

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2017

S. Previdi, Ed.
P. Psenak
C. Filsfils
Cisco Systems, Inc.
H. Gredler
RtBrick Inc.
M. Chen
Huawei Technologies
J. Tantsura
Individual
October 30, 2016

BGP Link-State extensions for Segment Routing
draft-gredler-idr-bgp-ls-segment-routing-ext-04

Abstract

Segment Routing (SR) allows for a flexible definition of end-to-end paths within IGP topologies by encoding paths as sequences of topological sub-paths, called "segments". These segments are advertised by the link-state routing protocols (IS-IS, OSPF and OSPFv3).

This draft defines extensions to the BGP Link-state address-family in order to carry segment information via BGP.

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.

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 May 3, 2017.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. BGP-LS Extensions for Segment Routing	5
2.1. Node Attributes TLVs	5
2.1.1. SR-Capabilities TLV	5
2.1.2. SR-Algorithm TLV	6
2.1.3. SR Local Block TLV	7
2.1.4. SRMS Preference TLV	7
2.2. Link Attribute TLVs	8
2.2.1. Adjacency SID TLV	9
2.2.2. LAN Adjacency SID TLV	9
2.3. Prefix Attribute TLVs	10
2.3.1. Prefix-SID TLV	11
2.3.2. IPv6 Prefix-SID TLV	12
2.3.3. IGP Prefix Attributes TLV	13
2.3.4. Source Router Identifier (Source Router-ID) TLV	14
2.3.5. Range TLV	14
2.3.6. Binding SID TLV	15
2.3.7. Binding SID SubTLVs	16
2.4. Equivalent IS-IS Segment Routing TLVs/Sub-TLVs	22
2.5. Equivalent OSPF/OSPFv3 Segment Routing TLVs/Sub-TLVs	23
3. Procedures	25
3.1. Advertisement of a IS-IS Prefix SID TLV	25
3.2. Advertisement of a OSPF/OSPFv3 Prefix-SID TLV	25
3.3. Advertisement of a range of prefix-to-SID mappings in OSPF	26
3.4. Advertisement of a range of IS-IS SR bindings	26
3.5. Advertisement of a path and its attributes from IS-IS protocol	26
3.6. Advertisement of a path and its attributes from	

OSPFv2/OSPFv3 protocol	27
4. IANA Considerations	27
4.1. TLV/Sub-TLV Code Points Summary	27
5. Manageability Considerations	28
5.1. Operational Considerations	28
5.1.1. Operations	28
6. Security Considerations	29
7. Contributors	29
8. Acknowledgements	29
9. References	29
9.1. Normative References	29
9.2. Informative References	30
9.3. URIs	31
Authors' Addresses	34

1. Introduction

Segment Routing (SR) allows for a flexible definition of end-to-end paths by combining sub-paths called "segments". A segment can represent any instruction, topological or service-based. A segment can have a local semantic to an SR node or global within a domain. Within IGP topologies an SR path is encoded as a sequence of topological sub-paths, called "IGP segments". These segments are advertised by the link-state routing protocols (IS-IS, OSPF and OSPFv3).

Two types of IGP segments are defined, Prefix segments and Adjacency segments. Prefix segments, by default, represent an ECMP-aware shortest-path to a prefix, as per the state of the IGP topology. Adjacency segments represent a hop over a specific adjacency between two nodes in the IGP. A prefix segment is typically a multi-hop path while an adjacency segment, in most of the cases, is a one-hop path. [I-D.ietf-spring-segment-routing].

When Segment Routing is enabled in a IGP domain, segments are advertised in the form of Segment Identifiers (SIDs). The IGP link-state routing protocols have been extended to advertise SIDs and other SR-related information. IGP extensions are described in: IS-IS [I-D.ietf-isis-segment-routing-extensions], OSPFv2 [I-D.ietf-ospf-segment-routing-extensions] and OSPFv3 [I-D.ietf-ospf-ospfv3-segment-routing-extensions]. Using these extensions, Segment Routing can be enabled within an IGP domain.

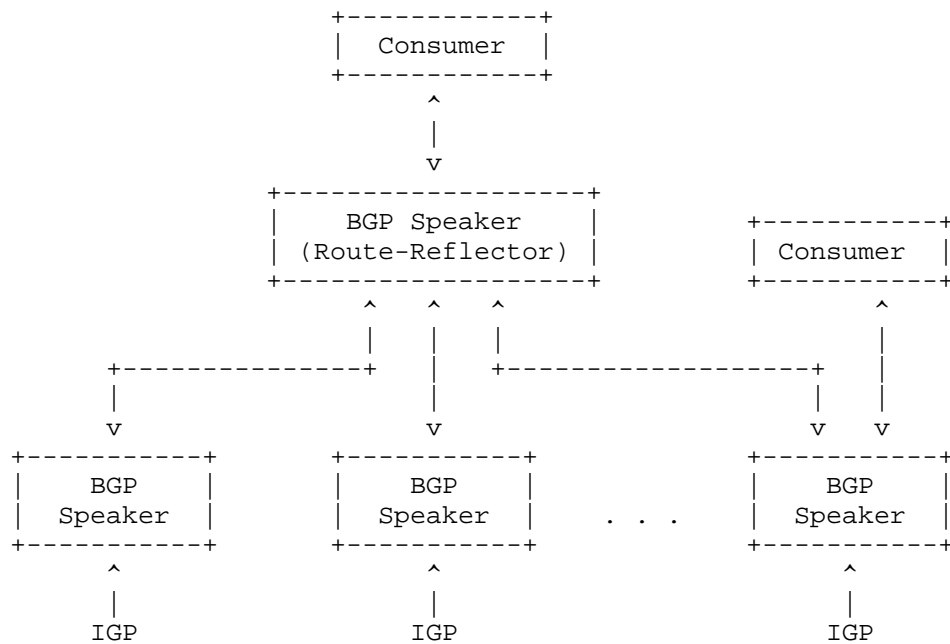


Figure 1: Link State info collection

Segment Routing (SR) allows advertisement of single or multi-hop paths. The flooding scope for the IGP extensions for Segment routing is IGP area-wide. Consequently, the contents of a Link State Database (LSDB) or a Traffic Engineering Database (TED) has the scope of an IGP area and therefore, by using the IGP alone it is not enough to construct segments across multiple IGP Area or AS boundaries.

In order to address the need for applications that require topological visibility across IGP areas, or even across Autonomous Systems (AS), the BGP-LS address-family/sub-address-family have been defined to allow BGP to carry Link-State information. The BGP Network Layer Reachability Information (NLRI) encoding format for BGP-LS and a new BGP Path Attribute called the BGP-LS attribute are defined in [RFC7752]. The identifying key of each Link-State object, namely a node, link, or prefix, is encoded in the NLRI and the properties of the object are encoded in the BGP-LS attribute. Figure Figure 1 describes a typical deployment scenario. In each IGP area, one or more nodes are configured with BGP-LS. These BGP speakers form an IBGP mesh by connecting to one or more route-reflectors. This way, all BGP speakers (specifically the route-reflectors) obtain Link-State information from all IGP areas (and from other ASes from EBGP peers). An external component connects to the route-reflector to obtain this information (perhaps moderated by

a policy regarding what information is or isn't advertised to the external component).

This document describes extensions to BGP-LS to advertise the SR information. An external component (e.g., a controller) then can collect SR information in the "northbound" direction across IGP areas or ASes and construct the end-to-end path (with its associated SIDs) that need to be applied to an incoming packet to achieve the desired end-to-end forwarding.

2. BGP-LS Extensions for Segment Routing

This document defines IGP SR extensions BGP-LS TLVs and Sub-TLVs. Section 2.4 and Section 2.5 illustrates the equivalent TLVs and Sub-TLVs in IS-IS, OSPF and OSPFv3 protocols.

BGP-LS [RFC7752] defines the BGP-LS NLRI that can be a Node NLRI, a Link NLRI or a Prefix NLRI. The corresponding BGP-LS attribute is a Node Attribute, a Link Attribute or a Prefix Attribute. BGP-LS [RFC7752] defines the TLVs that map link-state information to BGP-LS NLRI and the BGP-LS attribute. This document adds additional BGP-LS attribute TLVs in order to encode SR information.

2.1. Node Attributes TLVs

The following Node Attribute TLVs are defined:

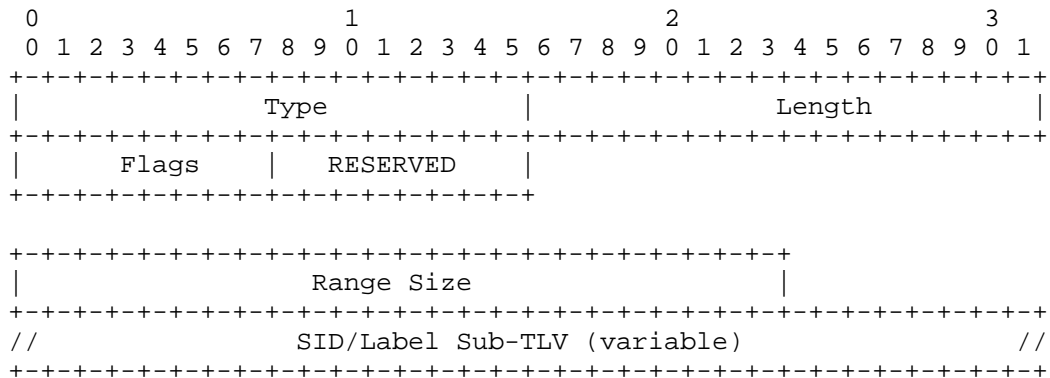
TLV Code Point	Description	Length	Section
1034	SR Capabilities	variable	Section 2.1.1
1035	SR Algorithm	variable	Section 2.1.2
1036	SR Local Block	variable	Section 2.1.3
1037	SRMS Preference	variable	Section 2.1.4

Table 1: Node Attribute TLVs

These TLVs can ONLY be added to the Node Attribute associated with the Node NLRI that originates the corresponding SR TLV.

2.1.1. SR-Capabilities TLV

The SR Capabilities sub-TLV has following format:



Type: TBD, suggested value 1034.

Length: Variable.

Flags: 1 octet of flags as defined in
 [I-D.ietf-isis-segment-routing-extensions] and
 [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

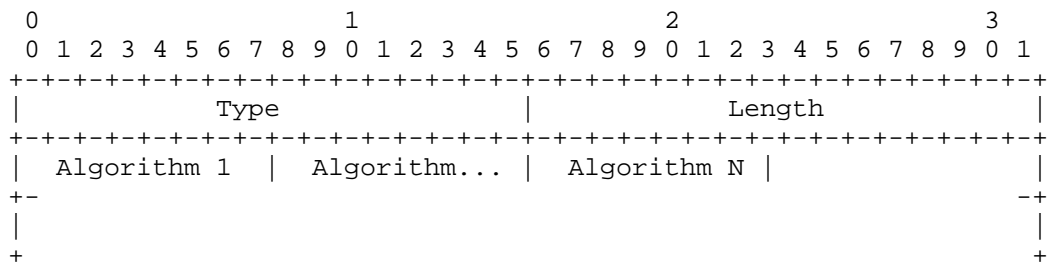
One or more entries, each of which have the following format:

Range Size: 3 octet value indicating the number of labels in
 the range.

SID/Label sub-TLV (as defined in Section 2.3.7.2).

2.1.1.2. SR-Algorithm TLV

The SR-Algorithm TLV has the following format:



where:

Type: TBD, suggested value 1035.

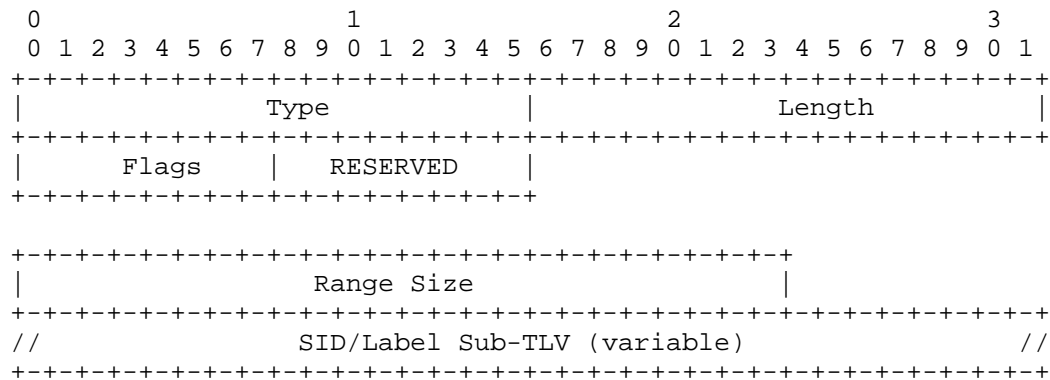
Length: Variable.

Algorithm: 1 octet identifying the algorithm.

2.1.3. SR Local Block TLV

The SR Local Block (SRLB) Sub-TLV contains the range of labels the node has reserved for local SIDs. Local SIDs are used, e.g., in IGP (IS-IS, OSPF) for Adjacency-SIDs, and may also be allocated by other components than IGP protocols. As an example, an application or a controller may instruct a node to allocate a specific local SID. Therefore, in order for such applications or controllers to know the range of local SIDs available, it is required that the node advertises its SRLB.

The SRLB TLV has the following format:



Type: TBD, suggested value 1036.

Length: Variable.

Flags: 1 octet of flags. None are defined at this stage.

One or more entries, each of which have the following format:

Range Size: 3 octet value indicating the number of labels in the range.

SID/Label sub-TLV (as defined in Section 2.3.7.2).

2.1.4. SRMS Preference TLV

The Segment Routing Mapping Server (SRMS) Preference sub-TLV is used in order to associate a preference with SRMS advertisements from a particular source.

The SRMS Preference sub-TLV has following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Type           |   Length       | Preference   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Type: TBD, suggested value 1037.

Length: 1.

Preference: 1 octet. Unsigned 8 bit SRMS preference.

The use of the SRMS Preference TLV is defined in [I-D.ietf-isis-segment-routing-extensions].

2.2. Link Attribute TLVs

The following Link Attribute TLVs are are defined:

TLV Code Point	Description	Length	Section
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	Section 2.2.1
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	Section 2.2.2

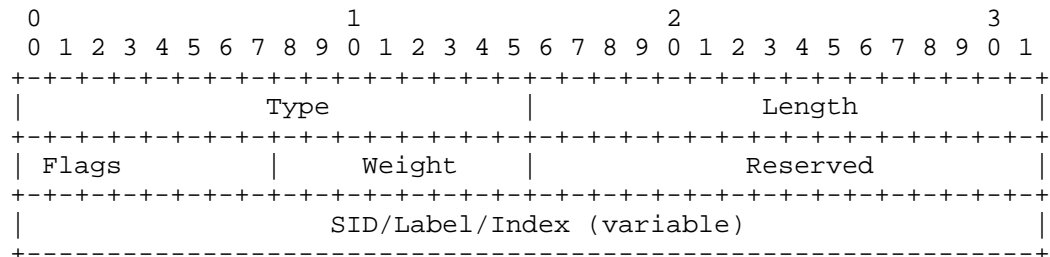
Table 2: Link Attribute TLVs

These TLVs can ONLY be added to the Link Attribute associated with the link whose local node originates the corresponding TLV.

For a LAN, normally a node only announces its adjacency to the IS-IS pseudo-node (or the equivalent OSPF Designated and Backup Designated Routers)[I-D.ietf-isis-segment-routing-extensions]. The LAN Adjacency Segment TLV allows a node to announce adjacencies to all other nodes attached to the LAN in a single instance of the BGP-LS Link NLRI. Without this TLV, the corresponding BGP-LS link NLRI would need to be originated for each additional adjacency in order to advertise the SR TLVs for these neighbor adjacencies.

2.2.1. Adjacency SID TLV

The Adjacency SID (Adj-SID) TLV has the following format:



where:

Type: TBD, suggested value 1099.

Length: Variable.

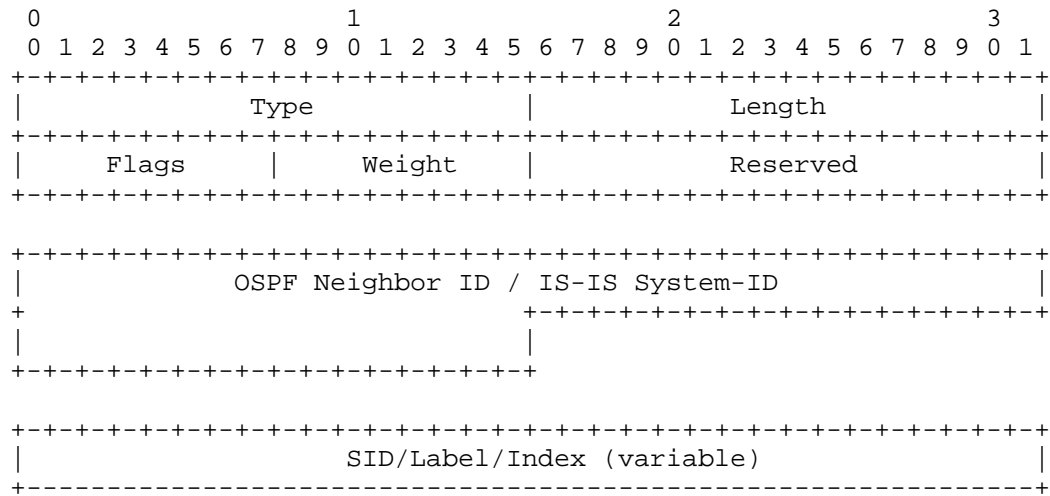
Flags. 1 octet field of following flags as defined in [I-D.ietf-isis-segment-routing-extensions], [I-D.ietf-ospf-segment-routing-extensions] and [I-D.ietf-ospf-ospfv3-segment-routing-extensions].

Weight: Weight used for load-balancing purposes.

SID/Index/Label: Label or index value depending on the flags setting as 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].

2.2.2. LAN Adjacency SID TLV

The LAN Adjacency SID (LAN-Adj-SID-SID) has the following format:



where:

Type: TBD, suggested value 1100.

Length: Variable.

Flags. 1 octet field of following flags as 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].

Weight: Weight used for load-balancing purposes.

SID/Index/Label: Label or index value depending on the flags setting as 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].

2.3. Prefix Attribute TLVs

The following Prefix Attribute TLVs and Sub-TLVs are defined:

TLV Code Point	Description	Length	Section
1158	Prefix SID	variable	Section 2.3.1
1159	Range	variable	Section 2.3.5
1160	Binding SID	variable	Section 2.3.6
1169	IPv6 Prefix SID	variable	Section 2.3.2
1170	IGP Prefix Attributes	variable	Section 2.3.3
1171	Source Router-ID	variable	Section 2.3.4

Table 3: Prefix Attribute TLVs

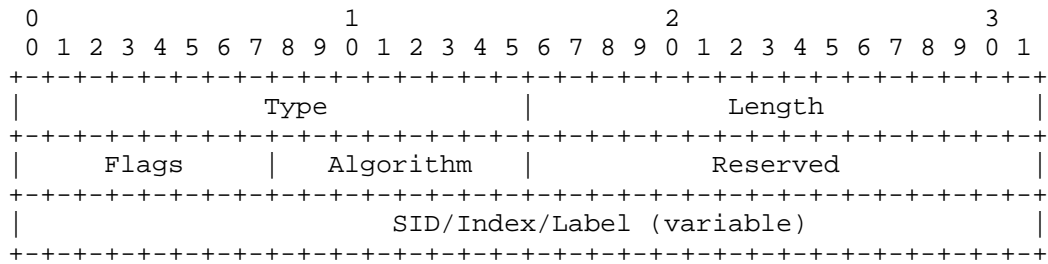
TLV Code Point	Description	Length	Section
1161	SID/Label TLV	variable	Section 2.3.7.2
1162	ERO Metric TLV	4 octets	Section 2.3.7.3
1163	IPv4 ERO TLV	8 octets	Section 2.3.7.4
1164	IPv6 ERO TLV	20 octets	Section 2.3.7.5
1165	Unnumbered Interface ID ERO TLV	12	Section 2.3.7.6
1166	IPv4 Backup ERO TLV	8 octets	Section 2.3.7.7
1167	IPv6 Backup ERO TLV	10 octets	Section 2.3.7.8
1168	Unnumbered Interface ID Backup ERO TLV	12	Section 2.3.7.9

Table 4: Prefix Attribute - Binding SID Sub-TLVs

2.3.1. Prefix-SID TLV

The Prefix-SID TLV can ONLY be added to the Prefix Attribute whose local node in the corresponding Prefix NLRI is the node that originates the corresponding SR TLV.

The Prefix-SID has the following format:



where:

Type: TBD, suggested value 1158.

Length: Variable

Algorithm: 1 octet value identify the algorithm.

SID/Index/Label: Label or index value depending on the flags setting as 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].

The Prefix-SID TLV includes a Flags field. In the context of BGP-LS, the Flags field format and the semantic of each individual flag MUST be taken from the corresponding source protocol (i.e.: the protocol of origin of the Prefix-SID being advertised in BGP-LS).

IS-IS Prefix-SID flags are defined in [I-D.ietf-isis-segment-routing-extensions] section 2.1.

OSPF Prefix-SID flags are defined in [I-D.ietf-ospf-segment-routing-extensions] section 5.

OSPFv3 Prefix-SID flags are defined in [I-D.ietf-ospf-segment-routing-extensions] section 5.

2.3.2. IPv6 Prefix-SID TLV

The IPv6 Prefix-SID TLV can ONLY be added to the Prefix Attribute whose local node in the corresponding Prefix NLRI is the node that originates the corresponding SR TLV.

The IPv6 Prefix-SID has the following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|               Type                 |               Length                 |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Flags                 |   Algorithm   |                   //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                                     Sub-TLVs                             //
//                                     //                                     //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

where:

Type: TBD, suggested value 1169.

Length: 3 + length of Sub-TLVs.

Flags: 2 octet field of flags. None of them is defined at this stage.

Algorithm: 1 octet value identify the algorithm as defined in [I-D.previdi-isis-ipv6-prefix-sid].

Sub-TLVs: additional information encoded into the IPv6 Prefix-SID Sub-TLV as defined in [I-D.previdi-isis-ipv6-prefix-sid].

The IPv6 Prefix-SID TLV is defined in [I-D.previdi-isis-ipv6-prefix-sid].

2.3.3. IGP Prefix Attributes TLV

The IGP Prefix Attribute TLV carries IPv4/IPv6 prefix attribute flags as defined in [RFC7684] and [RFC7794].

The IGP Prefix Attribute TLV has the following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|               Type                 |               Length                 |
+-----+-----+-----+-----+-----+-----+-----+-----+
//                                     Flags (variable)                             //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

where:

Type: TBD, suggested value 1170.

Length: variable.

Flags: a variable length flag field (according to the length field). Flags are routing protocol specific (OSPF and IS-IS). OSPF flags are defined in [RFC7684] and IS-IS flags are defined in [RFC7794]. The receiver of the BGP-LS update, when inspecting the IGP Prefix Attribute TLV, MUST check the Protocol-ID of the NLRI and refer to the protocol specification in order to parse the flags.

2.3.4. Source Router Identifier (Source Router-ID) TLV

The Source Router-ID TLV contains the IPv4 or IPv6 Router-ID of the originator as defined in [RFC7794]. While defined in the IS-IS protocol, the Source Router-ID TLV may be used to carry the OSPF Router-ID of the prefix originator.

The Source Router-ID TLV has the following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                         |                                         |
|                                         Type                                         Length                                         |
+-----+-----+-----+-----+-----+-----+-----+-----+
//                                         IPv4/IPv6 Address (Router-ID)                                         //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

where:

Type: TBD, suggested value 1171.

Length: 4 or 16.

IPv4/IPv6 Address: 4 octet IPv4 address or 16 octet IPv6 address.

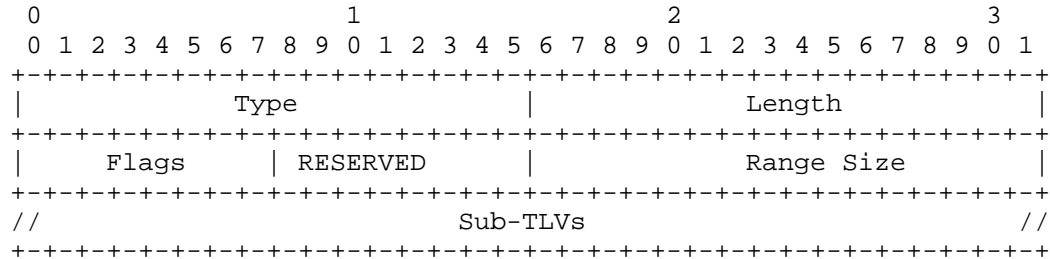
The semantic of the Source Router-ID TLV is defined in [RFC7794].

2.3.5. Range TLV

The Range TLV can ONLY be added to the Prefix Attribute whose local node in the corresponding Prefix NLRI is the node that originates the corresponding SR TLV.

When the range TLV is used in order to advertise a path to a prefix or a range of prefix-to-SID mappings, the Prefix-NLRI the Range TLV is attached to MUST be advertised as a non-routing prefix where no IGP metric TLV (TLV 1095) is attached.

The format of the Range TLV is as follows:



where:

Figure 2: Range TLV format

Type: 1159

Length is 4.

Flags: Only used when the source protocol is OSPF and defined in [I-D.ietf-ospf-segment-routing-extensions] section 4 and [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 4.

Range Size: 2 octets as defined in [I-D.ietf-ospf-segment-routing-extensions] section 4.

Within the Range TLV, the following SubTLVs are may be present:

Binding SID TLV, defined in Section 2.3.6

Prefix-SID TLV, defined in Section 2.3.1

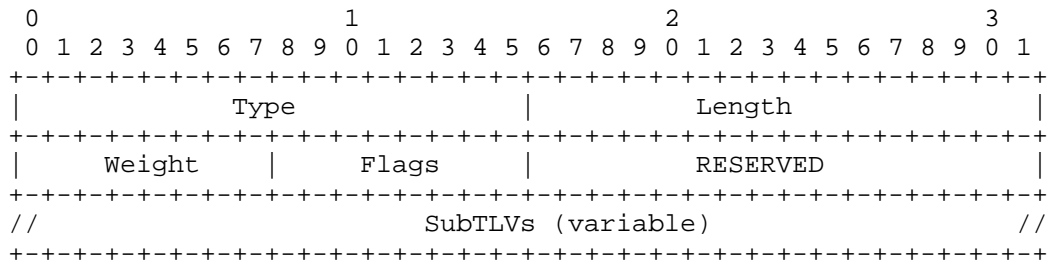
SID/Label TLV, defined in Section 2.3.7.2

2.3.6. Binding SID TLV

The Binding SID TLV can be used in two ways:

- o as a sub-TLV of the Range TLV
- o as a Prefix Attribute TLV

The format of the Binding SID TLV is as follows:



where:

Figure 3: Binding SID Sub-TLV format

Type is 1160

Length is variable

Weight and Flags are mapped to Weight and Flags defined in [I-D.ietf-isis-segment-routing-extensions] section 2.4, [I-D.ietf-ospf-segment-routing-extensions] section 4 and [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 4.

Sub-TLVs are defined in the following sections.

2.3.7. Binding SID SubTLVs

This section defines the Binding SID Sub-TLVs in BGP-LS to encode the equivalent Sub-TLVs 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].

All ERO (Explicit Route Object) Sub-TLVs must immediately follow the (SID)/Label Sub-TLV.

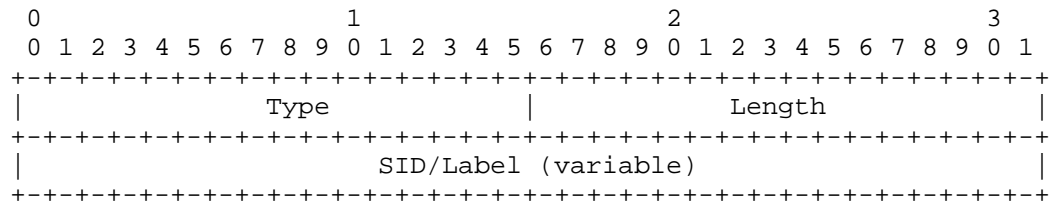
All Backup ERO Sub-TLVs must immediately follow the last ERO Sub-TLV.

2.3.7.1. Binding SID Prefix-SID Sub-TLV

When encoding IS-IS Mapping Server entries as defined in [I-D.ietf-isis-segment-routing-extensions] the Prefix-SID TLV defined in Section 2.3.1 is used as Sub-TLV in the Binding TLV.

2.3.7.2. SID/Label Sub-TLV

The SID/Label TLV has following format:



where:

Type: TBD, suggested value 1161.

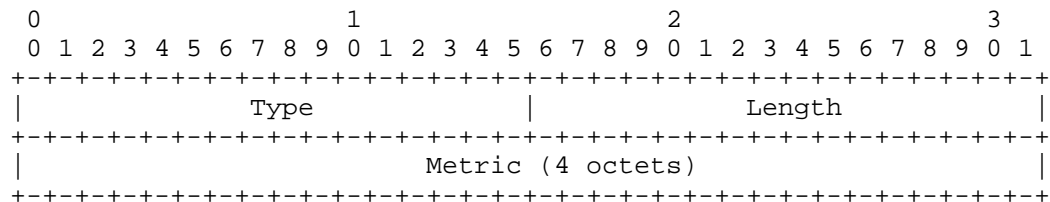
Length: Variable, 3 or 4 bytes

SID/Label: If length is set to 3, then the 20 rightmost bits represent a label. If length is set to 4, then the value represents a 32 bit SID.

The receiving router MUST ignore the SID/Label Sub-TLV if the length is other then 3 or 4.

2.3.7.3. ERO Metric Sub-TLV

The ERO Metric Sub-TLV has following format:



ERO Metric Sub-TLV format

where:

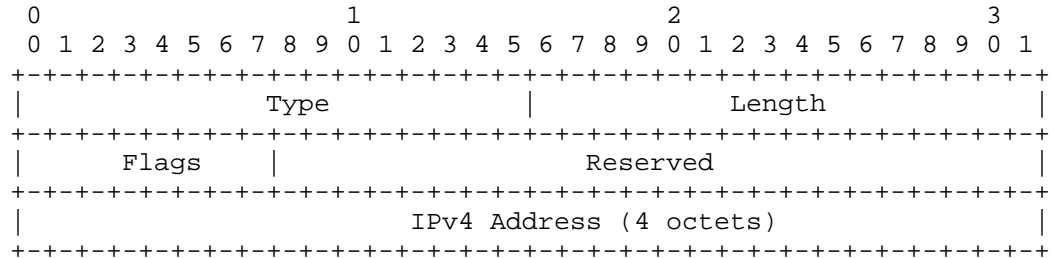
Type: TBD, suggested value 1162.

Length: Always 4

Metric: A 4 octet metric representing the aggregate IGP or TE path cost.

2.3.7.4. IPv4 ERO Sub-TLV

The ERO Sub-TLV has following format:



IPv4 ERO Sub-TLV format

where:

Type: TBD, suggested value 1163

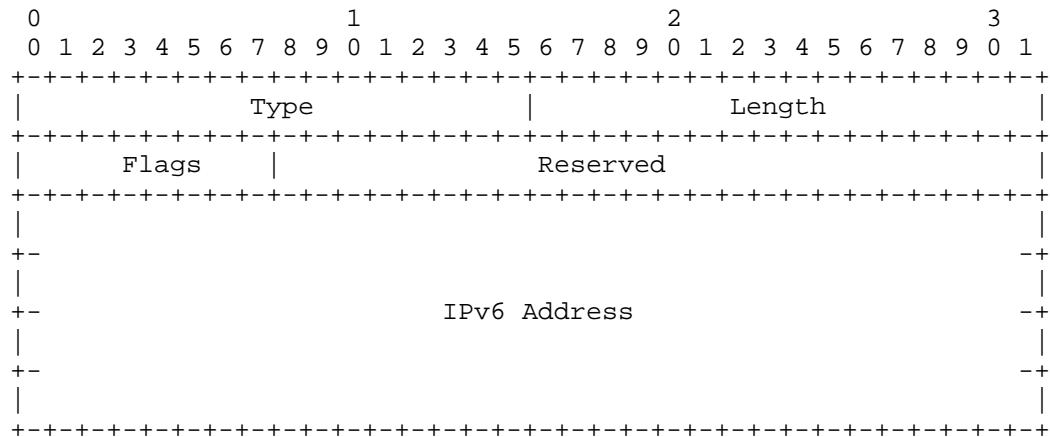
Length: 8 octets

Flags: 1 octet of flags as 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].

IPv4 Address - the address of the explicit route hop.

2.3.7.5. IPv6 ERO Sub-TLV

The IPv6 ERO Sub-TLV has following format:



IPv6 ERO Sub-TLV format

where:

Type: TBD, suggested value 1164

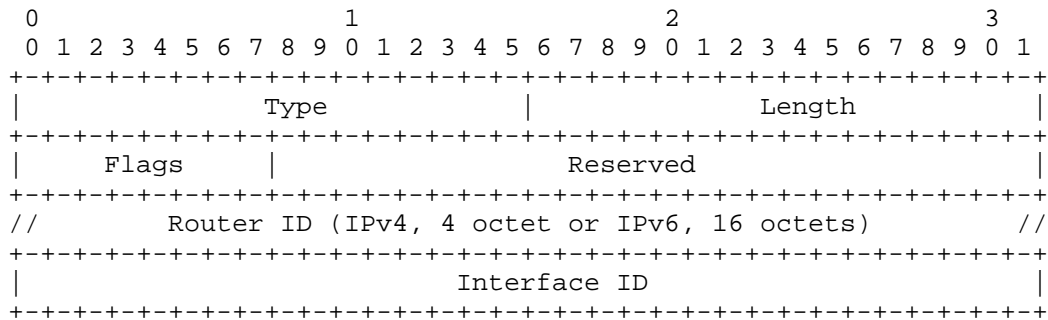
Length: 20 octets

Flags: 1 octet of flags as 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].

IPv6 Address - the address of the explicit route hop.

2.3.7.6. Unnumbered Interface ID ERO Sub-TLV

The Unnumbered Interface-ID ERO Sub-TLV has following format:



where:

Unnumbered Interface ID ERO Sub-TLV format

Type: TBD, suggested value 1165.

Length: Variable (12 for IPv4 Router-ID or 24 for IPv6 Router-ID).

Flags: 1 octet of flags as defined in:

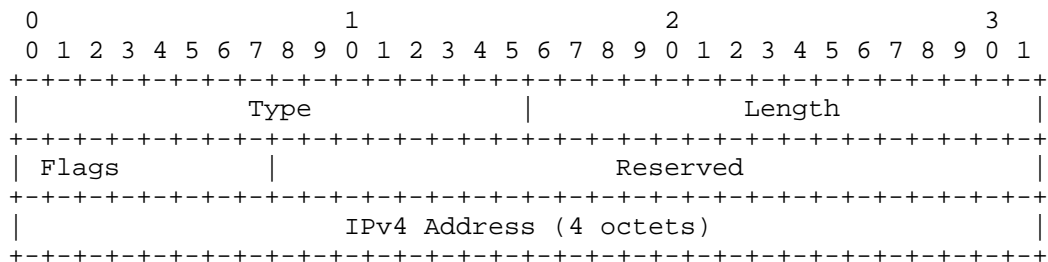
[I-D.ietf-isis-segment-routing-extensions],
[I-D.ietf-ospf-segment-routing-extensions] and
[I-D.ietf-ospf-ospfv3-segment-routing-extensions].

Router-ID: Router-ID of the next-hop.

Interface ID: is the identifier assigned to the link by the router specified by the Router-ID.

2.3.7.7. IPv4 Backup ERO Sub-TLV

The IPv4 Backup ERO Sub-TLV has following format:



IPv4 Backup ERO Sub-TLV format

where:

Type: TBD, suggested value 1166.

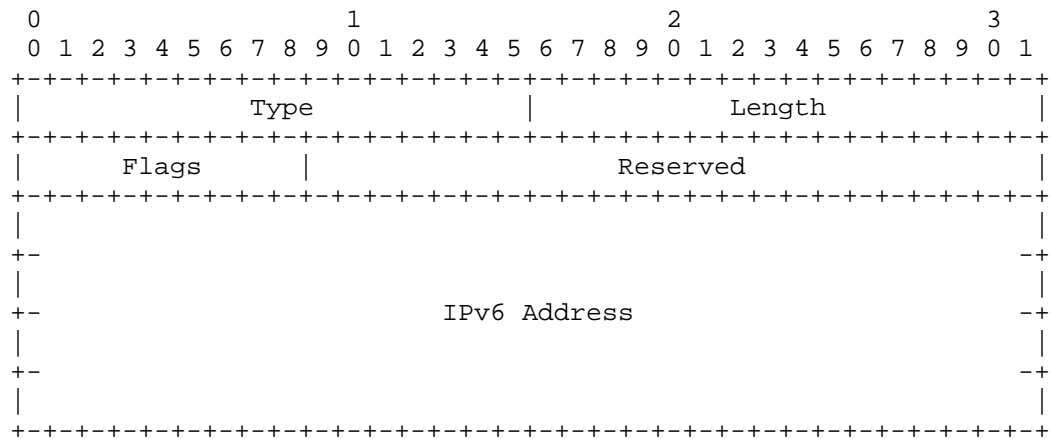
Length: 8 octets

Flags: 1 octet of flags as 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].

IPv4 Address: Address of the explicit route hop.

2.3.7.8. IPv6 Backup ERO Sub-TLV

The IPv6 Backup ERO Sub-TLV has following format:



IPv6 Backup ERO Sub-TLV format

where:

Type: TBD, suggested value 1167.

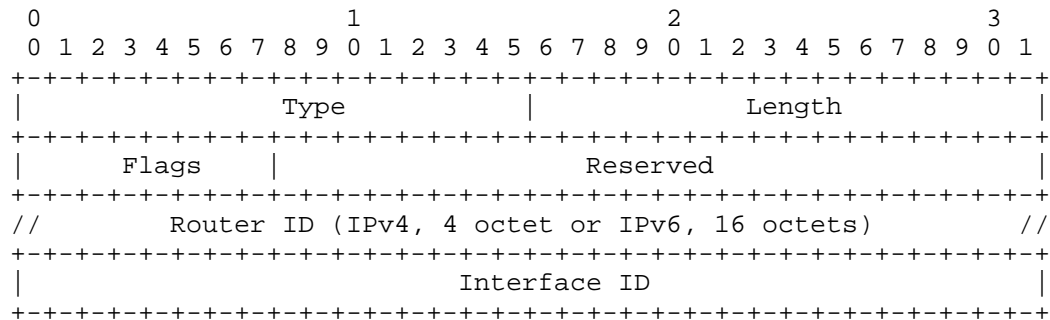
Length: 8 octets

Flags: 1 octet of flags as 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].

IPv6 Address: Address of the explicit route hop.

2.3.7.9. Unnumbered Interface ID Backup ERO Sub-TLV

The Unnumbered Interface-ID Backup ERO Sub-TLV has following format:



Unnumbered Interface ID Backup ERO Sub-TLV format

where:

Type: TBD, suggested value 1168.

Length: Variable (12 for IPv4 Router-ID or 24 for IPv6 Router-ID).

Flags: 1 octet of flags as 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].

Router-ID: Router-ID of the next-hop.

Interface ID: Identifier assigned to the link by the router specified by the Router-ID.

2.4. Equivalent IS-IS Segment Routing TLVs/Sub-TLVs

This section illustrate the IS-IS Segment Routing Extensions TLVs and Sub-TLVs mapped to the ones defined in this document.

The following table, illustrates for each BGP-LS TLV, its equivalence in IS-IS.

TLV Code Point	Description	Length	IS-IS TLV /Sub-TLV
1034	SR Capabilities	variable	2 [1]
1035	SR Algorithm	variable	19 [2]
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	31 [3]
1100	LAN Adjacency Segment Identifier (LAN-Adj-SID) TLV	variable	32 [4]
1158	Prefix SID	variable	3 [5]
1160	Binding SID	variable	149 [6]
1161	SID/Label TLV	variable	1 [7]
1162	ERO Metric TLV	4 octets	10 [8]
1163	IPv4 ERO TLV	5 octets	11 [9]
1164	IPv6 ERO TLV	17 octets	12 [10]
1165	Unnumbered Interface ID ERO TLV	variable	13 [11]
1166	IPv4 Backup ERO TLV	5 octets	14 [12]
1167	IPv6 Backup ERO TLV	17 octets	15 [13]
1168	Unnumbered Interface ID Backup ERO TLV	variable	16 [14]
1169	IPv6 Prefix SID	variable	5 [15]
1170	IGP Prefix Attributes	variable	4 [16]
1171	Source Router ID	variable	11/12 [17]

Table 5: IS-IS Segment Routing Extensions TLVs/Sub-TLVs

2.5. Equivalent OSPF/OSPFv3 Segment Routing TLVs/Sub-TLVs

This section illustrate the OSPF and OSPFv3 Segment Routing Extensions TLVs and Sub-TLVs mapped to the ones defined in this document.

The following table, illustrates for each BGP-LS TLV, its equivalence in OSPF and OSPFv3.

TLV Code Point	Description	Length	OSPF TLV /Sub-TLV
1034	SR Capabilities	variable	9 [18]
1035	SR Algorithm	variable	8 [19]
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	2 [20]
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	3 [21]
1158	Prefix SID	variable	2 [22]
1161	SID/Label TLV	variable	1 [23]
1162	ERO Metric TLV	4 octets	8 [24]
1163	IPv4 ERO TLV	8 octets	4 [25]
1165	Unnumbered Interface ID ERO TLV	12 octets	5 [26]
1166	IPv4 Backup ERO TLV	8 octets	6 [27]
1167	Unnumbered Interface ID Backup ERO TLV	12 octets	7 [28]
1167	Unnumbered Interface ID Backup ERO TLV	12 octets	7 [29]

Table 6: OSPF Segment Routing Extensions TLVs/Sub-TLVs

TLV Code Point	Description	Length	OSPFv3 TLV /Sub-TLV
1034	SR Capabilities	variable	9 [30]
1035	SR Algorithm	variable	8 [31]
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	5 [32]
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	6 [33]
1158	Prefix SID	variable	4 [34]
1161	SID/Label TLV	variable	3 [35]
1162	ERO Metric TLV	4 octets	8 [36]
1163	IPv4 ERO TLV	8 octets	9 [37]
1164	IPv6 ERO TLV	20 octets	8 [38]
1165	Unnumbered Interface ID ERO TLV	12 octets	11 [39]
1166	IPv4 Backup ERO TLV	8 octets	12 [40]
1167	IPv6 Backup ERO TLV	20 octets	13 [41]
1167	Unnumbered Interface ID Backup ERO TLV	12 octets	14 [42]

Table 7: OSPFv3 Segment Routing Extensions TLVs/Sub-TLVs

3. Procedures

The following sections describe the different operations for the propagation of SR TLVs into BGP-LS.

3.1. Advertisement of a IS-IS Prefix SID TLV

The advertisement of a IS-IS Prefix SID TLV has following rules:

The IS-IS Prefix-SID is encoded in the BGP-LS Prefix Attribute Prefix-SID as defined in Section 2.3.1. The flags in the Prefix-SID TLV have the semantic defined in [I-D.ietf-isis-segment-routing-extensions] section 2.1.

3.2. Advertisement of a OSPF/OSPFv3 Prefix-SID TLV

The advertisement of a OSPF/OSPFv3 Prefix-SID TLV has following rules:

The OSPF (or OSPFv3) Prefix-SID is encoded in the BGP-LS Prefix Attribute Prefix-SID as defined in Section 2.3.1. The flags in

the Prefix-SID TLV have the semantic defined in
[I-D.ietf-ospf-segment-routing-extensions] section 5 or
[I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 5.

3.3. Advertisement of a range of prefix-to-SID mappings in OSPF

The advertisement of a range of prefix-to-SID mappings in OSPF has following rules:

The OSPF/OSPFv3 Extended Prefix Range TLV is encoded in the BGP-LS Prefix Attribute Range TLV as defined in Section 2.3.5. The flags of the Range TLV have the semantic mapped to the definition in [I-D.ietf-ospf-segment-routing-extensions] section 4 or [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 4. The Prefix-SID from the original OSPF Prefix SID Sub-TLV is encoded using the BGP-LS Prefix Attribute Prefix-SID as defined in Section 2.3.1 with the flags set according to the definition in [I-D.ietf-ospf-segment-routing-extensions] section 5 or [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 5.

3.4. Advertisement of a range of IS-IS SR bindings

The advertisement of a range of IS-IS SR bindings has following rules:

In IS-IS the Mapping Server binding ranges are advertised using the Binding TLV. The IS-IS Binding TLV is encoded in the BGP-LS Prefix Attribute Range TLV as defined in Section 2.3.5 using the Binding Sub-TLV as defined in Section 2.3.6. The flags in the Range TLV are all set to zero on transmit and ignored on reception. The range value from the original IS-IS Binding TLV is encoded in the Range TLV "Range" field.

3.5. Advertisement of a path and its attributes from IS-IS protocol

The advertisement of a Path and its attributes is described in [I-D.ietf-isis-segment-routing-extensions] section 2.4 and has following rules:

The original Binding SID TLV (from IS-IS) is encoded into the BGP-LS Range TLV defined in Section 2.3.5 using the Binding Sub-TLV as defined in Section 2.3.6. The set of Sub-TLVs from the original IS-IS Binding TLV are encoded as Sub-TLVs of the BGP-LS Binding TLV as defined in Section 2.3.6. This includes the SID/Label TLV defined in Section 2.3.

3.6. Advertisement of a path and its attributes from OSPFv2/OSPFv3 protocol

The advertisement of a Path and its attributes is described in [I-D.ietf-ospf-segment-routing-extensions] section 6 and [I-D.ietf-ospf-ospfv3-segment-routing-extensions] section 6 and has following rules:

Advertisement of a path for a single prefix: the original Binding SID TLV (from OSPFv2/OSPFv3) is encoded into the BGP-LS Prefix Attribute Binding TLV as defined in Section 2.3.6. The set of Sub-TLVs from the original OSPFv2/OSPFv3 Binding TLV are encoded as Sub-TLVs of the BGP-LS Binding TLV as defined in Section 2.3.6. This includes the SID/Label TLV defined in Section 2.3.

Advertisement of an SR path for range of prefixes: the OSPF/OSPFv3 Extended Prefix Range TLV is encoded in the BGP-LS Prefix Attribute Range TLV as defined in Section 2.3.5. The original OSPFv2/OSPFv3 Binding SID TLV is encoded into the BGP-LS Binding Sub-TLV as defined in Section 2.3.6. The set of Sub-TLVs from the original OSPFv2/OSPFv3 Binding TLV are encoded as Sub-TLVs of the BGP-LS Binding TLV as defined in Section 2.3.6. This includes the SID/Label TLV defined in Section 2.3.

4. IANA Considerations

This document requests assigning code-points from the registry for BGP-LS attribute TLVs based on table Table 8.

4.1. TLV/Sub-TLV Code Points Summary

This section contains the global table of all TLVs/Sub-TLVs defined in this document.

TLV Code Point	Description	Length	Section
1034	SR Capabilities	variable	Section 2.1.1
1035	SR Algorithm	variable	Section 2.1.2
1036	SR Local Block	variable	Section 2.1.3
1037	SRMS Preference	variable	Section 2.1.4
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	Section 2.2.1
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	Section 2.2.2
1158	Prefix SID	variable	Section 2.3.1
1159	Range	variable	Section 2.3.5
1160	Binding SID	variable	Section 2.3.6
1161	SID/Label TLV	variable	Section 2.3.7.2
1162	ERO Metric TLV	4 octets	1 [43]
1163	IPv4 ERO TLV	8 octets	1 [44]
1164	IPv6 ERO TLV	20 octets	1 [45]
1165	Unnumbered Interface ID ERO TLV	12 octets	1 [46]
1166	IPv4 Backup ERO TLV	8 octets	1 [47]
1167	IPv6 Backup ERO TLV	20 octets	1 [48]
1168	Unnumbered Interface ID Backup ERO TLV	12 octets	1 [49]
1169	IPv6 Prefix SID	variable	Section 2.3.2
1170	IGP Prefix Attributes	variable	Section 2.3.3
1171	Source Router-ID	variable	Section 2.3.4

Table 8: Summary Table of TLV/Sub-TLV Codepoints

5. Manageability Considerations

This section is structured as recommended in [RFC5706].

5.1. Operational Considerations

5.1.1. Operations

Existing BGP and BGP-LS operational procedures apply. No additional operation procedures are defined in this document.

6. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See the 'Security Considerations' section of [RFC4271] for a discussion of BGP security. Also refer to [RFC4272] and [RFC6952] for analysis of security issues for BGP.

7. Contributors

The following people have substantially contributed to the editing of this document:

Acee Lindem
Cisco Systems
Email: acee@cisco.com

Saikat Ray
Individual
Email: raysaikat@gmail.com

8. Acknowledgements

The authors would like to thank Les Ginsberg for the review of this document.

9. References

9.1. Normative References

[I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and j. jeffrant@gmail.com, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-08 (work in progress), October 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.previdi-isis-ipv6-prefix-sid]
Previdi, S., Ginsberg, L., and C. Filsfils, "Segment Routing IPv6 Prefix-SID", draft-previdi-isis-ipv6-prefix-sid-02 (work in progress), May 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>>.
- [RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <<http://www.rfc-editor.org/info/rfc7684>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<http://www.rfc-editor.org/info/rfc7752>>.
- [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>>.

9.2. Informative 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-09 (work in progress), July 2016.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<http://www.rfc-editor.org/info/rfc4272>>.
- [RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, DOI 10.17487/RFC5706, November 2009, <<http://www.rfc-editor.org/info/rfc5706>>.

- [RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <<http://www.rfc-editor.org/info/rfc6952>>.

9.3. URIs

- [1] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-3.1>
- [2] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-3.2>
- [3] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.2.1>
- [4] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.2.2>
- [5] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.1>
- [6] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4>
- [7] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.3>
- [8] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.7>
- [9] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.8>
- [10] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.9>
- [11] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.10>
- [12] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.11>
- [13] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.12>

- [14] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.13>
- [15] <http://tools.ietf.org/html/draft-previdi-isis-ipv6-prefix-sid-01>
- [16] <http://tools.ietf.org/html/RFC7794>
- [17] <http://tools.ietf.org/html/RFC7794>
- [18] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-3.2>
- [19] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-3.1>
- [20] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-7.1>
- [21] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-7.2>
- [22] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-5>
- [23] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-2.1>
- [24] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.1>
- [25] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.1>
- [26] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.2>
- [27] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.3>
- [28] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.4>
- [29] <http://tools.ietf.org/html/draft-ietf-ospf-segment-routing-extensions-05#section-6.2.4>
- [30] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-3.2>

- [31] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-3.1>
- [32] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-7.1>
- [33] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-7.2>
- [34] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-5>
- [35] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-2.1>
- [36] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.1>
- [37] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.1>
- [38] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.2>
- [39] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.3>
- [40] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.4>
- [41] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.5>
- [42] <http://tools.ietf.org/html/draft-ietf-ospf-ospfv3-segment-routing-extensions-05#section-6.2.6>
- [43] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.7>
- [44] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.8>
- [45] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.9>
- [46] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.10>

- [47] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.11>
- [48] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.12>
- [49] <http://tools.ietf.org/html/draft-ietf-isis-segment-routing-extensions-05#section-2.4.13>

Authors' Addresses

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

Email: sprevidi@cisco.com

Peter Psenak
Cisco Systems, Inc.
Apollo Business Center
Mlynske nivy 43
Bratislava 821 09
Slovakia

Email: ppsenak@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Brussels
Belgium

Email: cfilsfil@cisco.com

Hannes Gredler
RtBrick Inc.

Email: hannes@rtbrick.com

Mach(Guoyi) Chen
Huawei Technologies
Huawei Building, No. 156 Beiqing Rd.
Beijing 100095
China

Email: mach.chen@huawei.com

Jeff Tantsura
Individual

Email: jefftant@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 2, 2017

J. Haas
Juniper Networks, Inc.
March 1, 2017

Extended Experimental Path Attributes for BGP
draft-haas-idr-extended-experimental-01

Abstract

BGP's primary feature extension mechanism, Optional-Transitive Path Attributes, has proven to be a successful mechanism to permit BGP to be extended. In order to ease various issues during the development of new BGP features, this document proposes an extended experimental Path Attribute to carry prototype features.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning.

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 September 2, 2017.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Extended Experimental Path Attribute	3
3. Usage	4
4. Error Handling	4
5. Moving to an Allocated Code Point	4
6. Security Considerations	5
7. IANA Considerations	5
8. References	5
8.1. Normative References	5
8.2. Informative References	6
Appendix A. Comparisons to Other Features	6
Appendix B. Discussion to this Date	7
Author's Address	8

1. Introduction

BGP's [RFC4271] primary feature extension mechanism, Optional-Transitive Path Attributes, has proven to be a successful mechanism to permit BGP to be extended. It permits implementations to propagate unknown Path Attributes without understanding their contents, so long as they are syntactically valid.

Path Attributes are encoded in BGP UPDATE messages using a single octet code-point. While this code-point space is relatively small, the rate at which new BGP features are introduced has proven to be slow enough that the potential for exhaustion has not been a significant concern more than twenty years into the deployment of BGP-4. This code point space is managed by IANA under the Standards Action policy [RFC5226], one of the more restrictive policies in IETF's repertoire. Early allocation [RFC7120] provides some latitude for allocation of these code points compared to the original RFC 5226 policy, but is reserved for features that are considered appropriately stable.

Development work on the BGP protocol often requires a code point be assigned to a feature in progress. While code point 255 has been

reserved to be Experimental ([RFC2042]), developers will often face collisions when attempting to do development on more than a single in-progress feature. Once the feature has reached a level of stability, early allocation should be strongly pursued. It may take some time, however, for features to reach that level of stability.

Due to the general difficulty of getting a public code point during the development process, code point "squatting" (use of a code point that has not been officially allocated) is unfortunately common. In many cases, this is done completely internally and has no impact on the Internet. But sometimes accidents happen and pre-release features ship. Prior to the deployment of the Revised BGP Error Handling Procedures [RFC7606], this could often be disastrous as different features, or different versions of the same feature, collided with each other and were interpreted as syntax errors and caused BGP peering sessions to reset per RFC 4271 error handling procedures. While it is less disastrous for such collisions to happen in terms of stability of the Internet, what's needed is a way for BGP protocol development to proceed with a little more safety.

This document proposes a new BGP Path Attribute, the BGP Extended Experimental Path Attribute. This Attribute is intended to be used solely for BGP Protocol development and is not intended to replace the allocation policies for the BGP Protocol.

2. Extended Experimental Path Attribute

The Extended Experimental Path Attribute is an Optional-Transitive Path Attribute with a code of TBD. Its contents are a series of TLVs in the following format:

```

  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
  +-----+-----+-----+-----+-----+-----+-----+-----+
  | Implementor IANA Private Enterprise Number (4 octets) |
  +-----+-----+-----+-----+-----+-----+-----+-----+
  | Implementor Feature Code Point Number (4 octets) |
  +-----+-----+-----+-----+-----+-----+-----+-----+
  | Version Number (2 octets) | Feature Length (2 octets) |
  +-----+-----+-----+-----+-----+-----+-----+-----+
  | Feature Data (0 or more octets) |
  +-----+-----+-----+-----+-----+-----+-----+-----+

```

- o Implementor IANA Private Enterprise Number is a Private Enterprise Number (PEN) assigned by IANA. [IANA-PEN]
- o Feature Code Point Number is a code point space under the control of the holder of the PEN.
- o Version Number is an unsigned number intended to convey the version of the feature covered by the Feature Code Point Number

for the implementor. Implementors are encouraged to sequentially number versions of their feature beginning at 1.

- o Feature Length is the length of the Feature Data.
- o Feature Data will be encoded as a BGP Path Attribute value for the experimental feature.

3. Usage

A BGP implementor intending to introduce a new standards oriented Path Attribute will select a code point number for their new Path Attribute and assign an initial Version Number. Whenever the format of the feature needs to change, the Version Number MUST also change. This prevents implementations understanding different versions of a pre-standards feature from improperly parsing the attribute.

BGP Experimental features MUST require explicit configuration to recognize a specific Feature Code Point Number, for a given Version Number, for a given PEN. If such configuration is not present, the TLV MUST be ignored.

BGP Experimental features SHOULD NOT carry more than one Version Number of the same Feature Code Point in a given UPDATE. Implementations are encouraged to strip inconsistent Version Numbered TLVs for a given feature when appropriate. For example, if the BGP speaker is configured to support Version Number 2 of an experimental feature, it may discard all TLVs for the Feature Code Point Number that are not 2.

BGP implementations supporting the Extended Experimental Path Attribute SHOULD strip this attribute by default on external BGP sessions. Explicit configuration SHOULD be required to permit a given PEN+FCPN+VN tuple into the network.

4. Error Handling

If the Extended Experimental Path Attribute is determined to be syntactically invalid, the Attribute discard behavior from [RFC7606] MUST be used.

5. Moving to an Allocated Code Point

Once an evolving BGP protocol feature reaches a reasonable level of stability, implementations MUST move to a Path Attribute Code Point allocated using the IETF sanctioned procedures. Implementors that publish their PEN+FCN+VN allocations for a given version of their feature in progress are recommended to publish this binding as part of their allocation request to enable short term backward compatibility with their experimental work.

While it is possible for implementations of a new feature to rely on experimental deployment for some time, the procedures noted in Section 3 are intended to discourage this behavior by making inter-domain distribution of the experiment fail by default.

6. Security Considerations

This document does not introduce any new security considerations into the BGP-4 protocol. While the injection of unknown or badly formatted Optional-Transitive Path Attributes has been and remains an issue impacting the stability of the Internet, this proposal doesn't increase exposure to that issue. It is rather expected that this proposal helps remediate the accidental attack surface that incremental BGP protocol work exposes to the Internet at large.

[RFC7606] has mitigated the majority of the issues mentioned in the prior paragraph. See that RFC for further information on the history of the problem.

7. IANA Considerations

This document is primarily about issues related to IANA Considerations. At some point, IANA will be requested to assign a BGP Path Attribute Code number, referenced as TBD early in the document.

8. References

8.1. Normative References

- [IANA-PEN] "IANA Private Enterprise Number", <<http://pen.iana.org>>.
- [RFC2042] Manning, B., "Registering New BGP Attribute Types", RFC 2042, DOI 10.17487/RFC2042, January 1997, <<http://www.rfc-editor.org/info/rfc2042>>.
- [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>>.

- [RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <<http://www.rfc-editor.org/info/rfc7606>>.

8.2. Informative References

- [I-D.ietf-idr-bgp-attribute-announcement]
Patel, K., Uttaro, J., Decraene, B., Henderickx, W., and J. Haas, "Constrain Attribute announcement within BGP", draft-ietf-idr-bgp-attribute-announcement-00 (work in progress), July 2016.
- [ietf-97-idr-code-point-management-slides]
Haas, J., "Code Point Management - IETF 97 Slides", November 2016, <<https://www.ietf.org/proceedings/97/slides/slides-97-idr-code-point-management-02.pdf>>.
- [ietf-97-idr-extended-experimental-path-attribute-slides]
Haas, J., "Extended Experimental Path Attributes - IETF 97 Slides", November 2016, <<https://www.ietf.org/proceedings/97/slides/slides-97-idr-extended-experimental-path-attributes-00.pdf>>.
- [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>>.
- [RFC6368] Marques, P., Raszuk, R., Patel, K., Kumaki, K., and T. Yamagata, "Internal BGP as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 6368, DOI 10.17487/RFC6368, September 2011, <<http://www.rfc-editor.org/info/rfc6368>>.
- [RFC7120] Cotton, M., "Early IANA Allocation of Standards Track Code Points", BCP 100, RFC 7120, DOI 10.17487/RFC7120, January 2014, <<http://www.rfc-editor.org/info/rfc7120>>.

Appendix A. Comparisons to Other Features

Astute readers will note that this is not the first time BGP Path Attributes have been "tunneled" inside of other Path Attributes. [RFC6368] provided a mechanism by which an entire set of Path Attributes could be tunneled inside of attribute 128 for purposes of transparently passing received BGP Path Attributes in an Internet Layer 3 VPN context from one Customer Edge (CE) router to another.

[RFC6368] suffered from two issues:

1. During its initial development, 4-byte AS numbers were starting to be deployed. This led to a change in the packet format of the feature to accommodate the 4-byte ASes instead of the previous 2-byte versions.
2. While this feature was intended solely to be used in a VPN context, implementations that did not understand it similarly did not strip it. This caused the VPN routes to carry attribute 128 in an Internet context after they were delivered to the target CE router.

Due to these two issues, routes containing one version of this feature that "escaped into the wild" eventually to be received by other BGP speakers supporting a different version of the feature. Each version would treat their opposite's encoding as a syntax error. This resulted in BGP peering sessions being reset. This, and other similar issues, was a motivation for [RFC6368].

The second issue noted above is the motivation for [I-D.ietf-idr-bgp-attribute-announcement].

Appendix B. Discussion to this Date

This proposal was originally well-received on the IDR mailing list and during its presentation at IETF. Comments included comparison to existing mechanisms in LDP and IS-IS; Hannes Gredler notes that the IS-IS feature is not used.

Another set of comments revolved around the structured format of the PEN+FCN+VN and "why couldn't we simply have a very large first-come, first-served code space". While the author agrees that this would serve a very similar behavior, the author's belief after further consideration is that:

- o Involving IANA, even when the process is very light weight, is part of our existing issue. The Enterprise numbering space permits completely internal management during development of new features.
- o There is no fundamental "burden" of multiple implementors rendezvousing around a common PEN+FCN+VN during interoperability testing. The motivation after such testing should be to request a valid BGP Path Attribute code point using existing IETF procedures.

Another comment was about the possibility of utilizing this mechanism as a long-term private BGP Path Attribute feature. Such behavior may

be a valid use case, however, there remains a need to provide for automatic filtering of experimental work.

This brings the final comment that both this new Path Attribute and potentially each of the experiments in the Feature Data should be covered by [I-D.ietf-idr-bgp-attribute-announcement] or something similar. This would include additional procedure to provide for remote filtering of the TLVs defined in this document. Progressing this document, and the use case of long term private Path Attributes as noted in the prior section, should be considered after the attribute-announcement draft receives further feedback.

Author's Address

Jeffrey Haas
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089
US

Email: jhaas@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 09, 2017

K. Patel
A. Vyavaharkar
N. Fazlollahi
Cisco Systems
A. Przygienda
Juniper Networks
July 08, 2016

Extension to BGP's Route Refresh Message
draft-idr-bgp-route-refresh-options-00.txt

Abstract

[RFC2918] defines a route refresh capability to be exchanged between BGP speakers. BGP speakers that support this capability are advertising that they can resend the entire BGP Adj-RIB-Out on receipt of a refresh request. By supporting this capability, BGP speakers are more flexible in applying any inbound routing policy changes as they no longer have to store received routes in their unchanged form or reset the session when an inbound routing policy change occurs. The route refresh capability is advertised per AFI, SAFI combination.

There are newer AFI, SAFI types that have been introduced to BGP that support a variety of route types (e.g. IPv4/MVPN, L2VPN/EVPN). Currently, there is no way to request a subset of routes in a Route Refresh message for a given AFI, SAFI. This draft defines route refresh capability extensions that help BGP speakers to request a subset of routes for a given address family. This is expected to reduce the amount of update traffic being generated by route refresh requests as well as lessen the burden on the router servicing such requests.

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 January 09, 2017.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	3
1.1. Use Case Examples	3
2. Requirements Language	4
3. Route Refresh Options Capability	4
4. Route Refresh Sub-Types	4
5. Route Refresh Option format	5
6. Route Refresh Option Length	6
7. Route Refresh ID	6
8. Route Refresh Option Flags	7
9. Route Refresh Options	7
10. Operation	9
11. Error Handling	10
12. IANA Considerations	11
13. Security Considerations	11
14. Acknowledgements	11
15. References	12
15.1. Normative References	12

15.2. Information References	12
Authors' Addresses	13

1. Introduction

[RFC2918] defines a route refresh capability to be exchanged between BGP speakers. BGP speakers that support this capability are advertising that they can resend the entire BGP Adj-RIB-Out on receipt of a refresh request. By supporting this capability, BGP speakers are more flexible in applying inbound routing policy changes as they no longer have to store copies of received routes in their unchanged form or reset the session when an inbound routing policy change occurs. The route refresh capability is advertised per AFI, SAFI combination.

Route refresh allows routers to dynamically request a full Adj-RIB-Out update from their peers when there's an inbound routing policy change. This is useful because routers that mutually support this capability no longer have to flap the peering session or store an extra copy of received routes in their original form. This helps by reducing memory requirements as well as eliminating the unnecessary churn caused by session flaps. [RFC2918] does not define a way for routers to request a subset of the Adj-RIB-Out for a given AFI, SAFI.

This draft defines new extensions to route refresh that will allow requesting routers to ask for a subset of the Adj-RIB-Out for a given AFI, SAFI combination. For example, routers could ask for specific route types from those address families that support multiple route types or, they could ask for a specific prefix.

As part of the new extensions, this draft combines elements of [RFC7313] and [RFC5291] and adds a new set of options to the route refresh message that will specify filters that can be applied to limit the scope of the refresh being requested. The new option format will apply to all new option types that may be defined moving forward.

1.1. Use Case Examples

The authors acknowledge that while the extensions being proposed in this draft could potentially be addressed by Route Target Constrain described in [RFC4684] by using route targets to identify desired subset of routes, this proposal includes address families where RT Constrain extension is not supported and avoids the necessity to assign and manage the route targets per desired set of routes. The approach in this draft is intended to be a single-hop refresh only, i.e., propagation of the refreshes in a way similar to RT Constrain routes is NOT intended.

Several possible use cases are discernible today:

- o The capacity to refresh routes of a certain type within an address family is needed, e.g., auto discovery routes within the EVPN AF [RFC7432].
- o In VPN scenarios where RT Constrain is not supported or configured, RDs can be used.
- o In BGP LS [RFC7752] cases a speaker may choose to hold only a subset of routes and depending on configuration request a subset of routes. This document could provide further filters to support those use cases.
- o On changes in inbound policy, when previously configured filters have been removed, only the according subset of routes may be requested.

2. 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].

3. Route Refresh Options Capability

A BGP speaker will use the BGP Capabilities Advertisement [RFC5492] to advertise the Route Refresh Options Capability to its peers. This new capability will be advertised using the Capability code [TBD] with a capability length of 0.

By advertising the Route Refresh Options Capability to a peer, a BGP speaker indicates that it is capable of receiving and processing the route refresh options described below. This new capability can be advertised along with the Enhanced Route Refresh Capability described in [RFC7313]. However, if the Route Refresh Options Capability has been negotiated by both sides of the BGP session, then it will override the Enhanced Route Refresh Capability.

4. Route Refresh Sub-Types

[RFC7313] defines route refresh BGP message sub-types that utilize the "Reserved" field of the Route Refresh message originally defined in [RFC2918]. Currently, there are three sub-types defined and this draft proposes three additional sub-types which will be used to indicate a Route Refresh message that includes options before any ORF field of the Route Refresh message as well as BoRR and EoRR Route Refresh messages with options.

- 0 - Normal route refresh request [RFC2918]
with/without Outbound Route Filtering (ORF) [RFC5291]
- 1 - Demarcation of the beginning of a route refresh
(BoRR) operation
- 2 - Demarcation of the ending of a route refresh
(EoRR) operation
- + 3 - Route Refresh request with options and optional
ORF [RFC5291]
- + 4 - BoRR with options
- + 5 - EoRR with options
- 255 - Reserved

When the Route Refresh Options Capability has been negotiated by both sides of a BGP session, both peers MUST use message types 3, 4 and 5. The requesting speaker MUST use the refresh ID for all refresh requests including those without any options, i.e., requests for the full BGP Adj-RIB-Out.

The Route Refresh Request Message with options will now be formatted as shown below

```

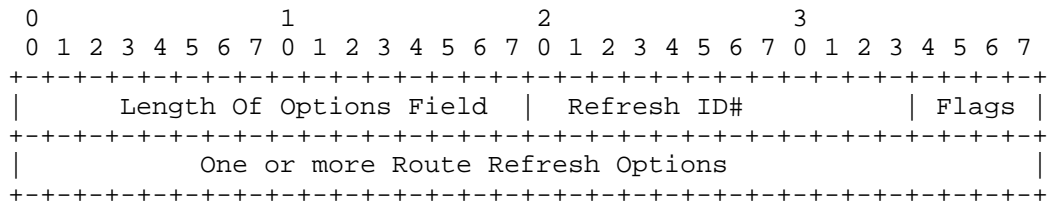
      0               1               2               3
      0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7
+-----+-----+-----+-----+-----+-----+-----+-----+
|               A F I               |   Res.   |   S A F I   |
+-----+-----+-----+-----+-----+-----+-----+-----+
|   Total Option Length   |   Refresh ID#   |   Flags   |
+-----+-----+-----+-----+-----+-----+-----+-----+
|   One or more Route Refresh Options   |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

5. Route Refresh Option format

[RFC2918] defines the route refresh BGP message that includes only the AFI, SAFI of the routes being requested. This draft proposes extending the basic message by including options that will indicate to the remote BGP speaker that a subset of the entire Adj-RIB-Out is being requested. The remote BGP speaker will select routes that match the specified options and the flag settings.

As described in the previous section, the options will be added to the Route Refresh message before the ORF field of the message. Outbound Route Filtering is described in [RFC5291]. The options will assume the following format



6. Route Refresh Option Length

The Option Length field will occupy the two octets immediately following the Route Refresh message containing the AFI, SAFI and sub-type. The purpose of this field is to allow the BGP speaker to calculate the length of any attached ORF fields by subtracting the Option Length from the Route Refresh message length.

7. Route Refresh ID

The Refresh ID field will occupy twelve bits following the Route Refresh Options Length. It is a value assigned by the requesting BGP speaker. It MUST be a strictly monotonically increasing number per peer AFI and SAFI and will be comparable using the calculations standardized in [RFC1982]. The purpose of this field is to allow the requesting BGP speaker to correlate concurrent, overlapping refresh requests and ultimately delete correct stale routes. The Refresh ID MUST be reflected in the BoRR and EoRR messages sent by the BGP speaker servicing the refresh request.

A Refresh ID value MUST NOT be reused until an EoRR with this ID has been received by the requesting speaker or the last resort time has expired. The behavior is unspecified otherwise. More specifically, defining the interval [LID, HID] by the values

$$\text{LID} = \text{MAX}(\text{lowest requested Refresh ID\# without BoRR,} \\ \text{lowest received BoRR without EoRR})$$

and

$$\text{HID} = \text{highest requested Refresh ID\#}$$

the requesting speaker MUST only use values V where $V > \text{LID}$ and $V > \text{HID}$ under [RFC1982].

Value of 0 SHOULD NOT be used as Refresh ID.

The sending speaker MUST NOT reorder the BoRR messages on sending in case it received multiple requests, i.e., the BoRRs MUST follow in the same sequence as the requested Route Refresh IDs.

8. Route Refresh Option Flags

This draft defines route refresh option flags to

- o specify whether the receiving BGP speaker MUST logically OR the attached options or logically AND them. When the flag is clear, the router on the receiving end SHOULD logically AND the options and only refresh routes that match all received options. If the option flag is set, the router SHOULD select routes that match using a logical OR of the options. In any case the set of routes sent between the according BoRR and EoRR MUST contain at least the logically requested set.
- o indicate that the receiving BGP speaker MUST clear immediately all the received Route Refresh Requests with Options, either pending or being processed. EoRRs MUST NOT be sent. The Refresh ID# on the request is free of restrictions and MUST be set as first number in the sequence number space per [RFC1982]. The C flag MUST NOT be set on BoRR or EoRR messages and CAN be used only with refresh requests.

The precise format is indicated below

```

  0 1 2 3 4 5 6 7
+---+---+---+---+
|   ....  |C|O| R |
+---+---+---+---+

```

C Clear pending requests and reset Refresh ID# space.

O Use logical OR of attached options

R Reserved bits

9. Route Refresh Options

This draft introduces new options carried within the Route Refresh message as shown in the following figure

```

      0               1               2               3
0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7

```

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|  Type      |                               Length                               |  Value      |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|                               Value (cont'd).                               |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+

```

The option Type is a 1 octet field that uniquely identifies individual options. The Length is a 2 octet field that contains the length of the option Value field in octets. The option Value is a variable length field that is interpreted according to the value of the option Type field.

The following types are being defined in this draft and additional types can be defined subsequently as needed

- + 1 - Route Type
- + 2 - NLRI Prefix
- + 3 - Route Distinguisher Prefix

The Route Type option would specify a particular route type that is being requested. This option applies specifically to those AFI/SAFI combinations that support multiple route types, e.g. L2VPN/EVPN and MUST be otherwise ignored. The value field would be the route type specifying which route type was being requested. The length of the option depends on the AFI/SAFI.

The NLRI Prefix option would specify a request for all matching address prefixes with their lengths equal to or greater than the specified prefix per AFI/SAFI definitions. The value field would contain the address prefix according to the NLRI specification of the AFI/SAFI contained in the Route Refresh message. For those AFI/SAFI combinations that specify NLRIs containing a type and/or RD, the value field MUST exclude the type and RD and SHOULD only include any remaining NLRI fields. If the requesting speaker expects its peer to also match the type and/or RD, the speaker CAN include the type and RD prefix options accordingly. The length field would contain the length of the value field in bits.

The Route Distinguisher prefix option would specify an RD prefix that is being requested for AFs that support it. The receiving BGP speaker would then refresh all routes in the specified AFI/SAFI that matched the requested RDs. The Value field would contain the RD, its length and the mask length of the RD prefix. This option applies specifically to those AFI/SAFI combinations that support route distinguishers and MUST be otherwise ignored.

10. Operation

A BGP speaker that understands and supports Route Refresh Options SHOULD advertise the Route Refresh Options Capability in its Open message. The following procedures for route refresh are only applicable if the BGP speaker originating the route refresh has received the route refresh options capability and supports it.

When originating a Route Refresh message, a BGP speaker SHOULD use and set these options if it wants to restrict the scope of updates being refreshed. The specific options being sent will be set according to the operator's command.

When a BGP speaker receives a route refresh message that includes any options, it MUST parse the options and strongly SHOULD use them to filter outgoing NLRIs when refreshing the Adj-RIB-Out to the requesting BGP speaker.

If a BGP speaker receives the route refresh message with the message subtype set to BoRR with options as described above, then it needs to process all the included options and MUST mark all matching routes as stale as described in [RFC7313].

If a BGP speaker receives the route refresh message with the message subtype set to EoRR with options as described above, then it needs to process all the included options and delete any remaining stale routes that match the options received with the EoRR as described in [RFC7313].

A BGP speaker responding to a route refresh request MUST set the message subtypes of the BoRR and EoRR messages so that each BoRR message has a matching EoRR message. This means a BoRR message without options SHOULD only be followed eventually by an EoRR message without options. Similarly, a BoRR message with options MUST eventually be followed by an EoRR message with the same options. If BoRR and EoRR message options do not match, the outcome is unpredictable as remaining staled routes pending a refresh may get inadvertently deleted. BGP speakers MUST NOT summarize EoRR messages by combining options in order to allow the requesting BGP speaker to uniquely identify the included sets of routes when concurrent refreshes are originated with overlapping sets of routes.

Observe that overlapping refreshes with different options are possible and in such case the according BoRR and EoRR messages are associated by using their Refresh ID#. The BGP speaker responding to the route refresh requests MAY perform the refreshes in parallel. In case of concurrent refreshes overlapping same routes, the responding speaker MUST ensure that the sent advertisements will result in

deletion of the omitted routes at the time all EoRRs have been received by the remote speaker or it MUST explicitly advertise withdrawals to correct any anomalies.

The BGP speaker requesting a refresh from its peers SHOULD maintain a locally configurable upper bound on how long it will keep matching stale routes once a BoRR has been received. Each subsequent BoRR SHOULD reset this period so that any remaining stale routes are only flushed after the last BoRR has been received in case there are multiple back-to-back refreshes being sent out and the last matching EoRR is never received or arrives too late. This is an implementation specific detail.

11. Error Handling

The handling of malformed options MUST follow the procedures mentioned in [RFC7606]. This draft obsoletes some of the error handling procedures in [RFC7313] if the Route Refresh Options Capability is sent. In addition, this draft mandates the following behavior at the receiver of the route refresh request upon detection of:

Length errors - If the message length minus the fixed-size message header is less than 4, the procedure in [RFC7313] MUST be followed. Also, if the overall length of all the options or any individual option length exceeds the total number of remaining bytes, the same procedure MUST be followed.

Option type errors - Any unknown option type CAN be ignored for AND'ed options. In case of OR'ed options the receiving speaker MUST ignore all the options and de-facto treat it as a full AFI/SAFI Adj-RIB-Out refresh. Such event SHOULD be logged in either case to notify the operator.

Option value errors - Length errors which cannot be distinguished from value field errors at the receiver are treated the same as value errors. The receiver MUST send a NOTIFICATION message with the Error Code "ROUTE-REFRESH Message Error" and the subcode of Invalid Message Length to the peer. The Data field of the NOTIFICATION message MUST contain the complete ROUTE-REFRESH message.

BoRR with unknown Refresh ID# - The receiver MUST discard all pending requests and issue a Route Refresh Request with Options. The options MUST be empty and the clear flag MUST be set to resynchronize the RIBs. "Unknown" means here a BoRR which is not in the interval

[MAX(lowest requested Refresh ID# without BoRR,
highest received BoRR+1 respecting [RFC1982]),

highest requested Refresh ID#]

EoRR with unknown Refresh ID# - Those SHOULD be ignored and a warning or error MUST be logged.

BoRR or EoRR with incorrect options - analogous to BoRR with unknown Refresh ID#.

EoRR with known Refresh ID# but without preceding BoRR - analogous to EoRR with unknown Refresh ID#. Observe that this can be caused by the peer expiring last resort timer and reusing the ID# for another request before the EoRR is received. This should be extremely unlikely given the size of the refresh ID space.

12. IANA Considerations

This draft defines a new route refresh options format for BGP Route Refresh messages.

This draft defines a new route refresh capability for BGP Route Refresh messages. We request IANA to record this capability to create a new registry under BGP Capability Codes as follows:

+74 Route Refresh Options Capability

This draft defines 3 new route refresh message subtypes for BGP Route Refresh messages. We request IANA to record these subtypes to create a new registry under BGP Route Refresh Subcodes as follows:

- + 3 - Route Refresh with options
- + 4 - BoRR with options
- + 5 - EoRR with options

13. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC7313] and [RFC4271].

14. Acknowledgements

The authors would like to thank Anant Utgikar for initial discussions resulting in this work. John Scudder and Jeff Hass provided further comments.

15. References

15.1. Normative References

- [RFC1982] Elz, R. and R. Bush, "Serial Number Arithmetic", RFC 1982, DOI 10.17487/RFC1982, August 1996, <<http://www.rfc-editor.org/info/rfc1982>>.
- [RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, DOI 10.17487/RFC2918, September 2000, <<http://www.rfc-editor.org/info/rfc2918>>.
- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <<http://www.rfc-editor.org/info/rfc4684>>.
- [RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <<http://www.rfc-editor.org/info/rfc5291>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.
- [RFC7313] Patel, K., Chen, E., and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", RFC 7313, DOI 10.17487/RFC7313, July 2014, <<http://www.rfc-editor.org/info/rfc7313>>.
- [RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <<http://www.rfc-editor.org/info/rfc7606>>.

15.2. Information 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>>.
- [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>>.

- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<http://www.rfc-editor.org/info/rfc7752>>.

Authors' Addresses

Keyur Patel
Cisco Systems
821 Alder Drive
Milpitas, CA 95035
USA

Email: keyupate@cisco.com

Aamod Vyavaharkar
Cisco Systems
821 Alder Drive
Milpitas, CA 95035
USA

Email: avyavaha@cisco.com

Niloofar Fazlollahi
Cisco Systems
821 Alder Drive
Milpitas, CA 95035
USA

Email: nifazlol@cisco.com

Tony Przygienda
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: prz@juniper.net

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: May 21, 2020

S. Litkowski
Individual
A. Simpson
Nokia
K. Patel
Arrcus, Inc
J. Haas
Juniper Networks
L. Yong
Huawei
November 18, 2019

Applying BGP flowspec rules on a specific interface set
draft-ietf-idr-flowspec-interfaceset-05

Abstract

The BGP Flow Specification (flowspec) Network Layer Reachability Information (BGP NLRI) extension (draft-ietf-idr-rfc5575bis) is used to distribute traffic flow specifications into BGP. The primary application of this extension is the distribution of traffic filtering policies for the mitigation of distributed denial of service (DDoS) attacks.

By default, flow specification filters are applied on all forwarding interfaces that are enabled for use by the BGP flowspec extension. A network operator may wish to apply a given filter selectively to a subset of interfaces based on an internal classification scheme. Examples of this include "all customer interfaces", "all peer interfaces", "all transit interfaces", etc.

This document defines BGP Extended Communities (RFC4360) that allow such filters to be selectively applied to sets of forwarding interfaces sharing a common group identifier. The BGP Extended Communities carrying this group identifier are referred to as the BGP Flowspec "interface-set" Extended Communities.

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

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 <https://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 May 21, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Use case	3
2. Interface specific filtering using BGP flowspec	3
3. Interface-set extended community	5
4. Scaling of per-interface rules	6
5. Deployment Considerations	6
5.1. Add-Paths	6
5.2. Inter-domain Considerations	6
6. Security Considerations	7
7. Acknowledgements	7
8. IANA Considerations	7
8.1. FlowSpec Transitive Extended Communities	7
8.2. FlowSpec Non-Transitive Extended Communities	7
8.3. FlowSpec interface-set Extended Community	8
8.4. Allocation Advice to IANA	8

9. Normative References	8
Authors' Addresses	9

1. Use case

While a network may provide connectivity to a homogenous class of users, it often provides connectivity to different groups of users. The nature of these different groups, and how they're classified, varies based on the purpose of the network. In an enterprise network, connectivity may exist between data centers, offices, and external connectivity. In a virtual private networking (VPN) network, it may consist of customers in different sites connected through a VPN, the provider core network, and external networks such as the Internet. In a traditional Internet service provider (ISP) network, the network may consist of points of presence (POPs), internal infrastructure networks, customer networks, peer networks, and transit networks.

The BGP flowspec extension permits traffic filters to be distributed to routers throughout a network. However, these filters often should not be uniformly applied to all network interfaces. As an example, a rate-limiting filter applied to the SMTP protocol may be applied to customer networks, but not other networks. Similarly, a DDoS attack on the SSH protocol may be deemed appropriate to drop at upstream peering routers but not customer routers.

By default, BGP flowspec filters are applied at all interfaces that permit flowspec filters to be installed. What is needed is a way to selectively apply those filters to subsets of interfaces in a network.

2. Interface specific filtering using BGP flowspec

The uses case detailed above require application of different BGP flowspec rules on different sets of interfaces.

We propose to introduce, within BGP flowspec, a traffic filtering scope that identifies a group of interfaces where a particular filter should be applied. Identification of interfaces within BGP flowspec will be done through group identifiers. A group identifier marks a set of interfaces sharing a common administrative property. Like a BGP community, the group identifier itself does not have any significance. It is up to the network administrator to associate a particular meaning to a group identifier value (e.g. group ID#1 associated to Internet customer interfaces). The group identifier is a local interface property. Any interface may be associated with one or more group identifiers using manual configuration.

When a filtering rule advertised through BGP flowspec must be applied only to particular sets of interfaces, the BGP flowspec BGP UPDATE will contain the identifiers associated with the relevant sets of interfaces. In addition to the group identifiers, it will also contain the direction the filtering rule must be applied in (see Section 3).

Configuration of group identifiers associated to interfaces may change over time. An implementation **MUST** ensure that the filtering rules (learned from BGP flowspec) applied to a particular interface are always updated when the group identifier mapping is changing.

As an example, we can imagine the following design :

- o Internet customer interfaces are associated with group-identifier 1.
- o VPN customer interfaces are associated with group-identifier 2.
- o All customer interfaces are associated with group-identifier 3.
- o Peer interfaces are associated with group-identifier 4.
- o Transit interfaces are associated with group-identifier 5.
- o All external provider interfaces are associated with group-identifier 6.
- o All interfaces are associated with group-identifier 7.

If the service provider wants to deploy a specific inbound filtering on external provider interfaces only, the provider can send the BGP flow specification using group-identifier 6 for the inbound direction.

There are some cases where nodes are dedicated to specific functions (Internet peering, Internet Edge, VPN Edge, Service Edge ...), in this kind of scenario, there is an interest for a constrained distribution of filtering rules that are using the interface specific filtering. Without the constrained route distribution, all nodes will received all the filters even if they are not interested in those filters. Constrained route distribution of flowspec filters would allow for a more optimized distribution.

3. Interface-set extended community

This document proposes a new BGP Route Target extended community called the "flowspec interface-set". This document expands the definition of the Route Target extended community to allow a new value of high order octet (Type field) to be 0x07 for the transitive flowspec interface-set extended community, or 0x47 for the non-transitive flowspec interface-set extended community. These are in addition to the values specified in [RFC4360].

This new BGP Route Target extended community is encoded as follows :

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| 0x07 or 0x47 |      0x02      | Autonomous System Number  :
+-----+-----+-----+-----+-----+-----+-----+-----+
: AS Number (cont.) | O | I | Group Identifier |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The flags are :

- o 0 : if set, the flow specification rule MUST be applied in outbound direction to the interface set referenced by the following group-identifier.
- o I : if set, the flow specification rule MUST be applied in inbound direction to the interface set referenced by the following group-identifier.

Both flags can be set at the same time in the interface-set extended community leading to flow rule to be applied in both directions. An interface-set extended community with both flags set to zero MUST be treated as an error and as consequence, the flowspec update MUST be discarded. As having no direction indicated as no sense, there is no need to propagate the filter informations in the network.

The Group Identifier is encoded as a 14-bit number, values 0..16383.

Multiple instances of the interface-set extended community may be present in a BGP update. This may occur if the flowspec rule needs to be applied to multiple sets of interfaces.

Multiple instances of the extended community in a BGP update MUST be interpreted as a "OR" operation. For example, if a BGP UPDATE

contains two interface-set extended communities with group ID 1 and group ID 2, the filter would need to be installed on interfaces belonging to Group ID 1 or Group ID 2.

Similar to using a Route Target extended community, route distribution of flowspec NLRI with interface-set extended communities may be subject to constrained distribution as defined in [RFC4684].

4. Scaling of per-interface rules

In the absence of an interface-set extended community, a flowspec filter is applied to all flowspec enabled interfaces. When interface-set extended communities are present, different interfaces may have different filtering rules, with different terms and actions. These differing rules may make it harder to share forwarding instructions within the forwarding plane.

Flowspec implementations supporting the interface-set extended community SHOULD take care to minimize the scaling impact in such circumstances. How this is accomplished is out of the scope of this document.

5. Deployment Considerations

5.1. Add-Paths

There are some cases where a particular BGP flowspec NLRI may be advertised to different interface groups with a different action. For example, a service provider may want to discard all ICMP traffic from customer interfaces to infrastructure addresses and want to rate-limit the same traffic when it comes from some internal platforms. These particular cases require ADD-PATH ([RFC7911]) to be deployed in order to ensure that all paths (NLRI+interface-set group-id+actions) are propagated within the BGP control plane. Without ADD-PATH, only a single "NLRI+interface-set group-id+actions" will be propagated, so some filtering rules will never be applied.

5.2. Inter-domain Considerations

The Group Identifier used by the interface-set extended community has local significance to its provisioning Autonomous System. While [I-D.ietf-idr-rfc5575bis] permits inter-as advertisement of flowspec NLRI, care must be taken to not accept these communities when they would result in unacceptable filtering policies.

Filtering of interface-set extended communities at Autonomous System border routers (ASBRs) may thus be desirable.

Note that the default behavior without the interface-set feature would to have been to install the flowspec filter on all flowspec enabled interfaces.

6. Security Considerations

This document extends the Security Considerations of [I-D.ietf-idr-rfc5575bis] by permitting flowspec filters to be selectively applied to subsets of network interfaces in a particular direction. Care must be taken to not permit the inadvertent manipulation of the interface-set extended community to bypass expected traffic manipulation.

7. Acknowledgements

Authors would like to thanks Wim Hendrickx and Robert Raszuk for their valuable comments.

8. IANA Considerations

8.1. FlowSpec Transitive Extended Communities

This document requests a new type from the "BGP Transitive Extended Community Types" extended community registry from the First Come First Served range. This type name shall be 'FlowSpec Transitive Extended Communities'. IANA has assigned the value 0x07 to this type.

This document requests creation of a new registry called "FlowSpec Transitive Extended Community Sub-Types". This registry contains values of the second octet (the "Sub-Type" field) of an extended community when the value of the first octet (the "Type" field) is the value allocated in this document. The registration procedure for values in this registry shall be First Come First Served.

8.2. FlowSpec Non-Transitive Extended Communities

This document requests a new type from the "BGP Non-Transitive Extended Community Types" extended community registry from the First Come First Served range. This type name shall be 'FlowSpec Non-Transitive Extended Communities'. IANA has assigned the value 0x47 to this type.

This document requests creation of a new registry called "FlowSpec Non-Transitive Extended Community Sub-Types". This registry contains values of the second octet (the "Sub-Type" field) of an extended community when the value of the first octet (the "Type" field) is the

value allocated in this document. The registration procedure for values in this registry shall be First Come First Served.

8.3. FlowSpec interface-set Extended Community

Within the two new registries above, this document requests a new subtype (suggested value 0x02). This sub-type shall be named "interface-set", with a reference to this document.

8.4. Allocation Advice to IANA

IANA is requested to allocate the values of the FlowSpec Transitive and Non-Transitive Extended Communities such that their values are identical when ignoring the second high-order bit (Transitive). See section 2, [RFC4360].

It is suggested to IANA that, when possible, allocations from the FlowSpec Transitive/Non-Transitive Extended Community Sub-Types registries are made for transitive or non-transitive versions of features (section 2, [RFC4360]) that their code point in both registries is identical.

9. Normative References

- [I-D.ietf-idr-rfc5575bis]
Loibl, C., Hares, S., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", draft-ietf-idr-rfc5575bis-17 (work in progress), June 2019.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <<https://www.rfc-editor.org/info/rfc4360>>.
- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <<https://www.rfc-editor.org/info/rfc4684>>.

- [RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder,
"Advertisement of Multiple Paths in BGP", RFC 7911,
DOI 10.17487/RFC7911, July 2016,
<<https://www.rfc-editor.org/info/rfc7911>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174,
May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

Authors' Addresses

Stephane Litkowski
Individual

Email: slitkows.ietf@gmail.com

Adam Simpson
Nokia

Email: adam.1.simpson@nokia.com

Keyur Patel
Arrcus, Inc

Email: keyur@arrcus.com

Jeffrey Haas
Juniper Networks

Email: jhaas@juniper.net

Lucy Yong
Huawei

Email: lucy.yong@huawei.com

INTERNET-DRAFT
Intended Status: Proposed Standard

W. Hao
Huawei Technologies
D. Eastlake
Futurewei Technologies
S. Litkowski
Cisco Systems
S. Zhuang
Huawei Technologies
April 18, 2022

Expires: October 17, 2022

BGP Dissemination of L2 Flow Specification Rules
draft-ietf-idr-flowspec-l2vpn-19

Abstract

This document defines a Border Gateway Protocol (BGP) Flow Specification (flowspec) extension to disseminate Ethernet Layer 2 (L2) and Layer 2 Virtual Private Network (L2VPN) traffic filtering rules either by themselves or in conjunction with L3 flowspecs. AFI/SAFI 6/133 and 25/134 are used for these purposes. New component types and two extended communities are also defined.

Status of This Document

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

Distribution of this document is unlimited. Comments should be sent to the authors or the IDR Working Group mailing list <idr@ietf.org>.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at <https://www.ietf.org/lid-abstracts.html>. The list of Internet-Draft Shadow Directories can be accessed at <https://www.ietf.org/shadow.html>.

Table of Contents

1. Introduction.....	3
1.1 Terminology.....	4
2. Layer 2 Flow Specification Encoding.....	5
2.1 L2 Component Types.....	6
2.1.1 Type 1 - Ethernet Type (EtherType).....	6
2.1.2 Type 2 - Source MAC.....	7
2.1.3 Type 3 - Destination MAC.....	7
2.1.4 Type 4 - DSAP (Destination Service Access Point).....	7
2.1.5 Type 5 - SSAP (Source Service Access Point).....	7
2.1.6 Type 6 - Control field in LLC.....	8
2.1.7 Type 7 - SNAP.....	8
2.1.8 Type 8 - VLAN ID.....	8
2.1.9 Type 9 - VLAN PCP.....	8
2.1.10 Type 10 - Inner VLAN ID.....	9
2.1.11 Type 11 - Inner VLAN PCP.....	9
2.1.12 Type 12 - VLAN DEI.....	9
2.1.13 Type 13 - Inner VLAN DEI.....	10
2.1.14 Type 14 - Source MAC Special Bits.....	10
2.1.15 Type 15 - Destination MAC Special Bits.....	10
2.2 Order of Traffic Filtering Rules.....	10
3. L2VPN Flow Specification Encoding in BGP.....	12
3.1 Order of L2VPN Filtering Rules.....	12
4. Ethernet Flow Specification Traffic Actions.....	13
4.1 VLAN-action.....	13
4.2 TPID-action.....	15
5. Flow Spec Validation.....	16
6. IANA Considerations.....	17
7. Security Considerations.....	19
8. Acknowledgements.....	19
9. Contributors.....	19
Normative References.....	20
Informative References.....	21
Authors' Addresses.....	22

1. Introduction

Border Gateway Protocol (BGP) Flow Specification [RFC8955] (flowspec) is an extension to BGP that supports the dissemination of traffic flow specifications and resulting actions to be taken on packets in a specified flow. It leverages the BGP Control Plane to simplify the distribution of ACLs (Access Control Lists). Using the Flow Specification extension new filter rules can be injected to all BGP peers simultaneously without changing router configuration. A typical application is to automate the distribution of traffic filter lists to routers for DDoS (Distributed Denial of Service) mitigation, access control, and similar applications.

BGP Flow Specification [RFC8955] defines a BGP Network Layer Reachability Information (NLRI) format used to distribute traffic flow specification rules. The NLRI for (AFI=1, SAFI=133) specifies IPv4 unicast filtering. The NLRI for (AFI=1, SAFI=134) specifies IPv4 BGP/MPLS VPN filtering [RFC7432]. The Flow Specification match part defined in [RFC8955] only includes L3/L4 information like IPv4 source/destination prefix, protocol, ports, and the like, so traffic flows can only be filtered based on L3/L4 information. This has been extended by [RFC8956] to cover IPv6 (AFI=2) L3/L4.

Layer 2 Virtual Private Networks (L2VPNs) have been deployed in an increasing number of networks. Such networks also have requirements to deploy BGP Flow Specification to mitigate DDoS attack traffic. Within an L2VPN network, both IP and non-IP Ethernet traffic may exist. For IP traffic filtering, the VPN Flow Specification rules defined in [RFC8955] and/or [RFC8956], which include match criteria and actions, can still be used. For non-IP Ethernet traffic filtering, Layer 2 related information like source/destination MAC and VLAN must be considered.

There are different kinds of L2VPN networks like EVPN [RFC7432], BGP VPLS [RFC4761], LDP VPLS [RFC4762] and border gateway protocol (BGP) auto discovery [RFC6074]. Because the Flow Specification feature relies on the BGP protocol to distribute traffic filtering rules, it can only be incrementally deployed in those L2VPN networks where BGP has already been used for auto discovery and/or signaling purposes such as BGP-based VPLS [RFC4761], EVPN, and LDP-based VPLS [RFC4762] with BGP auto-discovery [RFC6074].

This document defines new flowspec component types and two new extended communities to support L2 and L2VPN flowspec applications. The flowspec rules can be enforced on all border routers or on some interface sets of the border routers. SAFI=133 in [RFC8955] and [RFC8956] is extended for AFI=6 as specified in Section 2 to cover L2 traffic filtering information and in Section 3 SAFI=134 is extended for AFI=25 to cover the L2VPN environment.

1.1 Terminology

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

The following acronyms and terms are used in this document:

AFI - Address Family Identifier

ACL - Access Control List

DDoS - Distributed Denial of Service

DEI - Drop Eligible Indicator

EVPN - Ethernet VPN [RFC7432]

flowspec - BGP Flow Specification

L2 - Layer 2

L2VPN - Layer 2 VPN

L3 - Layer 3

L3VPN - Layer 3 VPN

NLRI - Network Layer Reachability Information

PCP - Priority Code Point [802.1Q]

SAFI - Subsequent Address Family Identifier

TPID - Tag Protocol ID, typically a VLAN ID

VLAN - Virtual Local Area Network

VPLS - Virtual Private Line Service [RFC4762]

VPN - Virtual Private Network

2. Layer 2 Flow Specification Encoding

[RFC8955] defines SAFI 133 and SAFI 134, with AFI=1, for "dissemination of IPv4 flow specification rules" and "dissemination of VPNv4 flow specification rules", respectively. [RFC8956] extends [RFC8955] to also allow AFI=2 thus making it applicable to both IPv4 and IPv6 applications. This document further extends SAFI=133 for AFI=6 and SAFI=134 for AFI=25 to make them applicable to L2 and L2VPN applications. This document also provides for the optional combination of L3 flow specifications with these L2 flow specifications.

This section specifies the L2 flowspec for AFI=6/SAFI=133. To simplify assignments, a new registry is used for L2 flowspec. Since it is frequently desirable to also filter on L3/L4 fields, provision is made for their inclusion along with an indication of the L3 protocol involved (IPv4 or IPv6).

The NLRI part of the MP_REACH_NLRI and MP_UNREACH_NLRI is encoded as a 1- or 2-octet total NLRI length field followed by several fields as described below.

total-length (0xnn or 0xfnnn)	2 or 3 octets
L3-AFI	2 octets
L2-length (0xnn or 0xfnnn)	2 or 3 octets
NLRI-value	variable

Figure 1: Flow Specification NLRI for L2

The fields show in Figure 1 are further specified below:

total-length: The length of the subsequent fields (L3 AFI, L2-length, and NLRI-value) encoded as provided in Section 4.1 of [RFC8955]. If this field is less than 4, which is the minimum valid value, then the NLRI is malformed in which case a NOTIFICATION message is sent and the BGP connection closed as provided in Section 6.3 of [RFC4271].

L3-AFI: If no L3/L4 filtering is desired, this two octet field MUST be zero which is a reserved AFI value. Otherwise L3-AFI indicates the L3 protocol involved by giving its AFI (0x0001 for IPv4 or 0x0002 for IPv6). If the receiver does not understand the value of the L3-AFI field, the MP_REACH or MP_UNREACH attribute is ignored.

L2-length: The length of the L2 components at the beginning of the NLRI-value field encoded as provided in Section 4.1 of [RFC8955]. If the value of this field indicates that the L2 components extend beyond the total-length, the NLRI is malformed in which case a NOTIFICATION message is sent and the BGP connection closed as provided in Section 6.3 of [RFC4271]. N2-length MAY be zero although, in that case, it would have been more efficient to encode the attribute as an L3 Flow spec unless it is desired to apply an L2 action (see Section 4). A null L2 flowspec always matches.

NLRI-value: This consists of the L2 flowspec, of length L2-length, followed by an optionally present L3 flowspec. The result can be treated in most ways as a single flowspec, matching the intersection (AND) of all the components except that the components in the initial L2 region are interpreted as L2 components and the remainder as L3 components per the L3-AFI field. This is necessary because there are different registries for the L2, L3 IPv4, and L3 IPv6 component types. If the L3 flowspec is null (length zero), it always matches.

2.1 L2 Component Types

The L2 flowspec portion of the NLRI-value consists of flowspec components as in [RFC8955] but using L2 components and types as specified below. All components start with a type octet followed by a length octet followed by any additional information needed. The length octet gives the length, in octets, of the information after the length octet. This structure applies to all new components to be defined in the L2 Flow-spec Component Registry (see Section 6) and to all existing components except Types 2 and 3 where the length is in bits.

2.1.1 Type 1 - Ethernet Type (EtherType)

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match the two-octet EtherType field. op is encoded as specified in Section 4.2.1.1 of [RFC8955]. Values are encoded as 2-octet quantities. Ethernet II framing defines the two-octet Ethernet Type (EtherType) field in an Ethernet frame, preceded by destination and source MAC addresses, that identifies an upper layer protocol encapsulating the frame data. The match fails if LLC encoding is being used rather than EtherType encoding.

2.1.2 Type 2 - Source MAC

Encoding: <type (1 octet), MAC Prefix length (1 octet), MAC Prefix>

Defines the source MAC Address prefix to match encoded as in BGP UPDATE messages [RFC4271]. Prefix length is in bits and the MAC Prefix is fill out with from 1 to 7 padding bits so that it is an integer number of octets. These padding bits are ignored for matching purposes.

2.1.3 Type 3 - Destination MAC

Encoding: <type (1 octet), MAC Prefix length (1 octet), MAC Prefix>

Defines the destination MAC Address to match encoded as in BGP UPDATE messages [RFC4271]. Prefix length is in bits and the MAC Prefix is fill out with from 1 to 7 padding bits so that it is an integer number of octets. These padding bits are ignored for matching purposes.

2.1.4 Type 4 - DSAP (Destination Service Access Point)

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match the 1-octet DSAP in the IEEE 802.2 LLC (Logical Link Control Header). Values are encoded as 1-octet quantities. op is encoded as specified in Section 4.2.1.1 of [RFC8955]. The match fails if EtherType L2 header encoding is being used rather than LLC encoding.

2.1.5 Type 5 - SSAP (Source Service Access Point)

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match the 1-octet SSAP in the IEEE 802.2 LLC. Values are encoded as 1-octet quantities. op is encoded as specified in Section 4.2.1.1 of [RFC8955]. The match fails if EtherType L2 header encoding is being used rather than LLC encoding.

2.1.6 Type 6 - Control field in LLC

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match the 1-octet control field in the IEEE 802.2 LLC. Values are encoded as 1-octet quantities. op is encoded as specified in Section 4.2.1.1 of [RFC8955]. The match fails if EtherType L2 header encoding is being used rather than LLC encoding.

2.1.7 Type 7 - SNAP

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match 5-octet SNAP (Sub-Network Access Protocol) field. Values are encoded as 8-octet quantities with the zero padded SNAP left justified. op is encoded as specified in Section 4.2.1.1 of [RFC8955]. The match fails if EtherType L2 header encoding is being used rather than LLC encoding.

2.1.8 Type 8 - VLAN ID

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match VLAN ID. Values are encoded as 2-octet quantities, where the four most significant bits are set to zero and ignored for matching and the 12 least significant bits contain the VLAN value. op is encoded as specified in Section 4.2.1.1 of [RFC8955].

In the virtual local-area network (VLAN) stacking case, the VLAN ID is the outer VLAN ID.

2.1.9 Type 9 - VLAN PCP

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match 3-bit VLAN PCP (priority code point) fields [802.1Q]. Values are encoded using a single octet, where the five most significant bits are set to zero and ignored for matching and the three least significant bits contain the VLAN PCP value. op is encoded as specified in Section 4.2.1.1 of [RFC8955].

In the virtual local-area network (VLAN) stacking case, the VLAN PCP is part of the outer VLAN tag.

2.1.10 Type 10 - Inner VLAN ID

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match the inner VLAN ID for virtual local-area network (VLAN) stacking or Q-in-Q use. Values are encoded as 2-octet quantities, where the four most significant bits are set to zero and ignored for matching and the 12 least significant bits contain the VLAN value. op is encoded as specified in Section 4.2.1.1 of [RFC8955].

In the single VLAN case, this component type MUST NOT be used. If it appears the match will fail.

2.1.11 Type 11 - Inner VLAN PCP

Encoding: <type (1 octet), length (1 octet), [op, value]+>

Defines a list of {operation, value} pairs used to match 3-bit inner VLAN PCP fields [802.1Q] for virtual local-area network (VLAN) stacking or Q-in-Q use. Values are encoded using a single octet, where the five most significant bits are set to zero and ignored for matching and the three least significant bits contain the VLAN PCP value. op is encoded as specified in Section 4.2.1.1 of [RFC8955].

In the single VLAN case, this component type MUST NOT be used. If it appears the match will fail.

2.1.12 Type 12 - VLAN DEI

Encoding: <type (1 octet), length (1 octet), op (1 octet)>

This type tests the DEI (Drop Eligible Indicator) bit in the VLAN tag. If op is zero, it matches if and only if the DEI bit is zero. If op is non-zero, it matches if and only if the DEI bit is one.

In the virtual local-area network (VLAN) stacking case, the VLAN DEI is part of the outer VLAN tag.

2.1.13 Type 13 - Inner VLAN DEI

Encoding: <type (1 octet), length (1 octet), op (1 octet)>

This type tests the DEI bit in the inner VLAN tag. If op is zero, it matches if and only if the DEI bit is zero. If op is non-zero, it matches if and only if the DEI bit is one.

In the single VLAN case, this component type MUST NOT be used. If it appears the match will fail.

2.1.14 Type 14 - Source MAC Special Bits

Encoding: <type (1 octet), length (1 octet), op (1 octet)>

This type tests the bottom nibble of the top octet of the Source MAC address. The two low order bits of that nibble have long been the local bit (0x2) and the group addressed bit (0x1). However, recent changes in IEEE 802 have divided the local address space into 4 quadrants specified by the next two bits (0x4 and 0x8) [RFC7042bis]. This flowspec component permits testing, for example, that a MAC is group addressed or is a local address in a particular quadrant. The encoding is as given in Section 4.2.1.2 of [RFC8955].

2.1.15 Type 15 - Destination MAC Special Bits

Encoding: <type (1 octet), length (1 octet), op (1 octet)>

As discussed in Section 2.1.14 but for the Destination MAC Address special bits.

2.2 Order of Traffic Filtering Rules

The existing rules in Section 5.1 of [RFC8955] and in [RFC8956] for the ordering of traffic filtering are extended as follows:

L2 flowspecs (AFI = 6, 25) take precedence over L3 flowspecs (AFI = 1, 2). Between two L2 flowspecs, precedence of the L2 portion is determined as specified in this section after this paragraph. If the L2 flowspec L2 portions are the same and the L3-AFI is nonzero, then the L3 portions are compared as specified in [RFC8955] or [RFC8956] as appropriate. Note: if the L3-AFI fields are different between two L2 flowspecs, they will never match the same packet so it will not be necessary to prioritize two flowspecs with different L3-AFI values.

The original definition for the order of traffic filtering rules can be reused for L2 with new consideration for the MAC Address offset. As long as the offsets are equal, the comparison is the same, retaining longest-prefix-match semantics. If the offsets are not equal, the lowest offset has precedence, as this flow matches the most significant bit.

Pseudocode:

```

flow_rule_L2_cmp (a, b)
{
    comp1 = next_component(a);
    comp2 = next_component(b);
    while (comp1 || comp2) {
        // component_type returns infinity on end-of-list
        if (component_type(comp1) < component_type(comp2)) {
            return A_HAS_PRECEDENCE;
        }
        if (component_type(comp1) > component_type(comp2)) {
            return B_HAS_PRECEDENCE;
        }

        if (component_type(comp1) == MAC_DESTINATION || MAC_SOURCE) {
            common = MIN(MAC Address length (comp1),
                        MAC Address length (comp2));
            cmp = MAC Address compare(comp1, comp2, common);
            // not equal, lowest value has precedence
            // equal, longest match has precedence
        } else {
            common =
                MIN(component_length(comp1), component_length(comp2));
            cmp = memcmp(data(comp1), data(comp2), common);
            // not equal, lowest value has precedence
            // equal, longest string has precedence
        }
    }
    return EQUAL;
}

```

3. L2VPN Flow Specification Encoding in BGP

The NLRI format for AFI=25/SAFI=134 (L2VPN), as with the other VPN flowspec AFI/SAFI pairs, is the same as the non-VPN Flow-Spec but with the addition of a Route Distinguisher to identify the VPN to which the flowspec is to be applied.

In addition, the IANA entry for SAFI 134 is slightly generalized as specified at the beginning of Section 6.

The L2VPN NLRI format is as follows:

total-length (0xnn or 0xfnnn)	2 or 3 octets
Route Distinguisher	8 octets
L3-AFI	2 octets
L2-length (0xnn or 0xfnnn)	2 or 3 octets
NLRI-value	variable

Figure 2: Flow Specification NLRI for L2VPN

The fields in Figure 2, other than the Route Distinguisher, are encoded as specified in Section 2 except that the minimum value for total-length is 12.

Flow specification rules received via this NLRI apply only to traffic that belongs to the VPN instance(s) into which it is imported. Flow rules are accepted as specified in Section 5.

3.1 Order of L2VPN Filtering Rules

The order between L2VPN filtering rules is determined as specified in Section 2.2. Note that if the Route Distinguisher is different between two L2VPN filtering rules, they will never both match the same packet so they need not be prioritized.

4. Ethernet Flow Specification Traffic Actions

The default action for an L2 traffic filtering flowspec is to accept traffic that matches that particular rule. The following extended community values per [RFC8955] can be used to specify particular actions in an L2 VPN network:

type	extended community	encoding
0x8006	traffic-rate	2-octet as#, 4-octet float
0x8007	traffic-action	bitmask
0x8008	redirect	6-octet Route Target
0x8009	traffic-marking	DSCP value

Redirect: The action should be redefined to allow the traffic to be redirected to a MAC or IP VRF routing instance that lists the specified route-target in its import policy.

Besides the above extended communities, this document also specifies the following BGP extended communities for Ethernet flows to extend [RFC8955]:

type	extended community	encoding
TBD1	VLAN-action	bitmask
TBD2	TPID-action	bitmask

4.1 VLAN-action

The VLAN-action extended community, as shown in the diagram below, consists of 6 octets that include action Flags, two VLAN IDs, and the associated PCP and DEI values. The action Flags fields are further divided into two parts which correspond to the first action and the second action respectively. Bit 0 to bit 7 give the first action while bit 8 to bit 15 give the second action. The bits of PO, PU, SW, RI and RO in each part represent the action of Pop, Push, Swap, Rewrite inner VLAN and Rewrite outer VLAN respectively. Through this method, more complicated actions also can be represented in a single VLAN-action extended community, such as SwapPop, PushSwap, etc. For example, SwapPop action is the sequence of two actions, the first action is Swap and the second action is Pop.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
PO1	PU1	SW1	RI1	RO1	Resv			PO2	PU2	SW2	RI2	RO2	Resv		
VLAN ID1												PCP1		DE1	
VLAN ID2												PCP2		DE2	

PO1: Pop action. If the PO1 flag is one, it indicates the outmost VLAN should be removed.

PU1: Push action. If PU1 is one, it indicates VLAN ID1 will be added, the associated PCP and DEI are PCP1 and DE1.

SW1: Swap action. If the SW1 flag is one, it indicates the outer VLAN and inner VLAN should be swapped.

PO2: Pop action. If the PO2 flag is one, it indicates the outmost VLAN should be removed.

PU2: Push action. If PU2 is one, it indicates VLAN ID2 will be added, the associated PCP and DEI are PCP2 and DE2.

SW2: Swap action. If the SW2 flag is one, it indicates the outer VLAN and inner VLAN should be swapped.

RI1 and RI2: Rewrite inner VLAN action. If the RIX flag is one (where "x" is "1" or "2"), it indicates the inner VLAN should be replaced by a new VLAN where the new VLAN is VLAN IDx and the associated PCP and DEI are PCPx and DEx. If the VLAN IDx is 0, the action is to only modify the PCP and DEI value of the inner VLAN.

RO1 and RO2: Rewrite outer VLAN action. If the ROx flag is one (where "x" is "1" or "2"), it indicates the outer VLAN should be replaced by a new VLAN where the new VLAN is VLAN IDx and the associated PCP and DEI are PCPx and DEx. If the VLAN IDx is 0, the action is to only modify the PCP and DEI value of the outer VLAN.

Resv: Reserved for future use. MUST be sent as zero and ignored on receipt.

Giving an example below: if the action of PUSH Inner VLAN 10 with PCP value 5 and DEI value 0 and PUSH Outer VLAN 20 with PCP value 6 and DEI value 0 is needed, the format of the VLAN-action extended community is as follows:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10												1	0	1	0
20												1	1	0	0

4.2 TPID-action

The TPID-action extended community consists of 6 octets which includes the fields of action Flags, TP ID1 and TP ID2.

0	15
TI TO	Resv
	TP ID1
	TP ID2

TI: Mapping inner TP ID action. If the TI flag is one, it indicates the inner TP ID should be replaced by a new TP ID, the new TP ID is TP ID1.

TO: Mapping outer TP ID action. If the TO flag is one, it indicates the outer TP ID should be replaced by a new TP ID, the new TP ID is TP ID2.

Resv: Reserved for future use. MUST be sent as zero and ignored on receipt.

5. Flow Spec Validation

Flow Specifications received over AFI=25/SAFI=134 are validated against routing reachability received over AFI=25/SAFI=128 as modified to conform to [RFC9117].

6. IANA Considerations

IANA is requested to change the description for SAFI 134 [RFC8955] to read as follows and to change the reference for it to [this document]:

134 VPN dissemination of flow specification rules

IANA is requested to create an L2 Flow Specification Component Type registry on the Flow Spec Component Types registries web page as follows:

Name: L2 Flow Specification Component Types

Reference: [this document]

Registration Procedures:

0 Reserved
 1-127 Specification Required
 128-255 First Come First Served

Initial contents:

type	Reference	description
0	[this document]	Reserved
1	[this document]	Ethernet Type
2	[this document]	Source MAC
3	[this document]	Destination MAC
4	[this document]	DSAP in LLC
5	[this document]	SSAP in LLC
6	[this document]	Control field in LLC
7	[this document]	SNAP
8	[this document]	VLAN ID
9	[this document]	VLAN PCP
10	[this document]	Inner VLAN ID
11	[this document]	Inner VLAN PCP
12	[this document]	VLAN DEI
13	[this document]	Inner VLAN DEI
14	[this document]	Source MAC Special Bits
15	[this document]	Destination MAC Special Bits
16-254	[this document]	unassigned
255	[this document]	Reserved

IANA is requested to assign two values from the "BGP Extended Communities Type - extended, transitive" registry [suggested value provided in square brackets]:

Type value	Name	Reference
TBD1[0x080A]	Flow spec VLAN action	[this document]
TBD2[0x080B]	Flow spec TPID action	[this document]

7. Security Considerations

For General BGP Flow Specification Security Considerations, see [RFC8955].

VLAN tagging identifies Layer 2 communities which are commonly expected to be isolated except when higher layer connection is provided, such as Layer 3 routing. Thus, the ability of the flowspec VLAN action to change the VLAN ID in a frame might compromise security.

8. Acknowledgements

The authors wish to acknowledge the important contributions and suggestions of the following:

Hannes Gredler, Xiaohu Xu, Zhenbin Li, Lucy Yong, and Feng Dong.

9. Contributors

Qiandeng Liang
Huawei Technologies
101 Software Avenue, Yuhuatai District
Nanjing 210012
China

Email: liangqiandeng@huawei.com

Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <<https://www.rfc-editor.org/info/rfc4761>>.
- [RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <<https://www.rfc-editor.org/info/rfc4762>>.
- [RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, DOI 10.17487/RFC6074, January 2011, <<https://www.rfc-editor.org/info/rfc6074>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8955] Loibl, C., Hares, S., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", RFC 8955, DOI 10.17487/RFC8955, December 2020, <<https://www.rfc-editor.org/info/rfc8955>>.
- [RFC8956] Loibl, C., Ed., Raszuk, R., Ed., and S. Hares, Ed., "Dissemination of Flow Specification Rules for IPv6", RFC 8956, DOI 10.17487/RFC8956, December 2020, <<https://www.rfc-editor.org/info/rfc8956>>.
- [RFC9117] Uttaro, J., Alcaide, J., Filsfils, C., Smith, D., and P. Mohapatra, "Revised Validation Procedure for BGP Flow Specifications", RFC 9117, DOI 10.17487/RFC9117, August 2021, <<https://www.rfc-editor.org/info/rfc9117>>.

Informative References

- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7042bis] Eastlake, D., and J. Abley, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", draft-eastlake-rfc7042bis, work in progress, March 2021.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Email: haoweiguo@huawei.com

Donald E. Eastlake, 3rd
Futurewei Technologies
2386 Panoramic Circle
Apopka, FL 32703
USA

Tel: +1-508-333-2270
Email: d3e3e3@gmail.com

Stephane Litkowski
Cisco Systems, Inc.

Email: slitkows.ietf@gmail.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Copyright, Disclaimer, and Additional IPR Provisions

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

IDR
Internet-Draft
Intended status: Standards Track
Expires: July 9, 2017

J. Heitz, Ed.
Cisco
J. Snijders, Ed.
NTT
K. Patel
Arrcus
I. Bagdonas
Equinix
N. Hilliard
INEX
January 5, 2017

BGP Large Communities
draft-ietf-idr-large-community-12

Abstract

This document describes the BGP Large Communities attribute, an extension to BGP-4. This attribute provides a mechanism to signal opaque information within separate namespaces to aid in routing management. The attribute is suitable for use with all Autonomous System Numbers including four-octet Autonomous System Numbers.

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 [RFC2119].

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 July 9, 2017.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. BGP Large Communities Attribute	3
3. Aggregation	4
4. Canonical Representation	4
5. Error Handling	4
6. Security Considerations	5
7. Implementation status - RFC EDITOR: REMOVE BEFORE PUBLICATION	5
8. IANA Considerations	6
9. Contributors	6
10. Acknowledgments	6
11. References	7
11.1. Normative References	7
11.2. Informative References	7
11.3. URIs	8
Authors' Addresses	8

1. Introduction

BGP [RFC4271] implementations typically support a routing policy language to control the distribution of routing information. Network operators attach BGP communities to routes to associate particular properties with these routes. These properties may include information such as the route origin location, or specification of a routing policy action to be taken, or one that has been taken, and is applied to all routes contained in a BGP Update Message where the Communities Attribute is included. Because BGP communities are optional transitive BGP attributes, BGP communities may be acted upon or otherwise used by routing policies in other Autonomous Systems (ASes) on the Internet.

BGP Communities attributes are a variable length attribute consisting of a set of one or more four-octet values, each of which specify a community [RFC1997]. Common use of the individual values of this attribute type split this single 32-bit value into two 16-bit values. The most significant word is interpreted as an Autonomous System Number (ASN) and the least significant word is a locally defined value whose meaning is assigned by the operator of the Autonomous System in the most significant word.

Since the adoption of four-octet ASNs [RFC6793], the BGP Communities attribute can no longer accommodate the above encoding, as a two-octet word cannot fit a four-octet ASN. The BGP Extended Communities attribute [RFC4360] is also unsuitable. The six-octet length of the Extended Community value precludes the common operational practise of encoding four-octet ASNs in both the Global Administrator and the Local Administrator sub-fields.

To address these shortcomings, this document defines a BGP Large Communities attribute encoded as an unordered set of one or more twelve-octet values, each consisting of a four-octet Global Administrator field and two four-octet operator-defined fields, each of which can be used to denote properties or actions significant to the operator of the Autonomous System assigning the values.

2. BGP Large Communities Attribute

This document defines the BGP Large Communities attribute as an optional transitive path attribute of variable length. All routes with the BGP Large Communities attribute belong to the communities specified in the attribute.

Each BGP Large Community value is encoded as a 12-octet quantity, as follows:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Global Administrator                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Local Data Part 1                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Local Data Part 2                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Global Administrator: A four-octet namespace identifier.

Local Data Part 1: A four-octet operator-defined value.

Local Data Part 2: A four-octet operator-defined value.

The Global Administrator field is intended to allow different Autonomous Systems to define BGP Large Communities without collision. This field SHOULD be an Autonomous System Number (ASN), in which case the Local Data Parts are to be interpreted as defined by the owner of the ASN. The use of Reserved ASNs (0 [RFC7607], 65535 and 4294967295 [RFC7300]) is NOT RECOMMENDED.

There is no significance to the order in which twelve-octet Large Community Attribute values are encoded in a Large Communities attribute. A BGP speaker can transmit them in any order.

Duplicate BGP Large Community values MUST NOT be transmitted. A receiving speaker MUST silently remove redundant BGP Large Community values from a BGP Large Community attribute.

3. Aggregation

If a range of routes is aggregated, then the resulting aggregate should have a BGP Large Communities attribute which contains all of the BGP Large Communities attributes from all of the aggregated routes.

4. Canonical Representation

The canonical representation of BGP Large Communities is three separate unsigned integers in decimal notation in the following order: Global Administrator, Local Data 1, Local Data 2. Numbers MUST NOT contain leading zeros; a zero value MUST be represented with a single zero. Each number is separated from the next by a single colon. For example: 64496:4294967295:2, 64496:0:0.

BGP Large Communities SHOULD be represented in the canonical representation.

5. Error Handling

The error handling of BGP Large Communities is as follows:

- o A BGP Large Communities attribute SHALL be considered malformed if the length of the BGP Large Communities Attribute value, expressed in octets, is not a non-zero multiple of 12.
- o A BGP Large Communities attribute SHALL NOT be considered malformed due solely to presence of duplicate community values.

- o A BGP UPDATE message with a malformed BGP Large Communities attribute SHALL be handled using the approach of "treat-as-withdraw" as described in section 2 [RFC7606].

The BGP Large Communities Global Administrator field may contain any value, and a BGP Large Communities attribute MUST NOT be considered malformed if the Global Administrator field contains an unallocated, unassigned or reserved ASN.

6. Security Considerations

This document does not change any underlying security issues associated with any other BGP Communities mechanism. Specifically, an AS relying on the BGP Large Communities attribute carried in BGP must have trust in every other AS in the path, as any intermediate Autonomous System in the path may have added, deleted, or altered the BGP Large Communities attribute. Specifying the mechanism to provide such trust is beyond the scope of this document.

BGP Large Communities do not protect the integrity of each community value. Operators should be aware that it is possible for a BGP speaker to alter BGP Large Community Attribute values in a BGP Update Message. Protecting the integrity of the transitive handling of BGP Large Community attributes in a manner consistent with the intent of expressed BGP routing policies falls within the broader scope of securing BGP, and is not specifically addressed here.

Network administrators should note the recommendations in Section 11 of BGP Operations and Security [RFC7454].

7. Implementation status - RFC EDITOR: REMOVE BEFORE PUBLICATION

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.

As of today these vendors have produced an implementation of BGP Large Communities:

- o Cisco IOS XR
- o ExaBGP
- o GoBGP
- o BIRD
- o OpenBGPD
- o pmacct
- o Quagga

The latest implementation news is tracked at <http://largebgpcommunities.net/> [1].

8. IANA Considerations

IANA has made an Early Allocation of the value 32 (LARGE_COMMUNITY) in the "BGP Path Attributes" registry under the "Border Gateway Protocol (BGP) Parameters" group and is now asked to make that Permanent.

9. Contributors

The following people contributed significantly to the content of the document:

John Heasley
NTT Communications
Email: heas@shrubbery.net

Adam Simpson
Nokia
Email: adam.1.simpson@nokia.com

10. Acknowledgments

The authors would like to thank Ruediger Volk, Russ White, Acee Lindem, Shyam Sethuram, Jared Mauch, Joel M. Halpern, Jeffrey Haas, Gunter van de Velde, Marco Marzetti, Eduardo Ascenco Reis, Mark Schouten, Paul Hoogsteder, Martijn Schmidt, Greg Hankins, Bertrand Duvivier, Barry O'Donovan, Grzegorz Janoszka, Linda Dunbar, Marco Davids, Gaurab Raj Upadhaya, Jeff Tantsura, Teun Vink, Adam Davenport, Theodore Baschak, Pier Carlo Chiodi, Nabeel Cocker, Ian Dickinson, Jan Baggen, Duncan Lockwood, David Farmer, Randy Bush, Wim Henderickx, Stefan Plug, Kay Rechthien, Rob Shakir, Warren Kumari,

Gert Doering, Thomas King, Mikael Abrahamsson, Wesley Steehouwer, Sander Steffann, Brad Dreisbach, Martin Millnert, Christopher Morrow, Jay Borkenhagen, Arnold Nipper, Joe Provo, Niels Bakker, Bill Fenner, Tom Daly, Ben Maddison, Alexander Azimov, Brian Dickson, Peter van Dijk, Julian Seifert, Tom Petch, Tom Scholl, Arjen Zonneveld, Remco van Mook, Adam Chappell, Jussi Peltola, Kristian Larsson, Markus Hauschild, Richard Steenbergen, David Freedman, Richard Hartmann, Geoff Huston, Mach Chen, and Alvaro Retana for their support, insightful review and comments.

11. References

11.1. 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>>.
- [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>>.
- [RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <<http://www.rfc-editor.org/info/rfc7606>>.

11.2. Informative References

- [RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <<http://www.rfc-editor.org/info/rfc1997>>.
- [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>>.
- [RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, DOI 10.17487/RFC6793, December 2012, <<http://www.rfc-editor.org/info/rfc6793>>.
- [RFC7300] Haas, J. and J. Mitchell, "Reservation of Last Autonomous System (AS) Numbers", BCP 6, RFC 7300, DOI 10.17487/RFC7300, July 2014, <<http://www.rfc-editor.org/info/rfc7300>>.

- [RFC7454] Durand, J., Pepelnjak, I., and G. Doering, "BGP Operations and Security", BCP 194, RFC 7454, DOI 10.17487/RFC7454, February 2015, <<http://www.rfc-editor.org/info/rfc7454>>.
- [RFC7607] Kumari, W., Bush, R., Schiller, H., and K. Patel, "Codification of AS 0 Processing", RFC 7607, DOI 10.17487/RFC7607, August 2015, <<http://www.rfc-editor.org/info/rfc7607>>.

11.3. URIs

- [1] <http://largebgpcommunities.net>

Authors' Addresses

Jakob Heitz (editor)
Cisco
170 West Tasman Drive
San Jose, CA 95054
USA

Email: jheitz@cisco.com

Job Snijders (editor)
NTT Communications
Theodorus Majofskistraat 100
Amsterdam 1065 SZ
The Netherlands

Email: job@ntt.net

Keyur Patel
Arrcus, Inc

Email: keyur@arrcus.com

Ignas Bagdonas
Equinix
80 Cheapside
London EC2V 6EE
United Kingdom

Email: ibagdona.ietf@gmail.com

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
IE

Email: nick@inex.ie

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 4, 2017

K. Patel
Arrcus, Inc.
A. Lindem
Cisco Systems
S. Zandi
Linkedin
G. Van de Velde
Nokia
October 31, 2016

Shortest Path Routing Extensions for BGP Protocol
draft-keyupate-idr-bgp-spf-01.txt

Abstract

Many Massively Scaled Data Centers (MSDCs) have converged on simplified layer 3 routing. Furthermore, requirements for operational simplicity have lead many of these MSDCs to converge on BGP as their single routing protocol for both their fabric routing and their Data Center Interconnect (DCI) routing. This document describes a solution which leverages BGP Link-State distribution and the Shortest Path First algorithm similar to Internal Gateway Protocols (IGPs) such as OSPF.

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 May 4, 2017.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1.	Introduction	3
1.1.	BGP Shortest Path First (SPF) Motivation	4
1.2.	Requirements Language	5
2.	BGP Peering Models	5
2.1.	BGP Single-Hop Peering on Network Node Connections	5
2.2.	BGP Peering Between Directly Connected Network Nodes	5
2.3.	BGP Peering in Route-Reflector or Controller Topology	6
3.	BGP-LS Shortest Path Routing (SPF) SAFI	6
4.	Extensions to BGP-LS	6
4.1.	Node NLRI Usage and Modifications	6
4.2.	Link NLRI Usage	7
4.3.	Prefix NLRI Usage	7
4.4.	BGP-LS Attribute Sequence-Number TLV	8
5.	Decision Process with SPF Algorithm	9
5.1.	Phase-1 BGP NLRI Selection	9
5.2.	Dual Stack Support	10
5.3.	NEXT_HOP Manipulation	10
5.4.	Error Handling	10
6.	IANA Considerations	11
7.	Security Considerations	11
7.1.	Acknowledgements	11
7.2.	Contributorss	11
8.	References	12
8.1.	Normative References	12

8.2. Information References	13
Authors' Addresses	13

1. Introduction

Many Massively Scaled Data Centers (MSDCs) have converged on simplified layer 3 routing. Furthermore, requirements for operational simplicity have lead many of these MSDCs to converge on BGP [RFC4271] as their single routing protocol for both their fabric routing and their Data Center Interconnect (DCI) routing. Requirements and procedures for using BGP are described in [RFC7938]. This document describes an alternative solution which leverages BGP-LS [RFC7752] and the Shortest Path First algorithm similar to Internal Gateway Protocols (IGPs) such as OSPF [RFC2328].

[RFC4271] defines the Decision Process that is used to select routes for subsequent advertisement by applying the policies in the local Policy Information Base (PIB) to the routes stored in its Adj-RIBs-In. The output of the Decision Process is the set of routes that are announced by a BGP speaker to its peers. These selected routes are stored by a BGP speaker in the speaker's Adj-RIBs-Out according to policy.

[RFC7752] describes a mechanism by which link-state and TE information can be collected from networks and shared with external components using BGP. This is achieved by defining NLRI carried within BGP-LS AFI and BGP-LS SAFIs. The BGP-LS extensions defined in [RFC7752] makes use of the Decision Process defined in [RFC4271].

This document augments [RFC7752] by replacing its use of the existing Decision Process. The BGP-LS-SPF and BGP-LS-SPF-VPN AFI/SAFI are introduced to insure backward compatibility. The Phase 1 and 2 decision functions of the Decision Process are replaced with the Shortest Path Algorithm (SPF) also known as the Dijkstra Algorithm. The Phase 3 decision function is also simplified since it is no longer dependent on the previous phases. This solution avails the benefits of both BGP and SPF-based IGPs. These include TCP based flow-control, no periodic link-state refresh, and completely incremental NLRI advertisement. These advantages can reduce the overhead in MSDCs where there is a high degree of Equal Cost Multi-Path (ECMPs) and the topology is very stable. Additionally, using a SPF-based computation can support fast convergence and the computation of Loop-Free Alternatives (LFAs) [RFC5286] in the event of link failures. Furthermore, a BGP based solution lends itself to multiple peering models including those incorporating route-reflectors [RFC4456] or controllers.

Support for Multiple Topology Routing (MTR) as described in [RFC4915] is an area for further study dependent on deployment requirements.

1.1. BGP Shortest Path First (SPF) Motivation

Given that [RFC7938] already describes how BGP could be used as the sole routing protocol in an MSDC, one might question the motivation for defining an alternate BGP deployment model when a mature solution exists. For both alternatives, BGP offers the operational benefits of a single routing protocol. However, BGP SPF offers some unique advantages above and beyond standard BGP distance-vector routing.

A primary advantage is that all BGP speakers in the BGP SPF routing domain will have a complete view of the topology. This will allow support of ECMP, IP fast-reroute (e.g., Loop-Free Alternatives), Shared Risk Link Groups (SRLGs), and other routing enhancements without advertisement of addition BGP paths or other extensions. In short, the advantages of an IGP such as OSPF [RFC2328] are availed in BGP.

With the simplified BGP decision process as defined in Section 5.1, NLRI changes can be disseminated throughout the BGP routing domain much more rapidly (equivalent to IGPs with the proper implementation).

Another primary advantage is a potential reduction in NLRI advertisement. With standard BGP distance-vector routing, a single link failure may impact 100s or 1000s prefixes and result in the withdrawal or re-advertisement of the attendant NLRI. With BGP SPF, only the BGP speakers corresponding to the link NLRI need withdraw the corresponding BGP-LS Link NLRI. This advantage will contribute to both faster convergence and better scaling.

With controller and route-reflector peering models, BGP SPF advertisement and distributed computation require a minimal number of sessions and copies of the NLRI since only the latest version of the NLRI from the originator is required. Given that verification of the adjacencies is done outside of BGP (see Section 2), each BGP speaker will only need as many sessions and copies of the NLRI as required for redundancy (e.g., one for SPF computation and another for backup). Functions such as Optimized Route Reflection (ORR) are supported without extension by virtue of the primary advantages. Additionally, a controller could inject topology that is learned outside the BGP routing domain.

Given that controllers are already consuming BGP-LS NLRI [RFC7752], reusing for the BGP-LS SPF leverages the existing controller implementations.

Another potential advantage of BGP SPF is that both IPv6 and IPv4 can be supported in the same address family using the same topology. Although not described in this version of the document, multi-topology extensions can be used to support separate IPv4, IPv6, unicast, and multicast topologies while sharing the same NLRI.

Finally, the BGP SPF topology can be used as an underlay for other BGP address families (using the existing model) and realize all the above advantages. A simplified peering model using IPv6 link-local addresses as next-hops can be deployed similar to [RFC5549].

1.2. 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. BGP Peering Models

Depending on the requirements, scaling, and capabilities of the BGP speakers, various peering models are supported. The only requirement is that all BGP speakers in the BGP SPF routing domain receive link-state NLRI on a timely basis, run an SPF calculation, and update their data plane appropriately. The content of the Link NLRI is described in Section 4.2.

2.1. BGP Single-Hop Peering on Network Node Connections

The simplest peering model is the one described in section 5.2.1 of [RFC7938]. In this model, EBGP single-hop sessions are established over direct point-to-point links interconnecting the network nodes. For the purposes of BGP SPF, Link NLRI is only advertised if a single-hop BGP session has been established and the Link-State/SPF address family capability has been exchanged [RFC4790] on the corresponding session. If the session goes down, the NLRI will be withdrawn.

2.2. BGP Peering Between Directly Connected Network Nodes

In this model, BGP speakers peer with all directly connected network nodes but the sessions may be multi-hop and the direct connection discovery and liveness detection for those connections are independent of the BGP protocol. How this is accomplished is outside the scope of this document. Consequently, there will be a single session even if there are multiple direct connections between BGP speakers. For the purposes of BGP SPF, Link NLRI is advertised as long as a BGP session has been established, the Link-State/SPF

address family capability has been exchanged [RFC4790] and the corresponding link is up and considered operational.

2.3. BGP Peering in Route-Reflector or Controller Topology

In this model, BGP speakers peer solely with one or more Route Reflectors [RFC4456] or controllers. As in the previous model, direct connection discovery and liveness detection for those connections are done outside the BGP protocol. For the purposes of BGP SPF, Link NLRI is advertised as long as the corresponding link is up and considered operational.

3. BGP-LS Shortest Path Routing (SPF) SAFI

In order to replace the Phase 1 and 2 decision functions of the existing Decision Process with an SPF-based Decision Process and streamline the Phase 3 decision functions in a backward compatible manner, this draft introduces a couple AFI/SAFIs for BGP LS SPF operation. The BGP-LS-SPF (AF 16388 / SAFI TBD1) and BGP-LS-SPF-VPN (AFI 16388 / SAFI TBD2) [RFC4790] are allocated by IANA as specified in the Section 6.

4. Extensions to BGP-LS

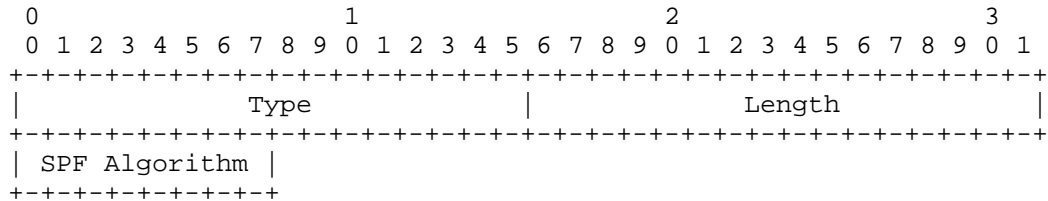
[RFC7752] describes a mechanism by which link-state and TE information can be collected from networks and shared with external components using BGP protocol. It contains two parts: definition of a new BGP NLRI that describes links, nodes, and prefixes comprising IGP link-state information and definition of a new BGP path attribute (BGP-LS attribute) that carries link, node, and prefix properties and attributes, such as the link and prefix metric or auxiliary Router-IDs of nodes, etc.

The BGP protocol will be used in the Protocol-ID field specified in table 1 of [I-D.ietf-idr-bgpls-segment-routing-epe]. The local and remote node descriptors for all NLRI will be the BGP Router-ID (TLV 516) and either the AS Number (TLV 512) [RFC7752] or the BGP Confederation Member (TLV 517) [I-D.ietf-idr-bgpls-segment-routing-epe]. However, if the BGP Router-ID is known to be unique within the BGP Routing domain, it can be used as the sole descriptor.

4.1. Node NLRI Usage and Modifications

The SPF capability is a new Node Attribute TLV that will be added to those defined in table 7 of [RFC7752]. The new attribute TLV will only be applicable when BGP is specified in the Node NLRI Protocol ID

field. The TBD TLV type will be defined by IANA. The new Node Attribute TLV will contain a single octet SPF algorithm field:



The SPF Algorithm may take the following values:

- 1 - Normal SPF
- 2 - Strict SPF

When computing the SPF for a given BGP routing domain, only BGP nodes advertising the SPF capability attribute will be included the Shortest Path Tree (SPT).

4.2. Link NLRI Usage

The criteria for advertisement of Link NLRI are discussed in Section 2.

Link NLRI is advertised with local and remote node descriptors as described above and unique link identifiers dependent on the addressing. For IPv4 links, the links local IPv4 (TLV 259) and remote IPv4 (TLV 260) addresses will be used. For IPv6 links, the local IPv6 (TLV 261) and remote IPv6 (TLV 262) addresses will be used. For unnumbered links, the link local/remote identifiers (TLV 258) will be used. For links supporting having both IPv4 and IPv6 addresses, both sets of descriptors may be included in the same Link NLRI. The link identifiers are described in table 5 of [RFC7752].

The link IGP metric attribute TLV (TLV 1095) as well as any others required for non-SPF purposes SHOULD be advertised. Algorithms such as setting the metric inversely to the link speed as done in the OSPF MIB [RFC4750] may be supported. However, this is beyond the scope of this document.

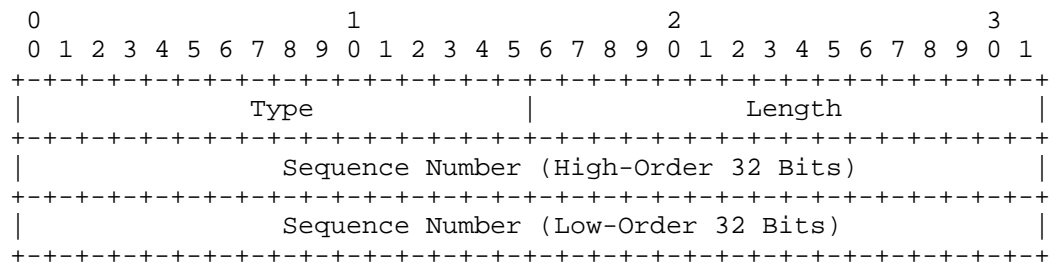
4.3. Prefix NLRI Usage

Prefix NLRI is advertised with a local descriptor as described above and the prefix and length used as the descriptors (TLV 265) as described in [RFC7752]. The prefix metric attribute TLV (TLV 1155) as well as any others required for non-SPF purposes SHOULD be

advertised. For loopback prefixes, the metric should be 0. For non-loopback, the setting of the metric is beyond the scope of this document.

4.4. BGP-LS Attribute Sequence-Number TLV

A new BGP-LS Attribute TLV to BGP-LS NLRI types is defined to assure the most recent version of a given NLRI is used in the SPF computation. The TBD TLV type will be defined by IANA. The new BGP-LS Attribute TLV will contain an 8 octet sequence number. The usage of the Sequence Number TLV is described in Section 5.1.



Sequence Number

The 64-bit strictly increasing sequence number is incremented for every version of BGP-LS NLRI originated. BGP speakers implementing this specification MUST use available mechanisms to preserve the sequence number's strictly increasing property for the deployed life of the BGP speaker (including cold restarts). One mechanism for accomplishing this would be to use the high-order 32 bits of the sequence number as a wrap/boot count that is incremented anytime the BGP Router router loses its sequence number state or the low-order 32 bits wrap.

When incrementing the sequence number for each self-originated NLRI, the sequence number should be treated as an unsigned 64-bit value. If the lower-order 32-bit value wraps, the higher-order 32-bit value should be incremented and saved in non-volatile storage. If by some chance the BGP Speaker is deployed long enough that there is a possibility that the 64-bit sequence number may wrap or a BGP Speaker completely loses its sequence number state (e.g, the BGP speaker hardware is replaced), the phase 1 decision function (see Section 5.1) rules should insure convergence, albeit, not immediately.

5. Decision Process with SPF Algorithm

The Decision Process described in [RFC4271] takes place in three distinct phases. The Phase 1 decision function of the Decision Process is responsible for calculating the degree of preference for each route received from a Speaker's peer. The Phase 2 decision function is invoked on completion of the Phase 1 decision function and is responsible for choosing the best route out of all those available for each distinct destination, and for installing each chosen route into the Loc-RIB. The combination of the Phase 1 and 2 decision functions is also known as a Path vector algorithm.

When BGP-LS-SPF NLRI is received, all that is required is to determine whether it is the best-path by examining the Node-ID and sequence number as described in Section 5.1. If the best-path NLRI had changed, it will be advertised to other BGP-LS-SPF peers. If the attributes have changed (other than the sequence number), a BGP SPF calculation will be scheduled. However, a changed best-path can be advertised to other peer immediately and propagation of changes can approach IGP convergence times.

The SPF based Decision process starts with selecting only those Node NLRI whose SPF capability TLV matches with the local BGP speaker's SPF capability TLV value. Since Link-State NLRI always contains the local descriptor [RFC7752], it will only be originated by a single BGP speaker in the BGP routing domain. These selected Node NLRI and their Link/Prefix NLRI are used to build a directed graph during the SPF computation. The best paths for BGP prefixes are installed as a result of the SPF process.

The Phase 3 decision function of the Decision Process [RFC4271] is also simplified since under normal SPF operation, a BGP speaker would advertise the NLRI selected for the SPF to all BGP peers with the BGP-LS/BGP-SPF AFI/SAFI. Application of policy would not be prevented but would normally not be necessary.

5.1. Phase-1 BGP NLRI Selection

The rules for NLRI selection are greatly simplified from [RFC4271].

1. If the NLRI is received from the BGP speaker originating the NLRI (as determined by the comparing BGP Router ID in the NLRI Node identifiers with the BGP speaker Router ID), then it is preferred over the same NLRI from non-originators.
2. If the Sequence-Number TLV is present in the BGP-LS Attribute, then the NLRI with the most recent, i.e., highest sequence number is selected. BGP-LS NLRI with a Sequence-Number TLV will be

considered more recent than NLRI without a BGP-LS or a BGP-LS Attribute that doesn't include the Sequence-Number TLV.

3. The final tie-breaker is the NLRI from the BGP Speaker with the numerically largest BGP Router ID.

The modified Decision Process with SPF algorithm uses the metric from Link and Prefix NLRI Attribute TLVs [RFC7752]. As a result, any attributes that would influence the Decision process defined in [RFC4271] like ORIGIN, MULTI_EXIT_DISC, and LOCAL_PREF attributes are ignored by the SPF algorithm. Furthermore, the NEXT_HOP attribute value is preserved and validated but otherwise ignored during the SPF or best-path.

5.2. Dual Stack Support

The SPF based decision process operates on Node, Link, and Prefix NLRIs that support both IPv4 and IPv6 addresses. Whether to run a single SPF instance or multiple SPF instances for separate AFs is a matter of a local implementation. Normally, IPv4 next-hops are calculated for IPv4 prefixes and IPv6 next-hops are calculated for IPv6 prefixes. However, an interesting use-case is deployment of [RFC5549] where IPv6 link-local next-hops are calculated for both IPv4 and IPv6 prefixes. As stated in Section 1, support for Multiple Topology Routing (MTR) is an area for future study.

5.3. NEXT_HOP Manipulation

A BGP speaker that supports SPF extensions MAY interact with peers that don't support SPF extensions. If the BGP Link-State address family is advertised to a peer not supporting the SPF extensions described herein, then the BGP speaker MUST conform to the NEXT_HOP rules mentioned in [RFC4271] when announcing the Link-State address family routes to those peers.

All BGP peers that support SPF extensions would locally compute the NEXT_HOP values as result of the SPF process. As a result, the NEXT_HOP attribute is always ignored on receipt. However BGP speakers should set the NEXT_HOP address according to the NEXT_HOP attribute rules mentioned in [RFC4271].

5.4. Error Handling

When a BGP speaker receives a BGP Update containing a malformed SPF Capability TLV in the Node NLRI BGP-LS Attribute [RFC7752], it MUST ignore the received TLV and the Node NLRI and not pass it to other BGP peers as specified in [RFC7606]. When discarding a Node NLRI

with malformed TLV, a BGP speaker SHOULD log an error for further analysis.

6. IANA Considerations

This document defines a couple AFI/SAFIs for BGP LS SPF operation and requests IANA to assign the BGP-LS-SPF AFI 16388 / SAFI TBD1 and the BGP-LS-SPF-VPN AFI 16388 / SAFI TBD2 as described in [RFC4750].

This document also defines two attribute TLV for BGP LS NLRI. We request IANA to assign TLVs for the SPF capability and the Sequence Number from the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" Registry. Additionally, IANA is requested to create a new registry for "BGP-LS SPF Capability Algorithms" for the value of the algorithm both in the BGP-LS Node Attribute TLV and the BGP SPF Capability. The initial assignments are:

Value(s)	Assignment Policy
0	Reserved (not to be assigned)
1	SPF
2	Strict SPF
3-254	Unassigned (IETF Review)
255	Reserved (not to be assigned)

BGP-LS SPF Capability Algorithms

7. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4724] and [RFC4271].

7.1. Acknowledgements

The authors would like to thank for the review and comments.

7.2. Contributorss

In addition to the authors listed on the front page, the following co-authors have contributed to the document.

Derek Yeung
Arrcus, Inc.
derek@arrcus.com

Abhay Roy
Cisco Systems
akr@cisco.com

Venu Venugopal
Cisco Systems
venuv@cisco.com

8. References

8.1. Normative References

- [I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Filsfils, C., Ray, S., Patel, K., Dong, J.,
and M. Chen, "Segment Routing Egress Peer Engineering BGP-
LS Extensions", draft-ietf-idr-bgpls-segment-routing-
epe-00 (work in progress), June 2015.
- [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>>.
- [RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K.
Patel, "Revised Error Handling for BGP UPDATE Messages",
RFC 7606, DOI 10.17487/RFC7606, August 2015,
<<http://www.rfc-editor.org/info/rfc7606>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and
S. Ray, "North-Bound Distribution of Link-State and
Traffic Engineering (TE) Information Using BGP", RFC 7752,
DOI 10.17487/RFC7752, March 2016,
<<http://www.rfc-editor.org/info/rfc7752>>.
- [RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of
BGP for Routing in Large-Scale Data Centers", RFC 7938,
DOI 10.17487/RFC7938, August 2016,
<<http://www.rfc-editor.org/info/rfc7938>>.

8.2. Information References

- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <<http://www.rfc-editor.org/info/rfc2328>>.
- [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>>.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <<http://www.rfc-editor.org/info/rfc4724>>.
- [RFC4750] Joyal, D., Ed., Galecki, P., Ed., Giacalone, S., Ed., Coltun, R., and F. Baker, "OSPF Version 2 Management Information Base", RFC 4750, DOI 10.17487/RFC4750, December 2006, <<http://www.rfc-editor.org/info/rfc4750>>.
- [RFC4790] Newman, C., Duerst, M., and A. Gulbrandsen, "Internet Application Protocol Collation Registry", RFC 4790, DOI 10.17487/RFC4790, March 2007, <<http://www.rfc-editor.org/info/rfc4790>>.
- [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>>.
- [RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, DOI 10.17487/RFC5286, September 2008, <<http://www.rfc-editor.org/info/rfc5286>>.
- [RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <<http://www.rfc-editor.org/info/rfc5549>>.

Authors' Addresses

Keyur Patel
Arrcus, Inc.

Email: keyur@arrcus.com

Acee Lindem
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: acee@cisco.com

Shawn Zandi
Linkedin
222 2nd Street
San Francisco, CA 94105
USA

Email: szandi@linkedin.com

Gunter Van de Velde
Nokia
Antwerp
Belgium

Email: gunter.van_de_velde@nokia.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 4, 2017

P. Lapukhov
Facebook
October 31, 2016

Equal-Cost Multipath Considerations for BGP
draft-lapukhov-bgp-ecmp-considerations-00

Abstract

BGP routing protocol defined in ([RFC4271]) employs tie-breaking logic to elect single best path among multiple possible. At the same time, it has been common in virtually all BGP implementations to allow for "equal-cost multipath" (ECMP) election and programming of multiple next-hops in routing tables. This documents summarizes some common considerations for the ECMP logic, with the intent of providing common reference on otherwise unstandardized feature.

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 May 4, 2017.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. AS-PATH attribute comparison	2
3. Multipath among eBGP-learned paths	3
4. Multipath among iBGP learned paths	3
5. Multipath among eBGP and iBGP paths	4
6. Multipath with AIGP	5
7. Best path advertisement	5
8. Multipath and non-deterministic tie-breaking	5
9. Weighted equal-cost multipath	5
10. Informative References	5
Author's Address	6

1. Introduction

Section 9.1.2.2 of [RFC4271] defines step-by-step procedure for selecting single "best-path" among multiple alternative available for the same NLRI (Network Layer Reachability Information) element. In order to improve efficiency in symmetric network topologies is has become common practice to allow for selecting multiple "equivalent" paths for the same prefix. Most commonly used approach is to abort the tie-breaking process after comparing the IGP cost for the NEXT_HOP attribute and selecting either all eBGP or all iBGP paths that remained equivalent under the tie-breaking rules (see [BGPMP] for a vendor document explaining the logic). Basically, the steps that compare the BGP identifier and BGP peer IP addresses (steps (f) and (g)) are ignored for the purpose of multipath routing. BGP implementations commonly have a configuration knob that specifies the maximum number of equivalent paths that may be programmed to the routing table. There is also common a knob to enable multipath separately for iBGP-learned or eBGP-learned paths.

2. AS-PATH attribute comparison

A mandatory requirement is for all paths that are candidates for ECMP selection to have the same AS_PATH length, computed using the standard logic defined in [RFC4271] and [RFC5065], i.e. ignoring the AS_SET, AS_CONFED_SEQUENCE, and AS_CONFED_SET segment lengths. The content of the latter attributes is used purely for loop detection. Assuming that AS_PATH lengths computed in this fashion are the same, many implementations require that content of AS_SEQUENCE segment MUST be the same among all equivalent paths. Two common configuration knobs are usually provided: one allowing only the length of AS_PATH to be the same, and another requiring that the first AS numbers in

first AS_SEQUENCE segment found in AS_PATH (often referred to as "peer AS" number) be the same as the one found in best path (determined by running the full tie-breaking algorithm). This document refer to those two as "multipath as-path relaxed" and "multipath same peer-as" knobs.

3. Multipath among eBGP-learned paths

Step (d) in Section 9.1.2.2 of [RFC4271] instructs to remove all iBGP paths from considerations if an eBGP path is present in the candidate set. This leaves the BGP process with just eBGP paths. At this point, the mandatory BGP NEXT_HOP attribute value most commonly belongs to the IP subnet that the BGP speaker shares with advertising neighbor. In this case, it is common for implementation to treat all NEXT_HOP values as having the same "internal cost" to reach them per the guidance of step (e) of Section 9.1.2.2. In some cases, either static routing or an IGP routing protocol could be running between the BGP speakers peering over eBGP session. An implementation may use the metric discovered from the above sources to perform tie-breaking even for eBGP paths.

Notice that in case when MED attribute is present in some paths, the set of allowed multipath routes will most likely be reduced to the ones coming from the same peer AS, per step (c) of Section 9.1.2.2. This is unless the implementation provided a configuration knob to always compare MED attributes across all paths, as recommended in [RFC4451]. In the latter case, the presence of MED attribute does not automatically narrow the candidate path set only to the same peer AS.

4. Multipath among iBGP learned paths

When all paths for a prefix are learned via iBGP, the tie-breaking commonly occurs based on IGP metric of the NEXT_HOP attribute, since in most cases iBGP is used along with an underlying IGP. It is possible, in some implementations, to ignore the IGP cost as well, if all of the paths are reachable via some kind of tunneling mechanism, such as MPLS ([RFC3031]). This is enabled via a knob referred to as "skip igp check" in this document. Notice that there is no standard way for a BGP speaker to detect presence of such tunneling techniques other than relying on configuration settings.

When iBGP is deployed with BGP route-reflectors per [RFC4456] the path attribute list may include the CLUSTER_LIST attribute. Most implementations commonly ignore it for the purpose of ECMP route selection, assuming that IGP cost along should be sufficient for loop prevention. This assumption may not hold when IGP is not deployed, and instead iBGP session are configured to reset the NEXT_HOP

attribute to self on every node (this also assumes the use of directly connected link addresses for session formation). In this case, ignoring CLUSTER_LIST length might lead to routing loops. It is therefore recommended for implementations to have a knob that enables accounting for CLUSTER_LIST length when performing multipath route selection. In this case, CLUSTER_LIST attribute length should be effectively used to replace the IGP metric.

Similar to the route-reflector scenario, the use of BGP confederations assumes presence of an IGP for proper loop prevention in multipath scenarios, and use the IGP metric as the final tie-breaker for multipath routing. In addition to this, and similar to eBGP case, implementation often require that equivalent paths belong to the same peer member AS as the best-path. It is useful to have two configuration knobs, one enabling "multipath same confederation member peer-as" and another enabling less restrictive "confed as-path multipath relaxed", which allows selecting multipath routes going via any confederation member peer AS. As mentioned above, the AS_CONFED_SEQUENCE value length is usually ignored for the purpose of AS_PATH length comparison, relying on IGP cost instead for loop prevention.

In case if IGP is not present with BGP confederation deployment, and similar to route-reflection case, it may be needed to consider AS_CONFED_SEQUENCE length when selecting the equivalent routes, effectively using it as a substitution for IGP metric. A separate configuration knob is needed to allow this behavior.

Per [RFC5065] the path learned over BGP intra-confederation peering sessions are treated as iBGP. There is no specification or operational document that defines how a mixed iBGP route-reflector and confederation based model would work together. Therefore, this document does not make recommendations or considers this case.

5. Multipath among eBGP and iBGP paths

The best-path selection algorithm explicitly prefers eBGP paths over iBGP (or learned from BGP confederation member AS, which is per [RFC5065] is treated the same as iBGP from perspective of best-path selection). In some case, allowing multipath routing between eBGP and iBGP learned paths might be beneficial. This is only possible if some sort of tunneling technique is used to reach both the eBGP and iBGP path. If this feature is enabled, the equivalent routes are selection by stopping the tie-breaking process prior at the MED comparison step (c) in Section 9.1.2.2 of [RFC4271].

6. Multipath with AIGP

AIGP attribute defined in [RFC7311] must be used for best-path selection prior to running any logic of Section 9.1.2.2. Only the paths with minimal value of AIGP metric are eligible for further consideration of tie-breaking rules. The rest of multipath selection logic remains the same.

7. Best path advertisement

Event though multiple equivalent paths may be selected for programming into the routing table, the BGP speaker always announces single best-path to its peers, unless BGP "Add-Path" feature has been enabled as described in [I-D.ietf-idr-add-paths]. The unique best-path is elected among the multi-path set using the standard tie-breaking rules.

8. Multipath and non-deterministic tie-breaking

Some implementations may implement non-standard tie-breaking using the oldest path rule. This is generally not recommended, and may interact with multi-path route selection on downstream BGP speakers. That is, after a route flap that affects the best-path upstream, the original best path would not be recovered, and the older path still be advertised, possibly affecting the tie-breaking rules on downstream device, for example if the AS_PATH contents are different from previous.

9. Weighted equal-cost multipath

The proposal in [I-D.ietf-idr-link-bandwidth] defines conditions where iBGP multipath feature might inform the routing table of the "weights" associated with the multiple paths. The document defines the applicability only in iBGP case, though there are implementations that apply it to eBGP multipath as well. The proposal does not change the equal-cost multipath selection logic, only associates additional load-sharing attributes with equivalent paths.

10. Informative References

- [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>>.

- [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>>.
- [RFC4451] McPherson, D. and V. Gill, "BGP MULTI_EXIT_DISC (MED) Considerations", RFC 4451, DOI 10.17487/RFC4451, March 2006, <<http://www.rfc-editor.org/info/rfc4451>>.
- [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>>.
- [RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, DOI 10.17487/RFC5065, August 2007, <<http://www.rfc-editor.org/info/rfc5065>>.
- [RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <<http://www.rfc-editor.org/info/rfc7311>>.
- [I-D.ietf-idr-add-paths]
Walton, D., Retana, A., Chen, E., and J. Scudder,
"Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-15 (work in progress), May 2016.
- [I-D.ietf-idr-link-bandwidth]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-06 (work in progress), January 2013.
- [BGPMP] "BGP Best Path Selection Algorithm",
<<http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.html>>.

Author's Address

Petr Lapukhov
Facebook
1 Hacker Way
Menlo Park, CA 94025
US

Email: petr@fb.com

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2019

Z. Li
Huawei
S. Aldrin
Google, Inc
J. Tantsura
Apstra
G. Mirsky
ZTE Corp.
S. Zhuang
Huawei
K. Talaulikar
Cisco Systems
February 22, 2019

BGP Link-State Extensions for Seamless BFD
draft-li-idr-bgp-ls-sbfd-extensions-03

Abstract

Seamless Bidirectional Forwarding Detection (S-BFD) defines a simplified mechanism to use Bidirectional Forwarding Detection (BFD) with large portions of negotiation aspects eliminated, thus providing benefits such as quick provisioning as well as improved control and flexibility to network nodes initiating the path monitoring. The link-state routing protocols (IS-IS and OSPF) have been extended to advertise the Seamless BFD (S-BFD) Discriminators.

This draft defines extensions to the BGP Link-state address-family to carry the S-BFD Discriminators information via BGP.

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

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 <https://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 26, 2019.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Problem and Requirement	3
4. BGP-LS Extensions for S-BFD Discriminator	4
5. IANA Considerations	6
6. Manageability Considerations	6
6.1. Operational Considerations	6
6.2. Management Considerations	6
7. Security Considerations	6
8. Acknowledgements	7
9. References	7
9.1. Normative References	7
9.2. Informative References	7
Authors' Addresses	8

1. Introduction

Seamless Bidirectional Forwarding Detection (S-BFD) [RFC7880] defines a simplified mechanism to use Bidirectional Forwarding Detection (BFD) [RFC5880] with large portions of negotiation aspects eliminated, thus providing benefits such as quick provisioning as well as improved control and flexibility to network nodes initiating the path monitoring.

For monitoring of a service path end-to-end via S-BFD, the headend/initiator node needs to know the S-BFD Discriminator of the destination/tail-end node of that service. The link-state routing protocols (IS-IS, OSPF and OSPFv3) have been extended to advertise the S-BFD Discriminators. With this a initiator node can learn the S-BFD discriminator for all nodes within its IGP area/level or optionally within the domain. With networks being divided into multiple IGP domains for scaling and operational considerations, the service endpoints that require end to end S-BFD monitoring often span across IGP domains.

BGP Link-State (BGP-LS) [RFC7752] enables the collection and distribution of IGP link-state topology information via BGP sessions across IGP areas/levels and domains. The S-BFD discriminator(s) of a node can thus be distributed along with the topology information via BGP-LS across IGP domains and even across multiple Autonomous Systems (AS) within an administrative domain.

This draft defines extensions to BGP-LS for carrying the S-BFD Discriminators information.

2. Terminology

This memo makes use of the terms defined in [RFC7880].

3. Problem and Requirement

Seamless MPLS [I-D.ietf-mpls-seamless-mpls] extends the core domain and integrates aggregation and access domains into a single MPLS domain. In a large network, the core and aggregation networks can be organized as different ASes. Although the core and aggregation networks are segmented into different ASes, an E2E LSP can be created using hierarchical BGP signaled LSPs based on iBGP labeled unicast within each AS, and eBGP labeled unicast to extend the LSP across AS boundaries. This provides a seamless MPLS transport connectivity for any two service end-points across the entire domain. In order to detect failures for such end to end services and trigger faster protection and/or re-routing, S-BFD MAY be used for the Service Layer (e.g. for MPLS VPNs, PW, etc.) or the Transport Layer monitoring. This brings up the need for setting up S-BFD session spanning across AS domains.

In a similar Segment Routing (SR) [RFC8402] multi-domain network, an end to end SR Policy [I-D.ietf-spring-segment-routing-policy] path may be provisioned between service end-points across domains either via local provisioning or by a controller or signalled from a Path Computation Engine (PCE). Monitoring using S-BFD can similarly be setup for such a SR Policy.

Extending the automatic discovery of S-BFD discriminators of nodes from within the IGP domain to across the administrative domain using BGP-LS enables setting up of S-BFD sessions on demand across IGP domains. The S-BFD discriminators for service end point nodes MAY be learnt by the PCE or a controller via the BGP-LS feed that it gets from across IGP domains and it can signal or provision the remote S-BFD discriminator on the initiator node on demand when S-BFD monitoring is required. The mechanisms for the signaling of the S-BFD discriminator from the PCE/controller to the initiator node and setup of the S-BFD session is outside the scope of this document.

Additionally, the service end-points themselves MAY also learn the S-BFD discriminator of the remote nodes themselves by receiving the BGP-LS feed via a route reflector (RR) or a centralized BGP Speaker that is consolidating the topology information across the domains. The initiator node can then itself setup the S-BFD session to the remote node without a controller/PCE assistance.

While this document takes examples of MPLS and SR paths, the S-BFD discriminator advertisement mechanism is applicable for any S-BFD use-case in general.

4. BGP-LS Extensions for S-BFD Discriminator

The BGP-LS [RFC7752] specifies the Node NLRI for advertisement of nodes and their attributes using the BGP-LS Attribute. The S-BFD discriminators of a node are considered as its node level attribute and advertised as such.

This document defines a new BGP-LS Attribute TLV called the S-BFD Discriminators TLV and its format is as follows:

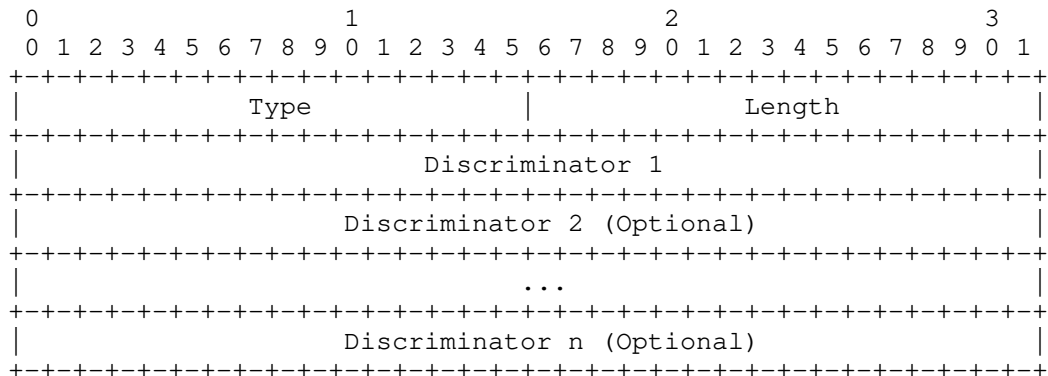


Figure 1: S-BFD Discriminators TLV

where:

- o Type: TBD (see IANA Considerations Section 5)
- o Length: variable. Minimum of 8 octets and increments of 4 octets there on for each additional discriminator
- o Discriminators : multiples of 4 octets, each carrying a S-BFD local discriminator value of the node. At least one discriminator MUST be included in the TLV.

The S-BFD Discriminators TLV can only be added to the BGP-LS Attribute associated with the Node NLRI that originates the corresponding underlying IGP TLV/sub-TLV as described below. This information is derived from the protocol specific advertisements as below..

- o IS-IS, as defined by the S-BFD Discriminators sub-TLV in [RFC7883].
- o OSPFv2/OSPFv3, as defined by the S-BFD Discriminators TLV in [RFC7884].

When the node is not running any of the IGPs but running a protocol like BGP, then the locally provisioned S-BFD discriminators of the node MAY be originated as part of the BGP-LS attribute within the Node NLRI corresponding to the local node.

5. IANA Considerations

This document requests assigning code-points from the registry "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" based on table below. The column "IS-IS TLV/Sub-TLV" defined in the registry does not require any value and should be left empty.

Code Point	Description	Length
TBD	S-BFD Discriminators TLV	variable

6. Manageability Considerations

This section is structured as recommended in [RFC5706].

The new protocol extensions introduced in this document augment the existing IGP topology information that was distributed via [RFC7752]. Procedures and protocol extensions defined in this document do not affect the BGP protocol operations and management other than as discussed in the Manageability Considerations section of [RFC7752]. Specifically, the malformed NLRIs attribute tests in the Fault Management section of [RFC7752] now encompass the new TLVs for the BGP-LS NLRI in this document.

6.1. Operational Considerations

No additional operation considerations are defined in this document.

6.2. Management Considerations

No additional management considerations are defined in this document.

7. Security Considerations

The new protocol extensions introduced in this document augment the existing IGP topology information that was distributed via [RFC7752]. Procedures and protocol extensions defined in this document do not affect the BGP security model other than as discussed in the Security Considerations section of [RFC7752]. More specifically the aspects related to limiting the nodes and consumers with which the topology information is shared via BGP-LS to trusted entities within an administrative domain.

Advertising the S-BFD Discriminators via BGP-LS makes it possible for attackers to initiate S-BFD sessions using the advertised

information. The vulnerabilities this poses and how to mitigate them are discussed in [RFC7752].

8. Acknowledgements

The authors would like to thank Nan Wu for his contributions to this work.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<https://www.rfc-editor.org/info/rfc7752>>.
- [RFC7880] Pignataro, C., Ward, D., Akiya, N., Bhatia, M., and S. Pallagatti, "Seamless Bidirectional Forwarding Detection (S-BFD)", RFC 7880, DOI 10.17487/RFC7880, July 2016, <<https://www.rfc-editor.org/info/rfc7880>>.
- [RFC7883] Ginsberg, L., Akiya, N., and M. Chen, "Advertising Seamless Bidirectional Forwarding Detection (S-BFD) Discriminators in IS-IS", RFC 7883, DOI 10.17487/RFC7883, July 2016, <<https://www.rfc-editor.org/info/rfc7883>>.
- [RFC7884] Pignataro, C., Bhatia, M., Aldrin, S., and T. Ranganath, "OSPF Extensions to Advertise Seamless Bidirectional Forwarding Detection (S-BFD) Target Discriminators", RFC 7884, DOI 10.17487/RFC7884, July 2016, <<https://www.rfc-editor.org/info/rfc7884>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

9.2. Informative References

- [I-D.ietf-mpls-seamless-mpls]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.
- [I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Sivabalan, S., daniel.voyer@bell.ca, d., bogdanov@google.com, b., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-02 (work in progress), October 2018.
- [RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, DOI 10.17487/RFC5706, November 2009, <<https://www.rfc-editor.org/info/rfc5706>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.

Authors' Addresses

Zhenbin Li
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Sam Aldrin
Google, Inc

Email: aldrin.ietf@gmail.com

Jeff Tantsura
Apstra

Email: jefftant.ietf@gmail.com

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Shunwan Zhuang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Ketan Talaulikar
Cisco Systems
India

Email: ketant@cisco.com

IDR
Internet-Draft
Updates: 4271, 4360, 7153 (if approved)
Intended status: Standards Track
Expires: September 4, 2018

Z. Li
China Mobile
J. Dong
Huawei Technologies
March 3, 2018

Carry congestion status in BGP community
draft-li-idr-congestion-status-extended-community-07

Abstract

To aid BGP receiver to steer the AS-outgoing traffic among the exit links, this document introduces a new BGP community, congestion status community, to carry the link bandwidth and utilization information, especially for the exit links of one AS. If accepted, this document will update RFC4271, RFC4360 and RFC7153.

The introduced congestion status community is not used to impact the decision process of BGP specified in section 9.1 of RFC4271, but can be used by route policy to impact the data forwarding behavior.

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 <https://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 September 4, 2018.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

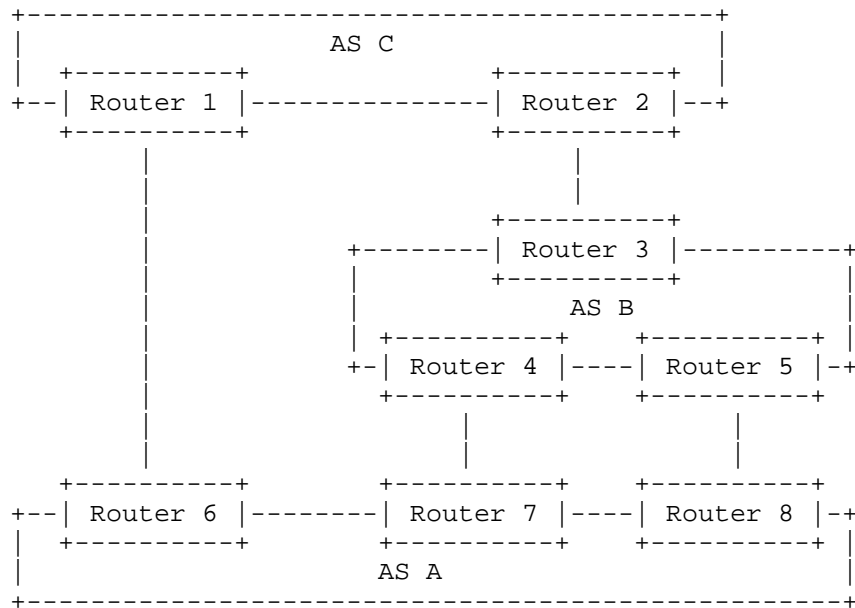
Table of Contents

1. Introduction	2
2. Requirements Language	4
3. Previous Work	4
4. Solution Alternative 1: Extended Community	4
5. Solution Alternative 2: Large Community	6
6. Solution Alternative 3: Community Container	6
7. Deployment Considerations	8
8. Security Considerations	9
9. IANA Considerations	9
10. Acknowledgments	9
11. References	9
11.1. Normative References	10
11.2. Informative References	10
Appendix A. Bandwidth Values	11
Authors' Addresses	12

1. Introduction

Knowing the congestion status (bandwidth and utilization) of the AS exit links is useful for traffic steering, especially for steering the AS outgoing traffic among the exit links. Section 7 of [I-D.gredler-idr-bgplu-epe] explicitly specifies this kind of requirement, which is also needed in our field network.

The following figure is used to illustrate the benefits of knowing the congestion status of the AS exit links. AS A has multiple exit links connected to AS B. Both AS A and B has exit link to AS C, and AS B provides transit service for AS A. Due to cost or some other reasons, AS A prefers using AS B to transmit its' traffic to AS C, not the directly connected link between AS A and C. If the exit routers, Router 7 and 8, in AS A tell their iBGP peers the congestion status of the exit links, the peers in turn can steer some outgoing traffic toward the less loaded exit link. If AS A knows the link between AS B and AS C is congested, it can steer some traffic towards AS C from AS B to the directly connected link by applying some route policies.



This document introduces new BGP extensions to deliver the congestion status of the exit link to other BGP speakers. The BGP receiver can then use this community to deploy route policy, thus steer AS outgoing traffic according to the congestion status of the exit links. This mechanism can be used by both iBGP and eBGP.

In this version, we provide three solution alternatives according to the discussion in the face to face meetings and mail list. After adoption, one solution will be selected as the final solution based on the working group consensus.

In a network deployed SDN (Software Defined Network) controller, congestion status extended community can be used by the controller to steer the AS outgoing traffic among all the exit links from the perspective of the whole network.

For the network with Route Reflectors (RRs) [RFC4456], RRs by default only advertise the best route for a specific prefix to their clients. Thus RR clients has no opportunity to compare the congestion status among all the exit links. In this situation, to allow RR clients learning all the routes for a specific prefix from all the exit links, RRs are RECOMMENDED to enable add-path functionality [RFC7911].

To emphasize, the introduced new BGP extensions have no impact on the decision process of BGP specified in section 9.1 of [RFC4271], but can be used by route policy to impact the data forwarding behavior.

2. 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 [RFC2119].

3. Previous Work

In [constrained-multiple-path], authors from France Telecom also specified the requirement to know the congestion status of a link.

To aid a router to perform unequal cost load balancing, experts from Cisco introduced Link Bandwidth Extended Community in [link-bandwidth-community] to carry the cost to reach the external BGP neighbor. The cost can be either configured per neighbor or derived from the bandwidth of the link that connects the router to a directly connected external neighbor. This document was accepted by the IDR working group, but expired in 2013.

Link Bandwidth Extended Community only carries the link bandwidth of the exit link. The method provided in our document can carry the link bandwidth together with the link utilization information. What the BGP receiver needs to impact its traffic steering policy is the up-to-date unused link bandwidth, which can be derived from the link bandwidth and link utilization. Since Link Bandwidth Extended Community is expired, the BGP speaker who receives update message with both Link Bandwidth Extended Community and Congestion Status Community SHOULD ignore the Link Bandwidth Extended Community and use the Congestion Status Community.

4. Solution Alternative 1: Extended Community

As described in [RFC4360], the extended community attribute is an 8-octet value with the first one or two octets to indicate the type of this attribute. Since congestion status community needs to be delivered from one AS to other ASes, and used by the BGP speakers both in other ASes and within the same AS as the sender, it MUST be a transitive extended community, i.e. the T bit in the first octet MUST be zero.

We only define the congestion status community for four-octet AS number [RFC6793], since all the BGP speakers can handle four-octet AS number now and the two-octet AS numbers can be mapped to four-octet

AS numbers by setting the two high-order octets of the four-octet field to zero, as per [RFC6793].

Congestion status community is a sub-type allocated from Transitive Four-Octet AS-Specific Extended Community Sub-Types defined in section 5.2.4 of [RFC7153]. Its format is as Figure 1.

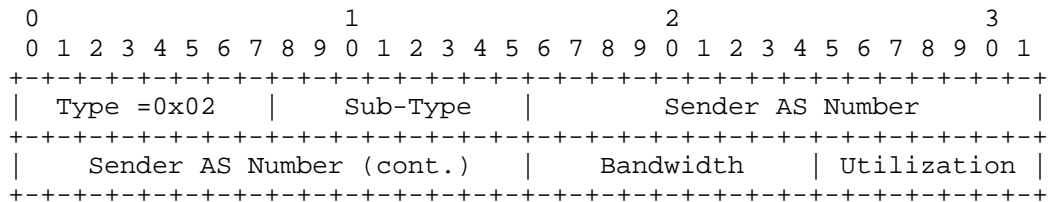


Figure 1: Congestion status extended community

Type: 1 octet. This field MUST be 0x02 to indicate this is a Transitive Four-Octet AS-Specific Extended Community.

Sub-Type: 1 octet. It is used to indicate this is a Congestion Status Extended Community. Its value is to be assigned by IANA.

Sender AS Number: 4 octets. Its value is the AS number of the BGP speaker who generates this congestion status extended community. If the generator has 2-octet AS number, it MUST encode its AS number in the last (low order) two bytes and set the first (high order) two bytes to zero, as per [RFC6793].

Bandwidth: 1 octet. Its value is the bandwidth of the exit link in unit of 10 gbps (gigabits per second). The link with bandwidth less than 10 gbps is not suitable to use this feature. To reflect the practice that sometimes the traffic is rate limited to a capacity smaller than the physical link, the value of the bandwidth can be the configured capacity of the link. The available configured capacity can be calculated from this field together with Utilization field. Zero means the bandwidth is unknown or is not advertised to other peers.

Utilization: 1 octet. Its value is the utilization of the exit link in unit of percent. A value bigger than 100 means the incoming traffic is higher than the link capacity. We can use the "Utilization" field together with the "Bandwidth" field to calculate the traffic load that we can further steer to this exit link.

5. Solution Alternative 2: Large Community

As described in [RFC8092], the BGP large community attribute is an optional transitive path attribute of variable length, consisting of 12-octet values. The BGP large community attribute is mainly used to extend the size of BGP Community [RFC1997] and Extended Community [RFC4360], thus to accommodate at least two four-octet ASNs [RFC6793]. As shown in the following figure, the format of the 12-octet BGP Large Community value is not suitable to be used to define new type for congestion status community.

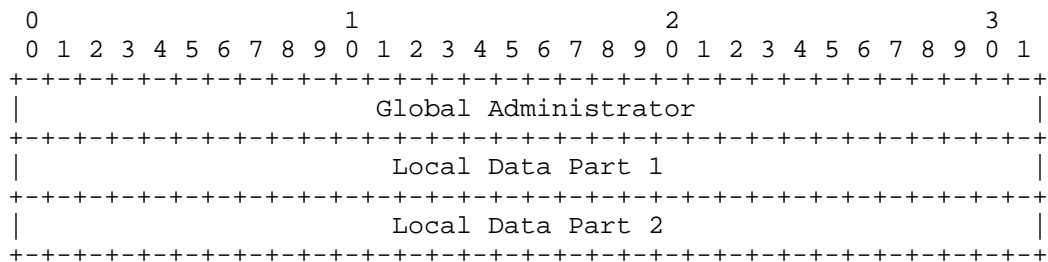


Figure 2

Global Administrator: A four-octet namespace identifier.

Local Data Part 1: A four-octet operator-defined value.

Local Data Part 2: A four-octet operator-defined value.

6. Solution Alternative 3: Community Container

As described in [I-D.ietf-idr-wide-bgp-communities], the BGP Community Container has flexible encoding format, which we can use to define the congestion status community.

A new type of the BGP Community Container is defined for the congestion status community, which has the same common header as the BGP Community Container with the following encoding format.

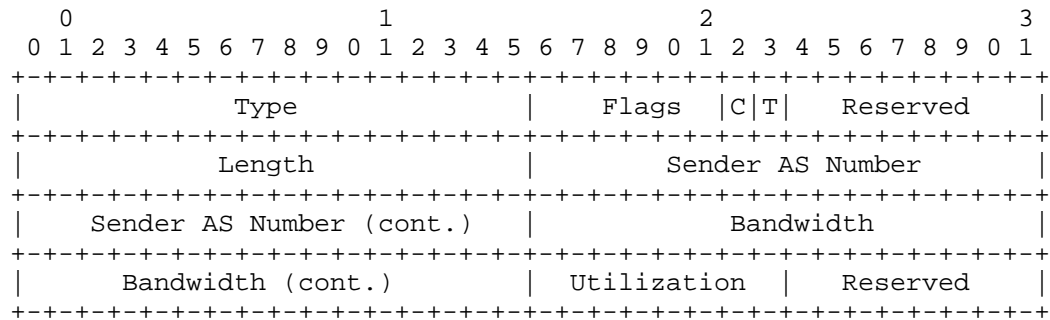


Figure 3

Type: 2 octets. Its value is to be assigned by IANA from the registry "BGP Community Container Types" to indicate this is the Congestion Status Community.

Flags: 1 octet. C and T bits MUST be set to indicate the Congestion Status Community is transitive across confederation and AS boundaries. The other bits in Flags field MUST be set to zero when originated and SHOULD be ignored upon receipt.

Reserved: Reserved fields are reserved for future definition, which MUST be set to zero when originated and SHOULD be ignored upon receipt.

Length: 2 octets. This field represents the total length of a given container's contents in octets.

Sender AS Number: 4 octets. Its value is the AS number of the BGP speaker who generates this congestion status community. If the generator has 2-octet AS number, it MUST encode its AS number in the last (low order) two bytes and set the first (high order) two bytes to zero, as per [RFC6793].

Bandwidth: 4 octets. Its value is the bandwidth of the exit link in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Zero means the bandwidth is unknown or is not advertised to other peers. Appendix A lists some typical bandwidth values, most of which are extracted from Section 3.1.2 of [RFC3471].

To reflect the practice that sometimes the traffic is rate limited to a capacity smaller than the physical link, the value of the bandwidth can be the configured capacity of the link. The available configured capacity can be calculated from this field together with Utilization field.

Utilization: 1 octet. Its value is the utilization of the exit link in unit of percent. A value bigger than 100 means the incoming traffic is higher than the link capacity. We can use the "Utilization" field together with the "Bandwidth" field to calculate the traffic load that we can further steer to this exit link.

7. Deployment Considerations

o To avoid route oscillation

The exit router SHOULD set a threshold. When the utilization change reaches the threshold, the exit router SHOULD generate a BGP update message with congestion status community.

Implementations SHOULD further reduce the BGP update messages triggered by link utilization change using the method similar to BGP Route Flap Damping [RFC2439]. When link utilization change by small amounts that fall under thresholds that would cause the announcement of BGP update message, implementations SHOULD suppress the announcement and set the penalty value accordingly.

To reduce the update churn introduced, when one BGP router needs to re-advertise a BGP path due to attribute changes, it SHOULD update its Congestion Status Community at the same time. Supposing there are N ASes on the way from the far end egress BGP speaker to the final ingress BGP speaker, this allows reducing the update churn as the final ingress BGP speaker will receive a single UPDATE refreshing the N communities, rather than N UPDATES, each refreshing one community.

o To avoid traffic oscillation

Traffic oscillation means more traffic than expected is attracted to the low utilized link, and some traffic has to be steered back to other links.

Route policy is RECOMMENDED to be set at the exit router. Congestion status community is only conveyed for some specific routes or only for some specific BGP peers.

Congestion status community can also be used in a SDN network. The SDN controller uses the exit link utilization information to steer the Internet access traffic among all the exit links from the perspective of the whole network.

o Other Consens

To avoid forwarding loops incremental deployment issues, complications in error handling, the reception of such community over IBGP session SHOULD NOT influence routing decision unless tunneling is used to reach the BGP Next-Hop.

8. Security Considerations

This document defines a new BGP community to carry the congestion status of the exit link. It is up to the BGP receiver to trust the congestion status communities or not. Following deployment models can be considered.

The BGP receiver may choose to only trust the congestion status communities generated by some specific ASes or containing bandwidth greater than a specific value.

You can filter the congestion status communities at the border of your trust/administrative domain. Hence all the ones you receive are trusted.

You can record the communities received over time, monitor the congestion e.g. via probing, detect inconsistency and choose to not trust anymore the ASes which advertise fake news.

9. IANA Considerations

For solution alternative 1, one sub-type is solicited to be assigned from Transitive Four-Octet AS-Specific Extended Community Sub-Types registry to indicate the Congestion Status Community defined in this document.

For solution alternative 3, one community value is solicited to be assigned from the registry "Registered Type 1 BGP Wide Community Community Types" to indicate the Congestion Status Community defined in this document.

10. Acknowledgments

We appreciate the constructive suggestions received from Bruno Decraene. Many thanks to Rudiger Volk, Susan Hares, John Scudder, Randy Bush for their review and comments to improve this document.

11. References

11.1. Normative References

- [I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S.,
and P. Jakma, "BGP Community Container Attribute", draft-
ietf-idr-wide-bgp-communities-04 (work in progress), March
2017.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119,
DOI 10.17487/RFC2119, March 1997,
<<https://www.rfc-editor.org/info/rfc2119>>.
- [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,
<<https://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, <<https://www.rfc-editor.org/info/rfc4360>>.
- [RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP
Extended Communities", RFC 7153, DOI 10.17487/RFC7153,
March 2014, <<https://www.rfc-editor.org/info/rfc7153>>.
- [RFC8092] Heitz, J., Ed., Snijders, J., Ed., Patel, K., Bagdonas,
I., and N. Hilliard, "BGP Large Communities Attribute",
RFC 8092, DOI 10.17487/RFC8092, February 2017,
<<https://www.rfc-editor.org/info/rfc8092>>.

11.2. Informative References

- [constrained-multiple-path]
Boucadair, M. and C. Jacquenet, "Constrained Multiple BGP
Paths", October 2010, <[https://www.ietf.org/archive/id/
draft-boucadair-idr-constrained-multiple-path-00.txt](https://www.ietf.org/archive/id/draft-boucadair-idr-constrained-multiple-path-00.txt)>.
- [I-D.gredler-idr-bgplu-epe]
Gredler, H., Vairavakkalai, K., R, C., Rajagopalan, B.,
Aries, E., and L. Fang, "Egress Peer Engineering using
BGP-LU", draft-gredler-idr-bgplu-epe-11 (work in
progress), October 2017.

[link-bandwidth-community]

Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", January 2013, <<https://www.ietf.org/archive/id/draft-ietf-idr-link-bandwidth-06.txt>>.

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <<https://www.rfc-editor.org/info/rfc1997>>.

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, DOI 10.17487/RFC2439, November 1998, <<https://www.rfc-editor.org/info/rfc2439>>.

[RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, DOI 10.17487/RFC3471, January 2003, <<https://www.rfc-editor.org/info/rfc3471>>.

[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, <<https://www.rfc-editor.org/info/rfc4456>>.

[RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, DOI 10.17487/RFC6793, December 2012, <<https://www.rfc-editor.org/info/rfc6793>>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <<https://www.rfc-editor.org/info/rfc7911>>.

Appendix A. Bandwidth Values

Some typical bandwidth values encoded in 32-bit IEEE floating point format are enumerated below.

Link Type	Bit-rate (Mbps)	Bandwidth Value (Bytes/Sec) (32-bit IEEE Floating point)
-----	-----	-----
E1	2.048	0x487A0000
Ethernet	10.00	0x49989680
Fast Ethernet	100.00	0x4B3EBC20
OC-3/STM-1	155.52	0x4B9450C0
OC-12/STM-4	622.08	0x4C9450C0
GigE	1000.00	0x4CEE6B28
OC-48/STM-16	2488.32	0x4D9450C0
OC-192/STM-64	9953.28	0x4E9450C0
10GigE	10000.00	0x4E9502F9
OC-768/STM-256	39813.12	0x4F9450C0
100GigE	100000.00	0x503A43B7

Authors' Addresses

Zhenqiang Li
China Mobile
No.32 Xuanwumenxi Ave., Xicheng District
Beijing 100032
P.R. China

Email: li_zhenqiang@hotmail.com

Jie Dong
Huawei Technologies
Huawei Campus, No.156 Beiqing Rd.
Beijing 100095
P.R. China

Email: jie.dong@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2020

Z. Li
Huawei
L. Ou
Y. Luo
China Telcom Co., Ltd.
S. Lu
Tencent
H. Chen
Futurewei
S. Zhuang
H. Wang
Huawei
July 7, 2019

BGP Extensions for Routing Policy Distribution (RPD)
draft-li-idr-flowspec-rpd-05

Abstract

It is hard to adjust traffic and optimize traffic paths on a traditional IP network from time to time through manual configurations. It is desirable to have an automatic mechanism for setting up routing policies, which adjust traffic and optimize traffic paths automatically. This document describes BGP Extensions for Routing Policy Distribution (BGP RPD) to support this.

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.

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 <https://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 January 8, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Problem Statements	3
3.1. Inbound Traffic Control	3
3.2. Outbound Traffic Control	4
4. Protocol Extensions	5
4.1. Using a New AFI and SAFI	5
4.2. BGP Wide Community	6
4.2.1. New Wide Community Atoms	6
4.3. Capability Negotiation	12
5. Consideration	12
5.1. Route-Policy	12
6. Contributors	13
7. Security Considerations	13
8. Acknowledgements	14
9. IANA Considerations	14
10. References	15
10.1. Normative References	15
10.2. Informative References	16
Authors' Addresses	16

1. Introduction

It is difficult to optimize traffic paths on a traditional IP network because of:

- o Heavy configuration and error prone. Traffic can only be adjusted device by device. All routers that the traffic traverses need to be configured. The configuration workload is heavy. The

operation is not only time consuming but also prone to misconfiguration for Service Providers.

- o Complex. The routing policies used to control network routes are complex, posing difficulties to subsequent maintenance, high maintenance skills are required.

It is desirable to have an automatic mechanism for setting up routing policies, which can simplify the routing policies configuration. This document describes extensions to BGP for Routing Policy Distribution to resolve these issues.

2. Terminology

The following terminology is used in this document.

- o ACL: Access Control List
- o BGP: Border Gateway Protocol
- o FS: Flow Specification
- o PBR: Policy-Based Routing
- o RPD: Routing Policy Distribution
- o VPN: Virtual Private Network

3. Problem Statements

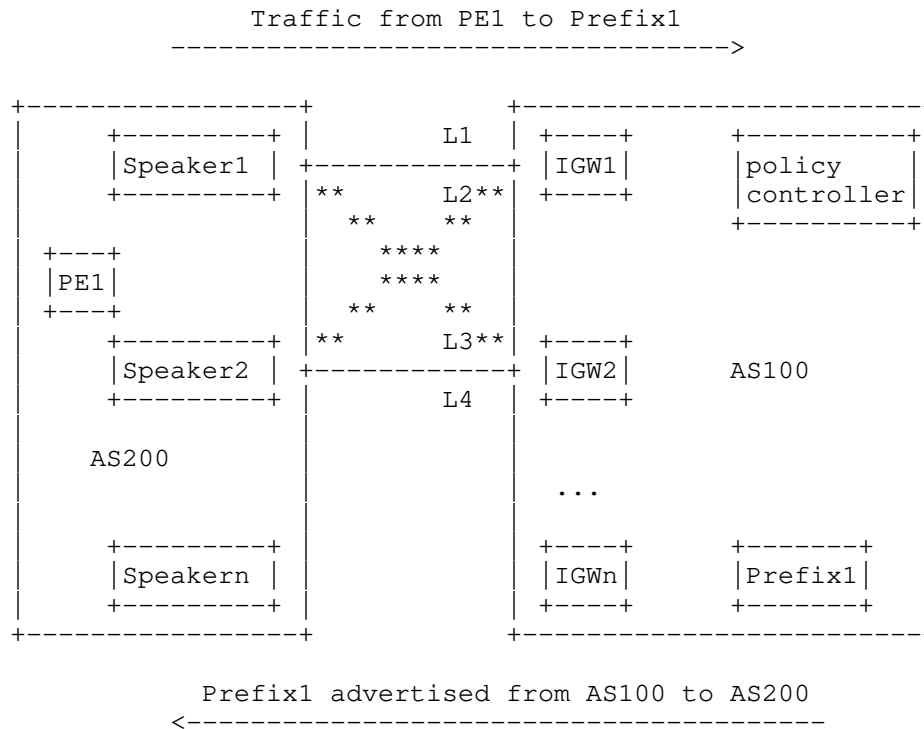
It is obvious that providers have the requirements to adjust their business traffic from time to time because:

- o Business development or network failure introduces link congestion and overload.
- o Network transmission quality is decreased as the result of delay, loss and they need to adjust traffic to other paths.
- o To control OPEX and CPEX, prefer the transit provider with lower price.

3.1. Inbound Traffic Control

In the scenario below, for the reasons above, the provider of AS100 saying P may wish the inbound traffic from AS200 enters AS100 through link L3 instead of the others. Since P doesn't have any

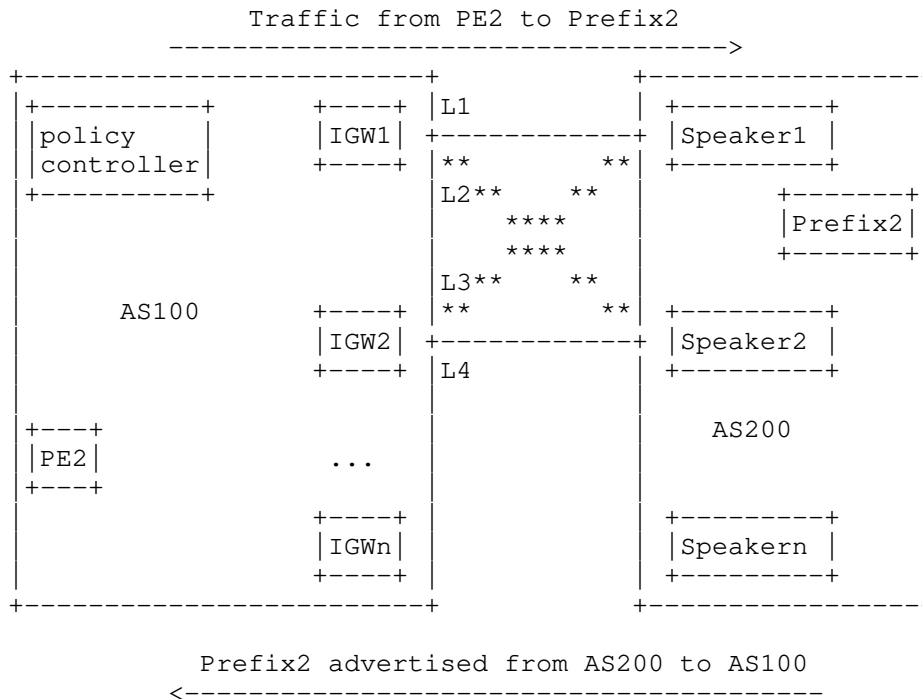
administration over AS200, so there is no way for P to modify the route selection criteria directly.



Inbound Traffic Control case

3.2. Outbound Traffic Control

In the scenario below, the provider of AS100 saying P prefers link L3 for the traffic to the destination Prefix2 among multiple exits and links. This preference can be dynamic and changed frequently because of the reasons above. So the provider P expects an efficient and convenient solution.



Outbound Traffic Control case

4. Protocol Extensions

A solution is proposed to use a new AFI and SAFI with the BGP Wide Community for encoding a routing policy.

4.1. Using a New AFI and SAFI

A new AFI and SAFI are defined: the Routing Policy AFI whose codepoint TBD1 is to be assigned by IANA, and SAFI whose codepoint TBD2 is to be assigned by IANA.

The AFI and SAFI pair uses a new NLRI, which is defined as follows:


```

      0                               1                               2                               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  NLRI Length  |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Policy Type  |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Distinguisher (4 octets) |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Peer IP (4/16 octets)      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Where:

NLRI Length: 1 octet represents the length of NLRI.

Policy Type: 1 octet indicates the type of a policy. 1 is for export policy. 2 is for import policy.

Distinguisher: 4 octet value uniquely identifies the policy in the peer.

Peer IP: 4/16 octet value indicates an IPv4/IPv6 peer.

The NLRI containing the Routing Policy is carried in a BGP UPDATE message, which MUST contain the BGP mandatory attributes and MAY also contain some BGP optional attributes.

When receiving a BGP UPDATE message, a BGP speaker processes it only if the peer IP address in the NLRI is the IP address of the BGP speaker or 0.

The content of the Routing Policy is encoded in a BGP Wide Community.

4.2. BGP Wide Community

The BGP wide community is defined in [I-D.ietf-idr-wide-bgp-communities]. It can be used to facilitate the delivery of new network services, and be extended easily for distributing different kinds of routing policies.

4.2.1. New Wide Community Atoms

A wide community Atom is a TLV (or sub-TLV), which may be included in a BGP wide community container (or BGP wide community for short) containing some BGP Wide Community TLVs. Three BGP Wide Community TLVs are defined in [I-D.ietf-idr-wide-bgp-communities], which are BGP Wide Community Target(s) TLV, Exclude Target(s) TLV, and

Parameter(s) TLV. Each of these TLVs comprises a series of Atoms, each of which is a TLV (or sub-TLV). A new wide community Atom is defined for BGP Wide Community Target(s) TLV and a few new Atoms are defined for BGP Wide Community Parameter(s) TLV. For your reference, the format of the TLV is illustrated below:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type      |      Length      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Value (variable) ~
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Format of Wide Community Atom TLV

A RouteAttr Atom TLV (or RouteAttr TLV/sub-TLV for short) is defined and may be included in a Target TLV. It has the following format.

```

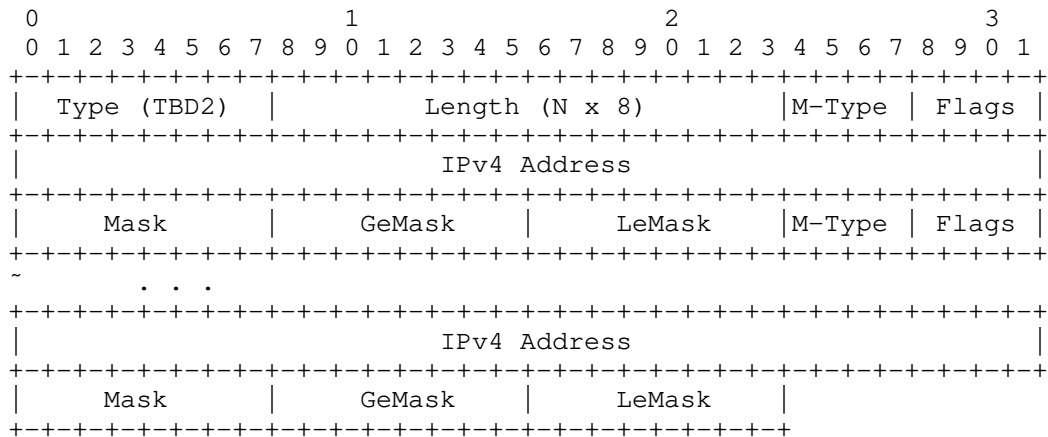
      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Type (TBD1) | Length (variable) |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     sub-TLVs ~
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Format of RouteAttr Atom TLV

The Type for RouteAttr is TBD1 (suggested value 48) to be assigned by IANA. In RouteAttr TLV, three sub-TLVs are defined: IP Prefix, AS-Path and Community sub-TLV.

An IP prefix sub-TLV gives matching criteria on IPv4 prefixes. Its format is illustrated below:



Format of IPv4 Prefix sub-TLV

Type: TBD2 (suggested value 1) for IPv4 Prefix is to be assigned by IANA.

Length: N x 8, where N is the number of tuples <M-Type, Flags, IPv4 Address, Mask, GeMask, LeMask>.

M-Type: 4 bits for match types, four of which are defined:

M-Type = 0: Exact match.

M-Type = 1: Match prefix greater and equal to the given masks.

M-Type = 2: Match prefix less and equal to the given masks.

M-Type = 3: Match prefix within the range of the given masks.

Flags: 4 bits. No flags are currently defined.

IPv4 Address: 4 octets for an IPv4 address.

Mask: 1 octet for the mask length.

GeMask: 1 octet for match range, must be less than Mask or be 0.

LeMask: 1 octet for match range, must be greater than Mask or be 0.

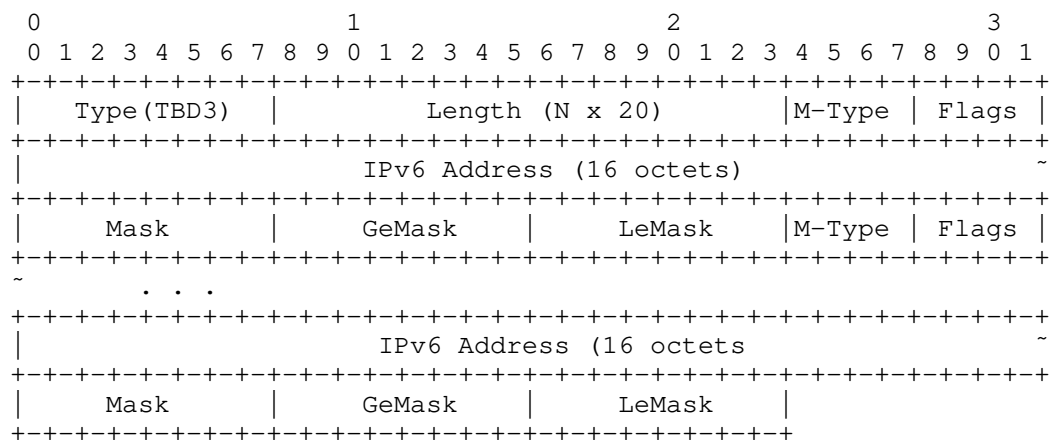
For example, tuple <M-Type=0, Flags=0, IPv4 Address = 1.1.0.0, Mask = 22, GeMask = 0, LeMask = 0> represents an exact IP prefix match for 1.1.0.0/22.

<M-Type=1, Flags=0, IPv4 Address = 16.1.0.0, Mask = 24, GeMask = 24, LeMask = 0> represents match IP prefix 1.1.0.0/24 greater-equal 24.

<M-Type=2, Flags=0, IPv4 Address = 17.1.0.0, Mask = 24, GeMask = 0, LeMask = 26> represents match IP prefix 17.1.0.0/24 less-equal 26.

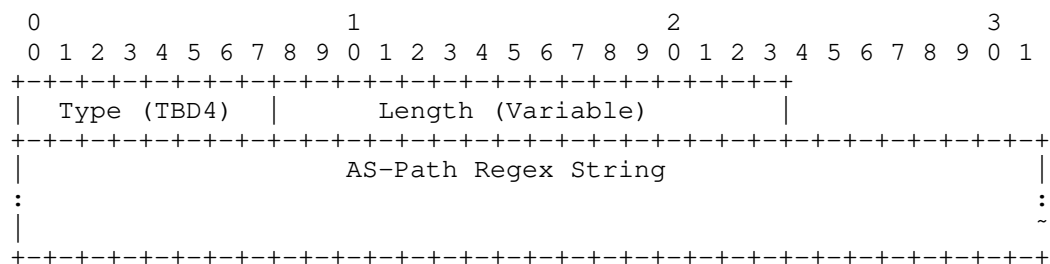
<M-Type=3, Flags=0, IPv4 Address = 18.1.0.0, Mask = 24, GeMask = 24, LeMask = 32> represents match IP prefix 18.1.0.0/24 greater-equal to 24 and less-equal 32.

Similarly, an IPv6 Prefix sub-TLV represents match criteria on IPv6 prefixes. Its format is illustrated below:



Format of IPv6 Prefix sub-TLV

An AS-Path sub-TLV represents a match criteria in a regular expression string. Its format is illustrated below:



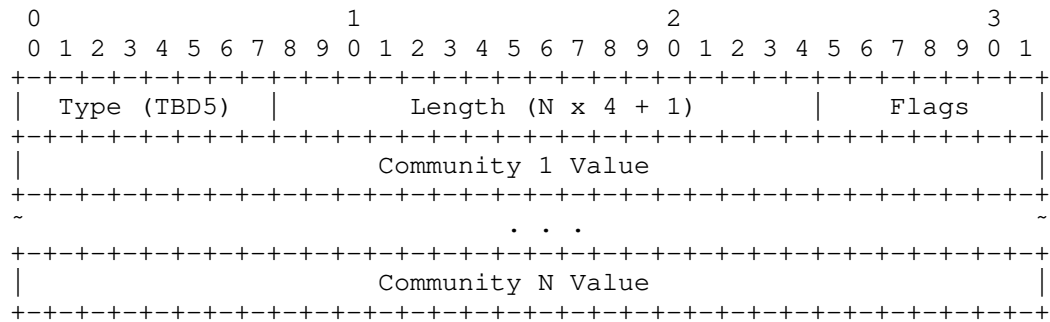
Format of AS Path sub-TLV

Type: TBD4 (suggested value 2) for AS-Path is to be assigned by IANA.

Length: Variable, maximum is 1024.

AS-Path Regex String: AS-Path regular expression string.

A community sub-TLV represents a list of communities to be matched all. Its format is illustrated below:



Format of Community sub-TLV

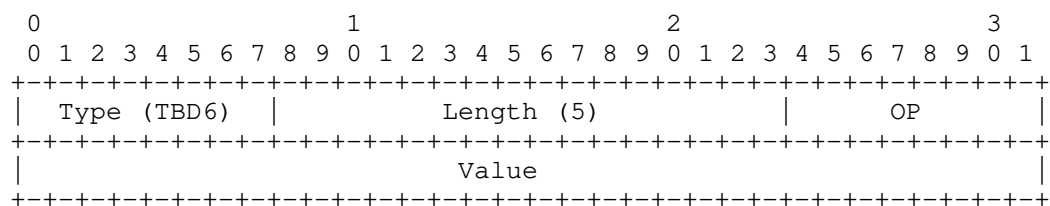
Type: TBD5 (suggested value 3) for Community is to be assigned by IANA.

Length: $N \times 4 + 1$, where N is the number of communities.

Flags: 1 octet. No flags are currently defined.

In Parameter(s) TLV, two action sub-TLVs are defined: MED change sub-TLV and AS-Path change sub-TLV. When the community in the container is MATCH AND SET ATTR, the Parameter(s) TLV includes some of these sub-TLVs. When the community is MATCH AND NOT ADVERTISE, the Parameter(s) TLV's value is empty.

A MED change sub-TLV indicates an action to change the MED. Its format is illustrated below:



Format of MED Change sub-TLV

Type: TBD6 (suggested value 1) for MED Change is to be assigned by IANA.

Length: 5.

OP: 1 octet. Three are defined:

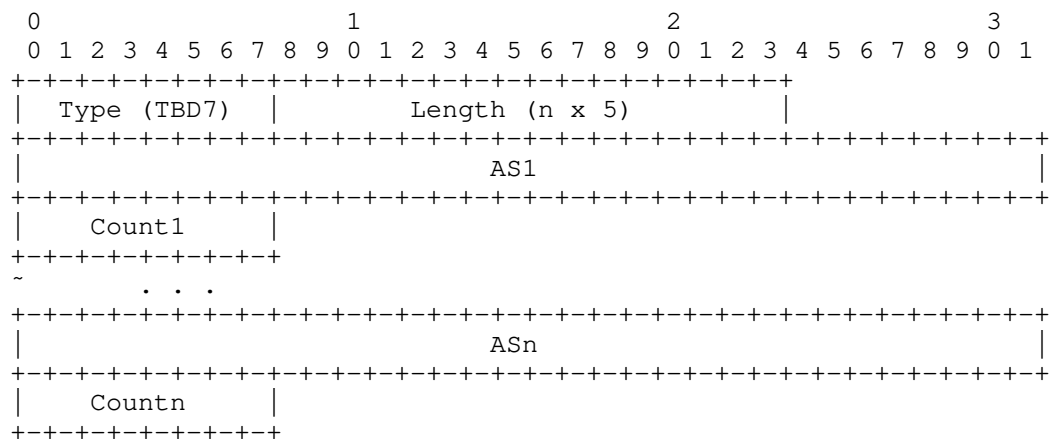
OP = 0: assign the Value to the existing MED.

OP = 1: add the Value to the existing MED. If the sum is greater than the maximum value for MED, assign the maximum value to MED.

OP = 2: subtract the Value from the existing MED. If the existing MED minus the Value is less than 0, assign 0 to MED.

Value: 4 octets.

An AS-Path change sub-TLV indicates an action to change the AS-Path. Its format is illustrated below:



Format of AS-Path Change sub-TLV

Type: TBD7 (suggested value 2) for AS-Path Change is to be assigned by IANA.

Length: n x 5.

ASi: 4 octet. An AS number.

Counti: 1 octet. ASi repeats Counti times.

The sequence of AS numbers are added to the existing AS Path.

4.3. Capability Negotiation

It is necessary to negotiate the capability to support BGP Extensions for Routing Policy Distribution (RPD). The BGP RPD Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. The Capability Length field of this capability is variable. The Capability Value field consists of one or more of the following tuples:

	Address Family Identifier (2 octets)	
	Subsequent Address Family Identifier (1 octet)	
	Send/Receive (1 octet)	

BGP RPD Capability

The meaning and use of the fields are as follows:

Address Family Identifier (AFI): This field is the same as the one used in [RFC4760].

Subsequent Address Family Identifier (SAFI): This field is the same as the one used in [RFC4760].

Send/Receive: This field indicates whether the sender is (a) willing to receive Routing Policies from its peer (value 1), (b) would like to send Routing Policies to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

5. Consideration

5.1. Route-Policy

Routing policies are used to filter routes and control how routes are received and advertised. If route attributes, such as reachability, are changed, the path along which network traffic passes changes accordingly.

When advertising, receiving, and importing routes, the router implements certain policies based on actual networking requirements to filter routes and change the attributes of the routes. Routing policies serve the following purposes:

- o Control route advertising: Only routes that match the rules specified in a policy are advertised.
- o Control route receiving: Only the required and valid routes are received. This reduces the size of the routing table and improves network security.
- o Filter and control imported routes: A routing protocol may import routes discovered by other routing protocols. Only routes that satisfy certain conditions are imported to meet the requirements of the protocol.
- o Modify attributes of specified routes: Attributes of the routes that are filtered by a routing policy are modified to meet the requirements of the local device.
- o Configure fast reroute (FRR): If a backup next hop and a backup outbound interface are configured for the routes that match a routing policy, IP FRR, VPN FRR, and IP+VPN FRR can be implemented.

Routing policies are implemented using the following procedures:

1. Define rules: Define features of routes to which routing policies are applied. Users define a set of matching rules based on different attributes of routes, such as the destination address and the address of the router that advertises the routes.
2. Implement the rules: Apply the matching rules to routing policies for advertising, receiving, and importing routes.

6. Contributors

The following people have substantially contributed to the definition of the BGP-FS RPD and to the editing of this document:

Peng Zhou
Huawei
Email: Jewpon.zhou@huawei.com

7. Security Considerations

Protocol extensions defined in this document do not affect the BGP security other than those as discussed in the Security Considerations section of [RFC5575].

8. Acknowledgements

The authors would like to thank Acee Lindem, Jeff Haas, Jie Dong, Lucy Yong, Qiandeng Liang, Zhenqiang Li for their comments to this work.

9. IANA Considerations

This document requests assigning a new AFI in the registry "Address Family Numbers" as follows:

Code Point	Description	Reference
TBD (36879 suggested)	Routing Policy AFI	This document

This document requests assigning a new SAFI in the registry "Subsequent Address Family Identifiers (SAFI) Parameters" as follows:

Code Point	Description	Reference
TBD(179 suggested)	Routing Policy SAFI	This document

This document defines a new registry called "Routing Policy NLRI". The allocation policy of this registry is "First Come First Served (FCFS)" according to [RFC8126].

Following code points are defined:

Code Point	Description	Reference
1	Export Policy	This document
2	Import Policy	This document

This document requests assigning a code-point from the registry "BGP Community Container Atom Types" as follows:

TLV Code Point	Description	Reference
TBD1 (48 suggested)	RouteAttr Atom	This document

This document defines a new registry called "Route Attributes Sub-TLV" under RouteAttr Atom TLV. The allocation policy of this registry is "First Come First Served (FCFS)" according to [RFC8126].

Following Sub-TLV code points are defined:

Code Point	Description	Reference
0	Reserved	
1	IP Prefix Sub-TLV	This document
2	AS-Path Sub-TLV	This document
3	Community Sub-TLV	This document
4 - 255	To be assigned in FCFS	

This document defines a new registry called "Attribute Change Sub-TLV" under Parameter(s) TLV. The allocation policy of this registry is "First Come First Served (FCFS)" according to [RFC8126].

Following Sub-TLV code points are defined:

Code Point	Description	Reference
0	Reserved	
1	MED Change Sub-TLV	This document
2	AS-Path Change Sub-TLV	This document
3 - 255	To be assigned in FCFS	

10. References

10.1. Normative References

- [I-D.ietf-idr-wide-bgp-communities]
 Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S.,
 and P. Jakma, "BGP Community Container Attribute", draft-
 ietf-idr-wide-bgp-communities-05 (work in progress), July
 2018.

- [RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <<https://www.rfc-editor.org/info/rfc1997>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<https://www.rfc-editor.org/info/rfc5492>>.
- [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, <<https://www.rfc-editor.org/info/rfc5575>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.

10.2. Informative References

- [I-D.ietf-idr-registered-wide-bgp-communities]
Raszuk, R. and J. Haas, "Registered Wide BGP Community Values", draft-ietf-idr-registered-wide-bgp-communities-02 (work in progress), May 2016.

Authors' Addresses

Zhenbin Li
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Liang Ou
China Telcom Co., Ltd.
109 West Zhongshan Ave, Tianhe District
Guangzhou 510630
China

Email: oul@gsta.com

Yujia Luo
China Telcom Co., Ltd.
109 West Zhongshan Ave, Tianhe District
Guangzhou 510630
China

Email: luoyuj@gsta.com

Sujian Lu
Tencent
Tengyun Building, Tower A , No. 397 Tianlin Road
Shanghai, Xuhui District 200233
China

Email: jasonlu@tencent.com

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

Shunwan Zhuang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Haibo Wang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: rainsword.wang@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2017

Z. Li
S. Zhuang
Huawei Technologies
S. Lu
Tencent
October 30, 2016

BGP Extensions for Service-Oriented MPLS Path Programming (MPP)
draft-li-idr-mpls-path-programming-04

Abstract

Service-oriented MPLS programming (SoMPP) is to provide customized service process based on flexible label combinations. BGP will play an important role for MPLS path programming to download programmed MPLS path and map the service path to the transport path. This document defines BGP extensions to support service-oriented MPLS path programming.

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.

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 May 3, 2017.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Architecture and Usecases of SoMPP	3
3.1. Architecture	3
3.2. Usecases	4
3.2.1. Deterministic ECMP	4
3.2.2. Centralized Mapping of Service to Tunnels	5
4. Advertising Label Stacks in BGP	5
4.1. Download of MPLS Path	7
4.2. Mapping Traffic to MPLS Path	7
5. Download of Mapping of Service Path to Transport Path	7
5.1. Specify Tunnel Type	7
5.2. Specify Specific Tunnel	8
6. Route Flag Extended Community	9
7. Destination Node Attribute	10
8. Capability Negotiation	11
9. Acknowledgments	12
10. IANA Considerations	12
11. Security Considerations	12
12. References	12
12.1. Normative References	12
12.2. Informative References	13
Authors' Addresses	13

1. Introduction

The label stack capability of MPLS would have been utilized well to implement flexible path programming to satisfy all kinds of service requirements. But in the distributed environment, the flexible programming capability is difficult to implement and always confined to reachability. As the introducing of central control in the network, the flexible MPLS programming capability becomes possible owing to two factors: 1. It becomes easier to allocate label for more purposes than reachability; 2. It is easy to calculate the MPLS path in a global network view. Moreover, the MPLS path programming capability can be utilized to satisfy more requirements of service

bearing in the service layer which is defined as Service-oriented MPLS path programming. BGP will play an important role for MPLS path programming to download programmed MPLS path and map the service path to the transport path. This document defines BGP extensions to support Service-oriented MPLS path programming.

2. Terminology

BGP: Border Gateway Protocol

EVPN: Ethernet VPN

L2VPN: Layer 2 VPN

L3VPN: Layer 3 VPN

MPP: MPLS Path Programming

MVPN: Multicast VPN

RR: Route Reflector

SR-Path: Segment Routing Path

NLRI: Network Layer Reachability Information

3. Architecture and Usecases of SoMPP

3.1. Architecture

The architecture of BGP-based MPLS path programming is shown in the Figure 1. Central control plays an important role in MPLS path programming. It can extend the MPLS path programming capability easily. The central controller can calculate path in a global network view and implement the MPLS path programming to satisfy different requirements of services. The result of MPLS path programming can be advertised from the central controller to the client nodes through BGP extensions to the ingress PEs. When client nodes receives the result of MPLS path programming, it will install the MPLS forwarding entry for the specified BGP prefix to implement the service process.

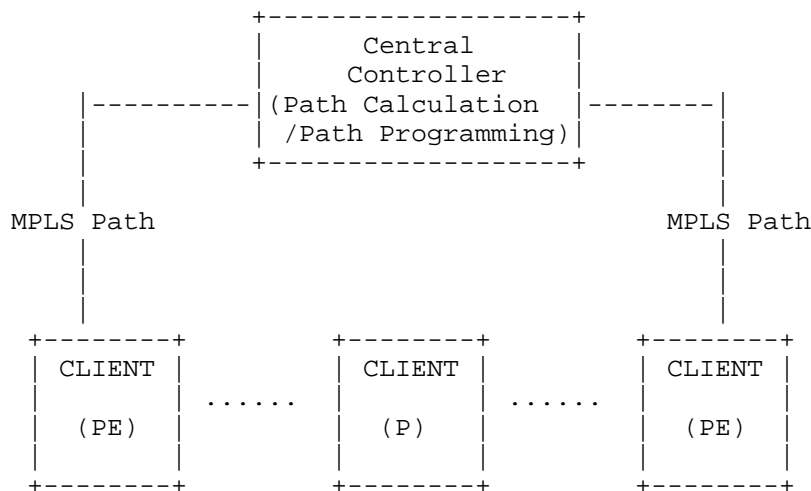


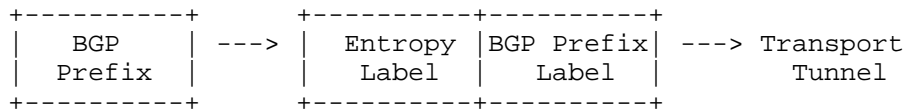
Figure 1 BGP-based MPLS Path Programming

3.2. Usecases

3.2.1. Deterministic ECMP

Entropy Label[RFC6790] is introduced to improve the ECMP capability by encapsulate the entropy label in the MPLS label stack. The existing implementation is always to calculate the entropy label based on the header of packets by specific hash algorithm in the ingress node. That is, the entropy label is determined locally by the ingress node. The method can improve the hash of packets in the network for load-sharing. But since the ingress node lacks the knowledge of the global traffic pattern of the network and calculates the entropy label by itself it may be not able to improve the ECMP capability accurately and in some cases it may deteriorate the imbalance of load-sharing.

With the central controlled MPLS path programming, the central controller can collect the global traffic pattern information of the network and based on the information deterministically calculate the entropy label for specific flows to help improve the load-sharing of the network. Then the central controller can download the label stack information with the deterministic entropy label to the ingress PEs for the specific BGP prefix. The ingress node can install the MPLS forwarding entry shown in the following figure to help optimize the ECMP of the flow specified by the BGP prefix, then optimize the ECMP of the whole network.



3.2.2. Centralized Mapping of Service to Tunnels

In the network there can be multiple tunnels to one specific destination which satisfy different constraints. In the traditional way, the tunnel is set up by the distributed forwarding nodes. As the PCE-initiated LSP setup [I-D.ietf-pce-pce-initiated-lsp] is introduced, the tunnel with different constraints can be set up in the central controlled way. In order to satisfy different service requirements, it is necessary to provide the capability to flexibly map the service to different tunnels which constraints can satisfy the required service requirement. Since the central controller has enough information of the whole network view, it can be an effective way to map the service (such as L3VPN and L2VPN) to the tunnel by the central controller and advertise the mapping information to the ingress PE of the service to guide the mapping in the forwarding node.

There can be two types of behaviors to map service to the tunnel:

1. Specify the tunnel type: with the method BGP will carry the tunnel type information for the BGP prefix. When the ingress PE receives the information, it will use the tunnel type and the nexthop address (or other specified target IP address) to search the corresponding tunnels to bear the flow specified by the BGP prefix. If there are more than one tunnels, the ingress PE will load share the traffic across all the tunnels.
2. Specify the specific tunnel: For MPLS TE/SR-TE tunnel, there can be multiple MPLS TE tunnels from one ingress PE to a specific destination with different constraints. BGP can carry the tunnel identifier information for the BGP prefix from the controller to the ingress node. When the ingress PE receives the information, it will use the tunnel identifier information to search the corresponding tunnels to bear the flow specified by the BGP prefix. If there are multiple tunnel identifiers, the ingress PE will load share the traffic across all the tunnels.

4. Advertising Label Stacks in BGP

According to the service requirements, the central controller can combine MPLS labels flexibly. Then it can download the service label combination for specific prefix. BGP extensions are necessary to advertise label stacks for the prefix in NLRI field.

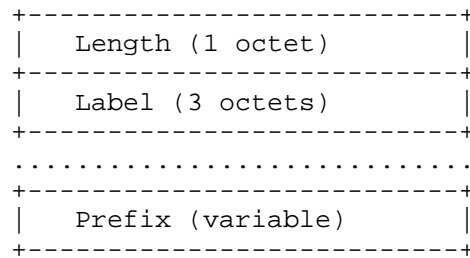


Figure 2: NLRI Definition in RFC3107

[RFC3107] defines above NLRI to advertise label binding for specific prefix. The label field can carry one or more labels. Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack". But for the other AFI/SAFIs using label binding such as IPv4 Flowspec, IPv6 Flowspec, VPNv4, VPNv6, EVPN, MVPN, etc., it dose not support the capability to carry more labels for the specific prefix. Moreover for the AFI/SAFIs which do not support label binding capability originally, but may possibly adopt MPLS path programming now, there is no label field in the NLRI. In order to support flexible MPLS path programming, this document defines and uses a new BGP attribute called the "Extended Label attribute". This is an optional transitive BGP attribute. The attribute type code is (TBA by IANA), the value field of this attribute is defined as follows:

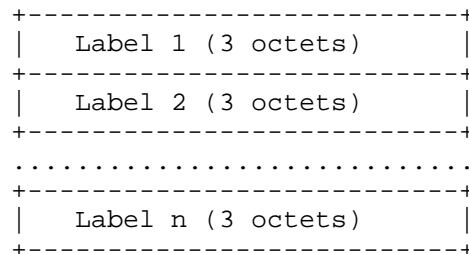


Figure 3: Extended Label Attribute

The Label field carries one or more labels (that corresponds to the stack of labels [[RFC3032]]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [[RFC3032]]). In the last label, the S bit MUST be "1"; in the other labels, the S bit MUST be "0".

The "Extended Label attribute" can be used for various BGP address families. Before using this attribute, firstly, it is necessary to

negotiate the capability between two nodes to support MPLS path programming for a specific BGP address family. If negotiation fails, a node MUST NOT send this attribute and MUST discard this attribute when it receives.

4.1. Download of MPLS Path

The Central Controller for MPLS path programming could build a route with Extended Label attribute and send it to the ingress routers.

Upon receiving such a route from the Central Controller, the ingress router SHOULD select such a route as the best path. If a packet comes into the ingress router and uses such a path, the ingress router will encapsulate the stack of labels which is derived from the Extended Label Attribute of the route into the packet and forward the packet along the path.

4.2. Mapping Traffic to MPLS Path

The Extended Label attribute can be used for BGP Flowspec address families. BGP advertises the Flowspec with the Extended Label attribute, so the flow packets can be redirected to the MPLS Path which is derived from the Extended Label Attribute.

5. Download of Mapping of Service Path to Transport Path

5.1. Specify Tunnel Type

[I-D.ietf-idr-tunnel-encaps] proposes the Tunnel Encapsulation Attribute which can be used without BGP Encapsulation SAFI to specify a set of tunnels. It defines a series of Encapsulation Sub-TLVs for particular tunnel types. It also defines the Remote Endpoint Attributes Sub-TLV to specify the remote tunnel endpoint address for each tunnel which can be different the BGP nexthop. The Tunnel Encapsulation Attributes can be reused for the MPLS path programming to specify the tunnel types, the encapsulation and the remote tunnel endpoint address which can determine a set of tunnels which the service can map to. Now the limited MPLS tunnel types are defined for the Tunnel Encapsulation Attributes. In order to support MPLS path programming, the following MPLS tunnel types are to be defined:

Value	Tunnel Type
-----	-----
TBD	LDP LSP
TBD	RSVP-TE LSP
TBD	MPLS-based Segment Routing Best-effort Path
TBD	MPLS-based Segment Routing Traffic Engineering Path

5.2. Specify Specific Tunnel

Besides specifying the tunnel types to determine the set of tunnels which the service traffic can map to, the specific tunnels can be specified directly by the tunnel identifiers when map the service traffic to the path. BGP extensions is necessary that through the community attribute of BGP the identifier of the transport path can be carried when advertise the specific prefix.

In order to support the application, this document defines a new BGP attribute called the "Extended Unicast Tunnel attribute". This is an optional transitive BGP attribute. The attribute type code is (TBA by IANA), the value field of this attribute is defined as follows:

```
+-----+
| First Tunnel entry (variable) |
+-----+
| Second Tunnel entry (variable) |
+-----+
| ... |
+-----+
| N-th Tunnel entry (variable) |
+-----+
```

The Tunnel entry is defined as follows:

```
+-----+
| Flags (1 octet) |
+-----+
| Tunnel Type (1 octets) |
+-----+
| Tunnel Identifier (variable) |
+-----+
| Tunnel Specific Attributes (Variable)(Optional) |
+-----+
```

The Flags is reserved and must be set as zero. The Tunnel Type identifies the type of the tunneling technology used for the unicast service path. The tunnel type determines the syntax and semantics of the Tunnel Identifier field. This document defines following Tunnel Types:

- + 0 - No tunnel information present
- + 1 - RSVP-TE LSP

+ 2 - MPLS-based Segment Routing Traffic Engineering Path

Tunnel Specific Attributes contains the attributes of the tunnel. The field is optional. The value depends on the tunnel type. It will be defined in the future versions.

When the Tunnel Type is set to "No tunnel information present", the Tunnel attribute carries no tunnel information (no Tunnel Identifier). when the type is used, the tunnel used for the service path is determined by the ingress router.

When the Tunnel Type is set to RSVP - Traffic Engineering (RSVP-TE) Label Switched Path (LSP), the Tunnel Identifier is <C-Type, Tunnel Sender Address, Tunnel ID, Tunnel End-point Address> as specified in [RFC3209] If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The other fields in the RSVP-TE LSP Identifier are the same as specified in [RFC3209].

When the Tunnel Type is set to MPLS-based Segment Routing Traffic Engineering Path, the Tunnel Identifier is <C-Type, Tunnel Sender Address, Tunnel ID, Tunnel End-point Address>. If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The tunnel identifier is similar as that of RSVP-TE LSP.

BGP can carry multiple Tunnel entries in one Extended Unicast Tunnel attribute for specific prefix. If there are multiple tunnel entries, the ingress PE can load share the traffic across all the specified tunnels for the service traffic determined by the specific BGP prefix, or selects the primary / Backup tunnels from the multiple tunnel entries.

The "Redirect-to-Tunnel Action" for BGP Flowspec has been described in[I-D.hao-idr-flowspec-redirect-tunnel]. This document reuses the tunnel identifier and defines it in the Extended Unicast Tunnel attribute which can be used for "Redirect-to-Tunnel Action".

6. Route Flag Extended Community

In order to make the MPLS path programming to take effect, the route advertised by the central controller after the MPLS Path Programming should be selected by the ingress PE over other routes for the same BGP prefix. There are two options of BGP extensions for the purpose:

Option 1: A new BGP Extended Community called as the "Route Flag Extended Community" can be introduced. The Type value is to be assigned by IANA.

The Route Flag Extended Community is used to carry the flag appointed by the BGP central controller.

The format of this extended community is defined as follows:

0	1	2	3	4	5	6	7
+-----+-----+-----+-----+-----+-----+-----+-----+							
Type		Reserved				Flag	
+-----+-----+-----+-----+-----+-----+-----+-----+							

Flag = 0, Treat as normal route

Flag = 1, Treat as best route

When a router receives a BGP route with a Route Flag Extended Community and the Flag set to "1", it SHOULD use the route as the best route when select the route from multiple routes for a specific prefix.

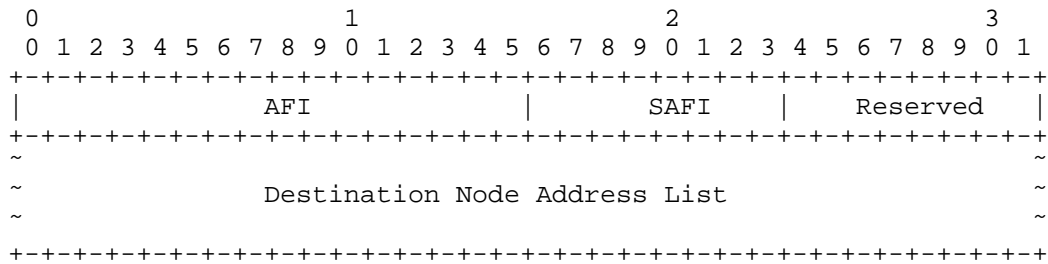
Option 2: [I-D.ietf-idr-custom-decision] defines a new Extended Community, called the Cost Community, which can be used in tie breaking during the best path selection process. The Cost Community can be reused by the MPLS path programming to set the "Point of Insertion" as 128 to make the route advertised by the central controller to be chosen.

7. Destination Node Attribute

This document defines and uses a new BGP attribute called as the "Destination Node attribute" which Type value is to be assigned by IANA. The Destination Node attribute is an optional non-transitive attribute that can be applied to any address family.

The Destination Node attribute is used to carry a list of node addresses, which are intended to be used to determine the nodes where the route with such attribute SHOULD be considered. If a node receives a BGP route with a Destination Node attribute, it MUST check the node address list. If one address of the list belongs to this node, the route MUST be used in this node. Otherwise the route MUST be ignored silently.

The format of this attribute is defined as follows:



AFI: Address Family Identifier (16 bits).

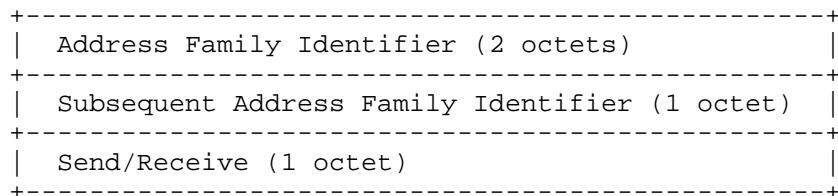
SAFI: Subsequent Address Family Identifier (8 bits).

Reserved: One octet reserved for special flags

Destination Node Address List: The list of IPv4 (AFI=1) or IPv6 (AFI=2) address.

8. Capability Negotiation

It is necessary to negotiate the capability to support MPLS path programming. The MPLS-Path-Programming Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. The Capability Length field of this capability is variable. The Capability Value field consists of one or more of the following tuples:



The meaning and use of the fields are as follows:

Address Family Identifier (AFI): This field is the same as the one used in [RFC4760].

Subsequent Address Family Identifier (SAFI): This field is the same as the one used in [RFC4760].

Send/Receive: This field indicates whether the sender is (a) willing to receive programming MPLS paths from its peer (value 1), (b) would

like to send programming MPLS paths to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

9. Acknowledgments

The authors of this document would like to thank Lucy Yong, Susan Hares, Eric Wu, Weiguo Hao, Pingan Li, Zhengqiang Li and Jie Dong for their reviews and comments of this document.

10. IANA Considerations

TBD.

11. Security Considerations

The security considerations of [RFC4271] and [RFC5575] are applicable.

12. References

12.1. Normative References

- [I-D.hao-idr-flowspec-redirect-tunnel]
Weiguo, H., Li, Z., and L. Yong, "BGP Flow-Spec Redirect to Tunnel Action", draft-hao-idr-flowspec-redirect-tunnel-01 (work in progress), March 2016.
- [I-D.ietf-idr-custom-decision]
Retana, A. and R. White, "BGP Custom Decision Process", draft-ietf-idr-custom-decision-07 (work in progress), November 2015.
- [I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-02 (work in progress), May 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>>.
- [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>>.

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<http://www.rfc-editor.org/info/rfc3209>>.
- [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>>.
- [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>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.
- [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>>.

12.2. Informative References

- [I-D.ietf-pce-pce-initiated-lsp] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-07 (work in progress), July 2016.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <<http://www.rfc-editor.org/info/rfc3107>>.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <<http://www.rfc-editor.org/info/rfc6790>>.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Sujian Lu
Tencent
Tengyun Building, Tower A ,No. 397 Tianlin Road
Shanghai, Xuhui District 200233
China

Email: jasonlu@tencent.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 24, 2017

S. Previdi, Ed.
C. Filsfils
Cisco Systems, Inc.
P. Mattes
Microsoft
E. Rosen
Juniper Networks
S. Lin
Google
June 22, 2017

Advertising Segment Routing Policies in BGP
draft-previdi-idr-segment-routing-te-policy-07

Abstract

This document defines a new BGP SAFI with a new NLRI in order to advertise a candidate path of a Segment Routing Policy (SR Policy). An SR Policy is a set of candidate paths consisting of one or more segment lists. The headend of an SR Policy may learn multiple candidate paths for an SR Policy. Candidate paths may be learned via a number of different mechanisms, e.g., CLI, NetConf, PCEP, or BGP. This document specifies the way in which BGP may be used to distribute candidate paths. New sub-TLVs for the Tunnel Encapsulation Attribute are defined.

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 December 24, 2017.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Requirements Language	5
2. SR TE Policy Encoding	5
2.1. SR TE Policy SAFI and NLRI	5
2.2. SR TE Policy and Tunnel Encapsulation Attribute	7
2.3. Remote Endpoint and Color	8
2.4. SR TE Policy Sub-TLVs	8
2.4.1. Preference sub-TLV	8
2.4.2. SR TE Binding SID Sub-TLV	9
2.4.3. Segment List Sub-TLV	10
3. Extended Color Community	21
4. SR Policy Operations	21
4.1. Configuration and Advertisement of SR TE Policies	22
4.2. Reception of an SR Policy NLRI	22
4.2.1. Acceptance of an SR Policy NLRI	22
4.2.2. Usable SR Policy NLRI	23
4.2.3. Passing a usable SR Policy NLRI to the SRTE Process	24
4.2.4. Propagation of an SR Policy	24
4.3. Flowspec and SR Policies	24
5. Contributors	24
6. Acknowledgments	25
7. Implementation Status	25
8. IANA Considerations	26
8.1. Existing Registry: Subsequent Address Family Identifiers (SAFI) Parameters	26
8.2. Existing Registry: BGP Tunnel Encapsulation Attribute Tunnel Types	26
8.3. Existing Registry: BGP Tunnel Encapsulation Attribute sub-TLVs	27
8.4. New Registry: SR Policy List Sub-TLVs	27
9. Security Considerations	27

10. References	27
10.1. Normative References	27
10.2. Informational References	28
Authors' Addresses	29

1. Introduction

Segment Routing (SR) allows a headend node to steer a packet flow along any path. Intermediate per-flow states are eliminated thanks to source routing [I-D.ietf-spring-segment-routing].

The headend node is said to steer a flow into an Segment Routing Policy (SR Policy).

The header of a packet steered in an SR Policy is augmented with the ordered list of segments associated with that SR Policy.

[I-D.filsfils-spring-segment-routing-policy] details the concepts of SR Policy and steering into an SR Policy. These apply equally to the MPLS and SRv6 instantiations of segment routing.

As highlighted in section 2 of [I-D.filsfils-spring-segment-routing-policy]:

- o an SR policy may have multiple candidate paths learned via various mechanisms (CLI, NetConf, PCEP or BGP);
- o the SRTE process selects the best candidate path for a Policy;
- o the SRTE process binds a BSID to the selected path of the Policy;
- o the SRTE process installs the selected path and its BSID in the forwarding plane.

This document specifies the way to use BGP to distribute one or more of the candidate paths of an SR policy to the headend of that policy. The SRTE process ([I-D.filsfils-spring-segment-routing-policy]) of the headend receives candidate paths from BGP, and possibly other sources as well, and the SRTE process then determines the selected path of the policy.

This document specifies a way of representing SR policies and their candidate paths in BGP UPDATE messages. BGP can then be used to propagate the SR policies and candidate paths. The usual BGP rules for BGP propagation and "bestpath selection" are used. At the headend of a specific policy, this will result in one or more candidate paths being installed into the "BGP table". These paths are then passed to the SRTE process. The SRTE process may compare

them to candidate paths learned via other mechanisms, and will choose one or more paths to be installed in the data plane. BGP itself does not install SRTE candidate paths into the data plane.

This document defines a new BGP address family (SAFI). In UPDATE messages of that address family, the NLRI identifies an SR policy, and the attributes specify candidate paths of that policy.

While for simplicity we may write that BGP advertises an SR Policy, it has to be understood that BGP advertises a candidate path of an SR policy and that this SR Policy might have several other candidate paths provided via BGP (via an NLRI with a different distinguisher as defined in this document), PCEP, NETCONF or local policy configuration.

Typically, a controller defines the set of policies and advertise them to policy head-end 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 a specific policy head-end. In this case the advertisement may sent on a BGP session to that head-end and not propagated any further.

Alternatively, a router (i.e.: an BGP egress router) advertises SR Policies representing paths to itself. In this case, it is possible to send the policy to each head-end over a BGP session to that head-end, without requiring any further propagation of the policy.

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

In some situations, it is undesirable for a controller or BGP egress router to have a BGP session to each policy head-end. In these situations, BGP Route Reflectors may be used to propagate the advertisements, or it may be necessary for the advertisement to propagate through a sequence of one or more ASes. To make this possible, an attribute needs to be attached to the advertisement that enables a BGP speaker to determine whether it is intended to be a head-end for the advertised policy. This is done by attaching one or more Route Target Extended Communities to the advertisement ([RFC4360]).

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

- o A new Subsequent Address Family Identifier (SAFI) whose NLRI identifies an SR Policy.

- o A set of new TLVs to be inserted into the Tunnel Encapsulation Attribute (as defined in [I-D.ietf-idr-tunnel-encaps]) specifying candidate paths of the SR policy, as well as other information about the SR policy.
- o One or more IPv4 address format route-target extended community ([RFC4360]) attached to the SR Policy advertisement and that indicates the intended head-end of such SR Policy advertisement.
- o The Color Extended Community (as defined in [I-D.ietf-idr-tunnel-encaps]) and used in order to steer traffic into an SR Policy, as described in [I-D.filsfils-spring-segment-routing-policy]. This document (Section 3) modifies the format of the Color Extended 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 8) from the "Subsequent Address Family Identifiers (SAFI) Parameters" registry).

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

+-----+	
NLRI Length	1 octet
+-----+	
Distinguisher	4 octets
+-----+	
Policy Color	4 octets
+-----+	
Endpoint	4 or 16 octets
+-----+	

where:

- o NLRI Length: 1 octet of length expressed in bits as defined in [RFC4760].

- o Distinguisher: 4-octet value uniquely identifying the policy in the context of <color, endpoint> tuple. The distinguisher has no semantic value and is solely used by the SR Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR Policy.
- o Policy Color: 4-octet value identifying (with the endpoint) the policy. The color is used to match the color of the destination prefixes to steer traffic into the SR Policy [I-D.filsfils-spring-segment-routing-policy].
- o 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 color and endpoint are used to automate the steering of BGP Payload prefixes on SR policy ([I-D.filsfils-spring-segment-routing-policy]).

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 73 (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 among the best paths as per the BGP best path selection algorithm. In other words, this document does not modify the BGP propagation or bestpath selection rules.

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, it is recommended that each NLRI use a different distinguisher. If BGP has installed into the BGP table two advertisements whose respective

NLRIs have the same color and endpoint, but different distinguishers, both advertisements are passed to the SRTE process.

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 8) 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:

- o SR Policy SAFI NLRI is defined in Section 2.1.
- o Tunnel Encapsulation Attribute is defined in [I-D.ietf-idr-tunnel-encaps].
- o Tunnel-Type is set to 15 (assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).
- o Preference, Binding SID, Segment-List, Weight and Segment are defined in this document.
- o Additional sub-TLVs may be defined in the future.

A Tunnel Encapsulation Attribute MUST NOT contain more than one TLV of type "SR Policy".

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.

- o Flags: 1 octet of flags. None are defined at this stage. Flags SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o Preference: a 4-octet value. The highest value is preferred.

2.4.2. SR TE Binding SID Sub-TLV

The Binding SID sub-TLV is not used by BGP. The contents of this sub-TLV are used by the SRTE process ([I-D.filsfils-spring-segment-routing-policy]).

The Binding SID sub-TLV is optional, MUST NOT appear more than once in the SR Policy and has the following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type      |      Length      |      Flags      |      RESERVED      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Binding SID (variable, optional)                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

where:

- o Type: TBD4 (to be assigned by IANA from the "BGP Tunnel Encapsulation Attribute sub-TLVs" registry).
- o Length: specifies the length of the value field not including Type and Length fields. Can be 2 or 6 or 18.
- o Flags: 1 octet of flags. None are defined at this stage. Flags SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o 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.

2.4.3. Segment List Sub-TLV

The Segment List TLV encodes 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 8 for details of codepoints allocation.

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

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

The Segment List sub-TLV has the following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type      |      Length      |  RESERVED  |
+-----+-----+-----+-----+-----+-----+-----+
//                               sub-TLVs                               //
+-----+-----+-----+-----+-----+-----+-----+

```

where:

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

2.4.3.1. Weight Sub-TLV

The Weight sub-TLV specifies the weight associated to a given candidate path (i.e.: a given segment list). The contents of this sub-TLV are used only by the SRTE process ([I-D.filsfils-spring-segment-routing-policy]).

The Weight sub-TLV is optional, MUST NOT appear more than once inside the Segment List sub-TLV, and has the following format:

0						1						2						3													
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
Type						Length						Flags						RESERVED													
Weight																															

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 are defined at this stage. Flags SHOULD be set to zero 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.

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.

The Segment sub-TLV does not have any effect on the BGP bestpath selection or propagation procedures. The contents of this sub-TLV are used only by the SRTE process ([I-D.filsfils-spring-segment-routing-policy]).

[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:

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
Type										Length										Flags										RESERVED									
Label																				TC			S	TTL															

where:

- o Type: 1 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- o Length is 6.
- o Flags: 1 octet of flags. None are defined at this stage. Flags SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o Label: 20 bits of label value.
- o TC: 3 bits of traffic class.
- o S: 1 bit of bottom-of-stack.
- o TTL: 1 octet of TTL.

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

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

- o If the originator wants the receiver to choose the TC value, it sets the TC field to zero.
- o If the originator wants the receiver to choose the TTL value, it sets the TTL field to 255.
- o If the originator wants to recommend a value for these fields, it puts those values in the TC and/or TTL fields.
- o 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:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type      |      Length      |      Flags      | RESERVED |
+-----+-----+-----+-----+-----+-----+-----+-----+
//                               IPv6 SID (16 octets)                               //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

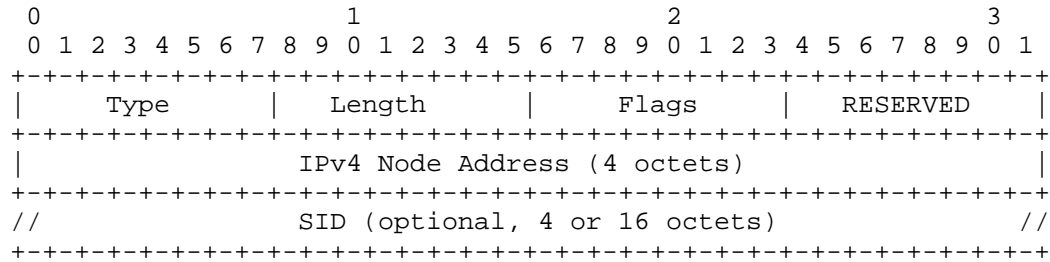
where:

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



where:

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

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

- o 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 address.
- o If length is 6, then only the IPv4 Node Address is present.

- o If length is 10, then the IPv4 Node Address and the MPLS SID are present.
- o 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:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type      |      Length      |      Flags      |  RESERVED  |
+-----+-----+-----+-----+-----+-----+-----+-----+
//                IPv6 Node Address (16 octets)                //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                SID (optional, 4 or 16 octets)                //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

where:

- o Type: 4 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- o Length is 18 or 22 or 34.
- o Flags: 1 octet of flags. None are defined at this stage. Flags SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o IPv6 Node Address: a 16 octet IPv6 address representing a node.
- o SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

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

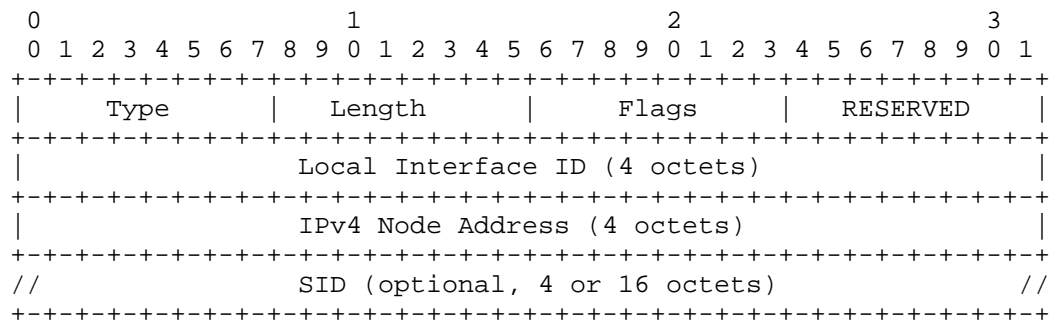
- o The IPv6 Node Address MUST be present.
- o The SID is optional and MAY be of one of the following formats:
 - * MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.

* IPV6 SID: a 16 octet IPv6 address.

- o If length is 18, then only the IPv6 Node Address is present.
- o If length is 22, then the IPv6 Node Address and the MPLS SID are present.
- o If length is 34, then the IPv6 Node Address and the IPv6 SID are present.

2.4.3.2.5. Type 5: IPv4 Address + Local Interface ID with optional SID

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



where:

- o Type: 5 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- o Length is 10 or 14 or 26.
- o Flags: 1 octet of flags. None are defined at this stage. Flags SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o Local Interface ID: 4 octets as defined in [I-D.ietf-pce-segment-routing].
- o IPv4 Node Address: a 4 octet IPv4 address representing a node.

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

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

- o The IPv4 Node Address MUST be present.
- o The Local Interface ID 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 Local Interface ID are present.
- o If length is 14, then the IPv4 Node Address, the Local Interface ID and the MPLS SID are present.
- o If length is 26, then the IPv4 Node Address, the Local Interface ID 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 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:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type      |      Length      |      Flags      |  RESERVED  |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Local IPv4 Address (4 octets)               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Remote IPv4 Address (4 octets)               |
+-----+-----+-----+-----+-----+-----+-----+-----+
//               SID (4 or 16 octets)                       //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

where:

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

- o Length is 10 or 14 or 26.
- o Flags: 1 octet of flags. None are defined at this stage. Flags SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o Local IPv4 Address: a 4 octet IPv4 address.
- o Remote IPv4 Address: a 4 octet IPv4 address.
- o SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

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

- o The Local IPv4 Address MUST be present and represents an adjacency local address.
- o The Remote IPv4 Address MUST be present and represents the remote end of the adjacency.
- 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 address.
- o If length is 10, then only the IPv4 Local and Remote addresses are present.
- o If length is 14, then the IPv4 Local address, IPv4 Remote address and the MPLS SID are present.
- o 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 + Local Interface ID with optional SID

The Type-7 Segment Sub-TLV encodes an IPv6 node address, a local interface identifier (Local Interface ID) and an optional SID in the form of either an MPLS label or an IPv6 address. The format is as follows:

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   Type   |   Length   |   Flags   |   RESERVED   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Local Interface ID (4 octets) |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
//                                     IPv6 Node Address (16 octets) //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
//                                     SID (optional, 4 or 16 octets) //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

where:

- o Type: 7 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- o Length is 22 or 26 or 38.
- o Flags: 1 octet of flags. None are defined at this stage. Flags SHOULD be set to zero on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o Local Interface ID: 4 octets of interface index.
- o IPv6 Node Address: a 16 octet IPv6 address representing a node.
- o SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

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

- o The IPv6 Node Address MUST be present.
- o The Local Interface ID 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 address.
- o If length is 22, then the IPv6 Node Address and Local Interface ID are present.

- o If length is 26, then the IPv6 Node Address, the Local Interface ID and the MPLS SID are present.
- o If length is 38, then the IPv6 Node Address, the Local Interface ID 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 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:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type      |      Length      |      Flags      | RESERVED |
+-----+-----+-----+-----+-----+-----+-----+-----+
//                Local IPv6 Address (16 octets)                //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                Remote IPv6 Address (16 octets)                //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                SID (4 or 16 octets)                          //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

where:

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

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

- o The Local IPv6 Address MUST be present and represents an adjacency local address.
- o The Remote IPv6 Address MUST be present and represents the remote end of the adjacency.
- 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 address.
- o If length is 34, then only the IPv6 Local and Remote addresses are present.
- o If length is 38, then the IPv6 Local address, IPv4 Remote address and the MPLS SID are present.
- o If length is 50, then the IPv6 Local address, IPv4 Remote address and the IPv6 SID are present.

3. Extended Color Community

The Color Extended Community as defined in [I-D.ietf-idr-tunnel-encaps] is used to steer traffic into a policy.

When the Color Extended Community is used for the purpose of steering the traffic into an SRTE policy, the RESERVED field (as defined in [I-D.ietf-idr-tunnel-encaps] is changed as follows:

```

      1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|C O|          RESERVED          |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

where CO bits are defined as the "Color-Only" bits.

[I-D.filsfils-spring-segment-routing-policy] defines the influence of these bits on the automated steering of BGP Payload traffic onto SRTE policies.

4. SR Policy Operations

As described in this document, the consumer of a SR Policy NLRI is not the BGP process. The BGP process is in charge of the origination and propagation of the SR Policy NLRI but its installation and use is

outside the scope of BGP
([I-D.filsfils-spring-segment-routing-policy]).

4.1. Configuration and Advertisement of SR TE Policies

Typically, but not limited to, an SR Policy is configured into a controller.

Multiple SR Policy NLRIs may be present with the same <color, endpoint> tuple but with different content when these SR policies are intended to different head-ends.

The distinguisher of each SR Policy NLRI prevents undesired BGP route selection among these SR Policy NLRIs and allow their propagation across route reflectors [RFC4456].

Moreover, one or more route-target SHOULD be attached to the advertisement, where each route-target identifies one or more intended head-ends for the advertised SR policy.

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 a BGP session) 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 NLRI

On reception of an SR Policy NLRI, a BGP speaker MUST determine if it's first acceptable, then it determines if it is usable.

4.2.1. Acceptance of an SR Policy NLRI

When a BGP speaker receives an SR Policy NLRI from a neighbor it has to determine if it's acceptable. The following applies:

- o The SR Policy NLRI MUST include a distinguisher, color and endpoint field which implies that the length of the NLRI MUST be either 12 or 24 octets (depending on the address family of the endpoint). If the NLRI is not one of the legal lengths, a router supporting this document and that imports the route MUST consider it to be malformed and MUST apply the "treat-as-withdraw" strategy of [RFC7606].
- o The SR Policy update MUST have either the NO_ADVERTISE community or at least one route-target extended community in IPv4-address format. If a router supporting this document receives an SR policy update with no route-target extended communities and no

NO_ADVERTISE community, the update MUST NOT be sent to the SRTE process. Furthermore, it SHOULD be considered to be malformed, and the "treat-as-withdraw" strategy of [RFC7606] applied.

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

A router that receives an SR Policy update that is not valid according to these criteria MUST treat the update as malformed. The route MUST NOT be passed to the SRTE process, and the "treat-as-withdraw" strategy of [RFC7606].

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

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

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

4.2.2. Usable SR Policy NLRI

If one or more route-targets are present, then at least one route-target MUST match one of the BGP Identifiers 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 any of the local BGP Identifiers, then, while the SR Policy NLRI is acceptable, it 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 NLRI to its neighbors as defined in Section 4.2.4.

Usable SR Policy NLRIs are sent to the Segment Routing Traffic Engineering (SRTE) process. The description of the SRTE process is outside the scope of this document and it's described in [I-D.filsfils-spring-segment-routing-policy].

4.2.3. Passing a usable SR Policy NLRI to the SRTE Process

Once BGP has determined that the SR Policy NLRI is usable, BGP passes the path to the SRTE process ([I-D.filsfils-spring-segment-routing-policy]).

The SRTE process 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.

4.2.4. Propagation of an SR Policy

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

However, a node MAY be explicitly configured to advertise a received SR Policy NLRI to neighbors according to normal BGP rules (i.e., EBGP propagation by an ASBR or iBGP propagation by a Route-Reflector).

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

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

4.3. 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 to steer them towards the next-hop as specified in the SR Policy SAFI NLRI.

5. Contributors

Arjun Sreekantiah
Cisco Systems
US

Email: asreekan@cisco.com

Dhanendra Jain
Cisco Systems
US

Email: dhjain@cisco.com

Acee Lindem
Cisco Systems
US

Email: acee@cisco.com

Siva Sivabalan
Cisco Systems
US

Email: msiva@cisco.com

Imtiyaz Mohammad
Arista Networks
India

Email: imtiyaz@arista.com

6. Acknowledgments

The authors of this document would like to thank Shyam Sethuram and John Scudder for their comments and review of this document.

7. 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:

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

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

8. IANA Considerations

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

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

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

8.1. 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

8.2. 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

8.3. 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

8.4. 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

9. Security Considerations

TBD.

10. References

10.1. 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-06 (work in progress), June 2017.

- [I-D.ietf-pce-segment-routing]
Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W.,
and J. Hardwick, "PCEP Extensions for Segment Routing",
draft-ietf-pce-segment-routing-09 (work in progress),
April 2017.
- [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", 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>>.
- [RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K.
Patel, "Revised Error Handling for BGP UPDATE Messages",
RFC 7606, DOI 10.17487/RFC7606, August 2015,
<<http://www.rfc-editor.org/info/rfc7606>>.

10.2. 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., Raza, K., Leddy, J., Field, B., daniel.voyer@bell.ca, d., daniel.bernier@bell.ca, d., Matsushima, S., Leung, I., Linkova, J., Aries, E., Kosugi, T., Vyncke, E., Lebrun, D., Steinberg, D., and R. Raszk, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-06 (work in progress), March 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-12 (work in progress), June 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.
IT

Email: stefano@previdi.net

Clarence Filsfils
Cisco Systems, Inc.
Brussels
BE

Email: cfilsfil@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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 29, 2017

A. Azimov
E. Bogomazov
Qrator Labs
R. Bush
Internet Initiative Japan
October 26, 2016

Route Leak Detection and Filtering using Roles in Update and Open
messages
draft-ymbk-idr-bgp-open-policy-01

Abstract

Route Leaks are propagation of BGP prefixes which violate assumptions of BGP topology relationships; e.g. passing a route learned from one peer to another peer or to a transit provider, passing a route learned from one transit provider to another transit provider or to a peer. Today, approaches to leak prevention rely on marking routes according to some configuration options without any check of the configuration corresponds to that of the BGP neighbor, or enforcement that the two BGP speakers agree on the relationship. This document enhances BGP Open to establish agreement of the (peer, customer, provider, internal) relationship of two BGP neighboring speakers to enforce appropriate configuration on both sides. Propagated routes are then marked with a eOTC and iOTC attributes according to agreed relationship allowing prevention and detection of route leaks.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in RFC 2119 [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning.

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 April 29, 2017.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. BGP Role	3
3. Role capability	4
4. Role correctness	4
4.1. Strict mode	5
5. Restrictions on the Complex role	5
6. BGP Internal Only To Customer attribute	5
7. BGP External Only To Customer attribute	6
8. Compatibility with BGPsec	6
9. Additional Considerations	6
10. IANA Considerations	7
11. Security Considerations	7
12. References	8
12.1. Normative References	8
12.2. Informative References	8
Authors' Addresses	9

1. Introduction

For the purposes of this document BGP route leaks are when a BGP route was learned from transit provider or peer and is announced to another provider or peer. See [RFC7908]. These are usually the result of misconfigured or absent BGP route filtering or lack of coordination between two BGP speakers.

[I-D.ietf-idr-route-leak-detection-mitigation] describes a method of marking and detecting leaks which relies on operator maintained markings. Unfortunately, in most cases, a leaking router will likely also be misconfigured to mark incorrectly. The proposed mechanism provides an opportunity to detect route leaks made by third parties but provides no support to prevent route leak creation. The leak prevention still relies on communities which are optional and often missed due to mistakes or misunderstanding of the BGP configuration process.

It has been suggested to use white list filtering, relying on knowing the prefixes in the customer cone as import filtering, in order to detect route leaks. Unfortunately, a large number of incidents is created medium size transit operators use a single prefix list as only the ACL for export filtering, without community tagging and paying attention to the source of a learned route. So, if they learn a customer's route from their provider or peer - they will announce it in all directions, including other providers or peers. This misconfiguration affects a limited number of prefixes; but such route leaks will obviously bypass customer cone import filtering made by upper level upstream providers.

Also, route tagging which relies on operator maintained policy configuration is too easily and too often misconfigured.

This document specifies a new BGP Capability Code, [RFC5492] Sec 4, which two BGP speakers MAY use to ensure that they MUST agree on their relationship; i.e. customer and provider or peers. Either or both may optionally be configured to require that this option be exchanged for the BGP Open to succeed.

Also this document specifies a way to mark routes according to BGP Roles and a way to create double-boundary filters for prevention and detection of route leaks via a two new BGP Path Attributes.

2. BGP Role

BGP Role is new mandatory configuration option which must be set per each address family. It reflects the real-world agreement between two BGP speakers about their business relationship.

Allowed Role values are:

- o Provider - sender is a transit provider to neighbor;
- o Customer - sender is customer of neighbor;
- o Peer - sender and neighbor are peers;

- o Internal - sender is part of an internal AS of an organization which has multiple ASs, is a confederation, ...
- o Complex - sender has non-standard agreement and wants to use manual policies.

Since BGP Role reflects the relationship between two BGP speakers, it could also be used for more than route leak mitigation.

3. Role capability

The TLV (type, length, value) of the BGP Role capability are:

- o Type - <TBD1>;
- o Length - 1 (octet);
- o Value - integer corresponding to speaker' BGP Role.

Value	Role name
0	Undefined
1	Sender is Peer
2	Sender is Provider
3	Sender is Customer
4	Sender is Internal
5	Sender is Complex

Table 1: Predefined BGP Role Values

4. Role correctness

Section 2 described how BGP Role is a reflection of the relationship between two BGP speakers. But the mere presence of BGP Role doesn't automatically guarantee role agreement between two BGP peers.

To enforce correctness, use the BGP Role check with a set of constraints on how speakers' BGP Roles MUST corresponded. Of course, each speaker MUST announce and accept the BGP Role capability in the BGP OPEN message exchange.

If a speaker receives a BGP Role capability, it SHOULD check value of the received capability with its own BGP Role. The allowed pairings are (first a sender's Role, second the receiver's Role):

Sender Role	Receiver Role
Peer	Peer
Provider	Customer
Customer	Provider
Internal	Internal
Complex	Complex

Table 2: Allowed Role Capabilities

In all other cases speaker MUST send a Role Mismatch Notification (code 2, sub-code <TBD2>).

4.1. Strict mode

A new BGP configuration option "strict mode" is defined with values of true or false. If set to true, then the speaker MUST refuse to establish a BGP session with peers which do not announce BGP Role capability in their OPEN message. If a speaker rejects a connection, it MUST send a Connection Rejected Notification [RFC4486] (Notification with error code 6, subcode 5). By default strict mode SHOULD be set to false for backward compatibility with BGP speakers, that do not yet support this mechanism.

5. Restrictions on the Complex role

Complex role should be set only if relations between BGP neighbors could not be described using simple Customer/Provider/Peer roles. For a example, if neighbor is literal peer, but for some prefixes it provides full transit, complex role SHOULD be set on both sides. In this case configuration of detection and filtering mechanisms (Section 6 and Section 7) should be set on per-prefix basis upon local policy.

6. BGP Internal Only To Customer attribute

The Internal Only To Customer (iOTC) attribute is a new optional, non-transitive BGP Path attribute with the Type Code <TBD3>. This attribute has zero length as it used only as a flag.

There are two rules for setting the iOTC attribute:

1. The iOTC attribute MUST be added to all incoming routes if the receiver's Role is Customer or Peer;

2. Routes with the iOTC attribute set MUST NOT be announced if the sender's Role is Customer or Peer;

These two rules provide mechanism that prevent route leak creation by an AS. In case of Complex role usage the way of iOTC process is not automated and upon local policy.

7. BGP External Only To Customer attribute

The External Only To Customer (eOTC) attribute is a new optional, transitive BGP Path attribute with the Type Code <TBD4>. This attribute has four bytes length and contain an AS number of AS, that added attribute to the route.

There are two rules for setting the eOTC attribute:

1. If eOTC is not set and sender's Role is Provider or Peer the eOTC attribute MUST be added with value equal to its ASN.
2. If eOTC is set, receiver's Role is Provider or Peer, and its value is not equal to neighbor ASN then such incoming route is route leak and MUST be given a lower local preference, or they MAY be dropped.

These two rules provide mechanism for route leak detection that is made by some party in ASPath. In case of Complex role usage the way of eOTC process is not automated and upon local policy.

8. Compatibility with BGPsec

In BGPsec [I-D.ietf-sidr-bgpsec-protocol] enabled routers eOTC attribute MUST be turned into one bit of Flags field of Secure_Path Segment and MUST NOT be added as separate attribute.

When route is transmitted from BGPsec enabled router to BGPsec disabled device, in addition to AS_PATH reconstruction MUST be performed eOTC attribute reconstruction. If corresponded bit was set in one of Secure_Path Segments, eOTC attribute SHOULD be added with value that equals to ASN in which segment it appears for the first time.

9. Additional Considerations

As BGP Role reflects the relationship between neighbors, it can also have other uses. As an example, BGP Role might affect route priority, or be used to distinguish borders of a network if a network consists of multiple AS.

Though such uses may be worthwhile, they are not the goal of this document. Note that such uses would require local policy control.

This document doesn't provide any security measures to check correctness of attributes usage in case of Complex role, so Complex role should be set with great caution.

10. IANA Considerations

This document defines a new Capability Codes option [to be removed upon publication: <http://www.iana.org/assignments/capability-codes/capability-codes.xhtml>] [RFC5492], named "BGP Role", assigned value <TBD1> . The length of this capability is 1.

The BGP Role capability includes a Value field, for which IANA is requested to create and maintain a new sub-registry called "BGP Role Value". Assignments consist of Value and corresponding Role name. Initially this registry is to be populated with the data in Table 1. Future assignments may be made by a standard action procedure [RFC5226].

This document defines new subcode, "Role Mismatch", assigned value <TBD2> in the OPEN Message Error subcodes registry [to be removed upon publication: <http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-6>] [RFC4271].

This document defines a new optional, non-transitive BGP Path Attributes option, named "Internal Only To Customer", assigned value <TBD3> [To be removed upon publication: <http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-2>] [RFC4271]. The length of this attribute is 0.

This document defines a new optional, transitive BGP Path Attributes option, named "External Only To Customer", assigned value <TBD4> [To be removed upon publication: <http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-2>] [RFC4271]. The length of this attribute is 4.

11. Security Considerations

This document proposes a mechanism for prevention and detection of route leaks, that are the result of BGP policy misconfiguration. That includes preventing route leaks created inside an AS (company), and route leak detection, if a route was leaked by third party.

Deliberate sending of a known conflicting BGP Role could be used to sabotage a BGP connection. This is easily detectable.

Deliberate mis-marking of the eOTC flag could be used to could affect BGP decision process but could not sabotage a route's propagation.

BGP Role is disclosed only to an immediate BGP speaker, so it will not itself reveal any sensitive information to third parties.

On the other hand, eOTC is a transitive BGP AS_PATH attribute which reveals a bit about a BGP speaker's business relationship. It will give a strong hint that some link isn't customer to provider, but will not help to distinguish if it is provider to customer or peer to peer. If eOTC is BGPsec signed, it can not be removed for business confidentiality.

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>>.
- [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>>.
- [RFC4486] Chen, E. and V. Gillet, "Subcodes for BGP Cease Notification Message", RFC 4486, DOI 10.17487/RFC4486, April 2006, <<http://www.rfc-editor.org/info/rfc4486>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.

12.2. Informative References

- [I-D.ietf-idr-route-leak-detection-mitigation] Sriram, K., Montgomery, D., Dickson, B., Patel, K., and A. Robachevsky, "Methods for Detection and Mitigation of BGP Route Leaks", draft-ietf-idr-route-leak-detection-mitigation-04 (work in progress), July 2016.
- [I-D.ietf-sidr-bgpsec-protocol] Lepinski, M. and K. Sriram, "BGPsec Protocol Specification", draft-ietf-sidr-bgpsec-protocol-18 (work in progress), August 2016.

- [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>>.
- [RFC7908] Sriram, K., Montgomery, D., McPherson, D., Osterweil, E., and B. Dickson, "Problem Definition and Classification of BGP Route Leaks", RFC 7908, DOI 10.17487/RFC7908, June 2016, <<http://www.rfc-editor.org/info/rfc7908>>.

Authors' Addresses

Alexander Azimov
Qrator Labs

Email: aa@qrator.net

Eugene Bogomazov
Qrator Labs

Email: eb@qrator.net

Randy Bush
Internet Initiative Japan

Email: randy@psg.com