1. Acceptance of a SR Policy Update

When a BGP speaker receives an SR Policy from a neighbor it has to determine if the SR Policy advertisement is acceptable. The following applies:

- The SR Policy NLRI MUST have a color value.
- The SR Policy NLRI MUST have either an IPv4 endpoint address or an IPv6 endpoint address or a zero-value (either IPv4 or IPv6 format).
- The SR Policy NLRI MUST have distinguisher field.

- The SR Policy update MUST have either the NO_ADV community or at least one route-target extended community in IPv4-address format.

- The Tunnel Encapsulation Attribute MUST be attached to the BGP Update and MUST have the Tunnel Type set to SR Policy (value to be assigned by IANA).

- Within the SR Policy, at least one Segment List sub-TLV MUST be present.

- Within the Segment List sub-TLV at least one Segment sub-TLV MUST be present.

The use of an endpoint address with a zero-value is described in Section 4.3.

The Remote Endpoint and Color sub-TLVs, as defined in [I-D.ietf-idr-tunnel-encaps], MAY also be present in the SR Policy encodings. If present, the Remote Endpoint sub-TLV MUST match the Endpoint of the SR Policy SAFI NLRI. If they don't match, the SR Policy advertisement MUST be considered as not acceptable. If present, the Color sub-TLV MUST match the Policy Color of the SR Policy SAFI NLRI. If they don't match, the SR Policy advertisement MUST be considered as not acceptable.

A non-acceptable SR Policy update that has a valid NLRI portion with invalid attribute portion MUST be considered as a withdraw of the SR Policy.

A non-acceptable SR Policy update that has an invalid NLRI portion MUST trigger a reset of the BGP session.

The receiver MUST check whether route-target or NO_ADVERTISE communities are attached to it. If no route-target is present and the NO_ADVERTISE community is present, then the SR Policy is usable.

If one or more route-targets are present, then at least one route-target MUST match the BGP Identifier (BGP Router-ID) of the receiver in order for the update to be considered usable. The BGP Identifier is defined in [RFC4271] as a 4 octet IPv4 address. Therefore the route-target extended community MUST be of the same format.

If one or more route-targets are present and no one matches the local BGP router-ID, then, while the SR Policy is acceptable, the SR Policy is not usable. It has to be noted that if the receiver has been explicitly configured to do so, it MAY propagate the SR Policy to its neighbors as defined in Section 4.2.3.

4.2.2. Passing an acceptable path to an SR Policy

Once BGP has determined that the path is acceptable, BGP passes the path to the SR Policy.

The SR Policy applies the rules defined in [I-D.filsfils-spring-segment-routing-policy] to determine whether a path is valid and to select the best path among the valid paths.

4.2.3. Propagation of an SR Policy

By default, a BGP node receiving an SR Policy MUST NOT not propagate it to any eBGP neighbor.

However, a node MAY be explicitly configured in order to advertise a received SR Policy update to neighbors according to normal BGP rules (iBGP and eBGP propagation), e.g., in the case the node is a Route-Reflector.

SR Policies that have been determined acceptable and valid can be propagated, even the ones that are not usable.

Only SR Policies that do not have the NO_ADVERTISE community attached to them can be propagated.

4.3. Steering Traffic into a SR Policy

The steering of a BGP route onto an SR Policy is defined in [I-D.filsfils-spring-segment-routing-policy].

4.4. Flowspec and SR Policies

The SR Policy can be carried in context of a Flowspec NLRI ([RFC5575]). In this case, when the redirect to IP next-hop is specified as in [I-D.ietf-idr-flowspec-redirect-ip], the tunnel to the next-hop is specified by the segment list in the Segment List sub-TLVs. The Segment List (e.g..: label stack or IPv6 segment list) is imposed to flows matching the criteria in the Flowspec route in order to steer them towards the next-hop as specified in the SR Policy SAFI NLRI.

5. Acknowledgments

The authors of this document would like to thank Dhanendra Jain, Shyam Sethuram, Acee Lindem, Imtiyaz Mohammad and John Scudder for their comments and review of this document.

6. Implementation Status

Note to RFC Editor: Please remove this section prior to publication, as well as the reference to RFC 7942.

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in [RFC7942]. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist.

According to [RFC7942], "this will allow reviewers and working groups to assign due consideration to documents that have the benefit of running code, which may serve as evidence of valuable experimentation and feedback that have made the implemented protocols more mature. It is up to the individual working groups to use this information as they see fit".

Several early implementations exist and will be reported in detail in a forthcoming version of this document. For purposes of early interoperability testing, when no FCFS code point was available, implementations have made use of the following values:

- Preference sub-TLV: 6
- Binding SID sub-TLV: 7
- Segment List sub-TLV: 128

When IANA-assigned values are available, implementations will be updated to use them.

7. IANA Considerations

This document defines new Sub-TLVs in following existing registries:

- Subsequent Address Family Identifiers (SAFI) Parameters
- BGP Tunnel Encapsulation Attribute Tunnel Types
- BGP Tunnel Encapsulation Attribute sub-TLVs

This document also defines a new registry: "SR Policy List Sub-TLVs".

7.1. Existing Registry: Subsequent Address Family Identifiers (SAFI) Parameters

This document defines a new SAFI in the registry "Subsequent Address Family Identifiers (SAFI) Parameters" that has been assigned by IANA:

Codepoint	Description	Reference
73	SR Policy SAFI	This document

7.2. Existing Registry: BGP Tunnel Encapsulation Attribute Tunnel Types

This document defines a new Tunnel-Type in the registry "BGP Tunnel Encapsulation Attribute Tunnel Types" that has been assigned by IANA:

Codepoint	Description	Reference
15	SR Policy Type	This document

7.3. Existing Registry: BGP Tunnel Encapsulation Attribute sub-TLVs

This document defines new sub-TLVs in the registry "BGP Tunnel Encapsulation Attribute sub-TLVs" to be assigned by IANA:

Codepoint	Description	Reference

TBD3	Preference sub-TLV	This document
TBD4	Binding SID sub-TLV	This document
TBD5	Segment List sub-TLV	This document

7.4. New Registry: SR Policy List Sub-TLVs

This document defines a new registry called "SR Policy List Sub-TLVs". The allocation policy of this registry is "First Come First Served (FCFS)" according to [RFC5226].

Following Sub-TLV codepoints are defined:

Value	Description	Reference
1	MPLS SID sub-TLV	This document
2	IPv6 SID sub-TLV	This document
3	IPv4 Node and SID sub-TLV	This document
4	IPv6 Node and SID sub-TLV	This document
5	IPv4 Node, index and SID sub-TLV	This document
6	IPv4 Local/Remote addresses and SID sub-TLV	This document
7	IPv6 Node, index and SID sub-TLV	This document
8	IPv6 Local/Remote addresses and SID sub-TLV	This document
9	Weight sub-TLV	This document

8. Security Considerations

TBD.

9. References

9.1. Normative References

[I-D.ietf-idr-tunnel-encaps]

Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-03 (work in progress), November 2016.

[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>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <http://www.rfc-editor.org/info/rfc4360>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <http://www.rfc-editor.org/info/rfc4760>.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <http://www.rfc-editor.org/info/rfc5226>.

[RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, DOI 10.17487/RFC5575, August 2009, <http://www.rfc-editor.org/info/rfc5575>.

9.2. Informational References

[I-D.filsfils-spring-segment-routing-policy] Filsfils, C., Sivabalan, S., Yoyer, D., Nanduri, M., Lin, S., bogdanov@google.com, b., Horneffer, M., Clad, F., Steinberg, D., Decraene, B., and S. Litkowski, "Segment Routing Policy for Traffic Engineering", draft-filsfils-spring-segment-routing-policy-00 (work in progress), February 2017.

[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-idr-flowspec-redirect-ip]
Uttaro, J., Haas, J., Texier, M., Andy, A., Ray, S., Simpson, A., and W. Henderickx, "BGP Flow-Spec Redirect to IP Action", draft-ietf-idr-flowspec-redirect-ip-02 (work in progress), February 2015.

[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-11 (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.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <http://www.rfc-editor.org/info/rfc4456>.

[RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <http://www.rfc-editor.org/info/rfc7942>.

Authors' Addresses

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

Email: sprevidi@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Brussels
BE

Email: cfilsfil@cisco.com

Arjun Sreekantiah
Cisco Systems, Inc.
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: asreekan@cisco.com

Siva Sivabalan
Cisco Systems, Inc.
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: msiva@cisco.com

Paul Mattes
Microsoft
One Microsoft Way
Redmond, WA 98052
USA

Email: pamattes@microsoft.com

Eric Rosen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
US

Email: erosen@juniper.net

Steven Lin
Google

Email: stevenlin@google.com