Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: September 25, 2019

G. Dawra, Ed.
LinkedIn
C. Filsfils
K. Talaulikar, Ed.
Cisco Systems
M. Chen
Huawei
D. Bernier
Bell Canada
J. Uttaro
AT&T
B. Decraene
Orange
H. Elmalky
Ericsson

March 24, 2019

BGP Link State Extensions for SRv6
draft-dawra-idr-bgpls-srv6-ext-06

Abstract

Segment Routing IPv6 (SRv6) allows for a flexible definition of end-to-end paths within various topologies by encoding paths as sequences of topological or functional sub-paths, called "segments". These segments are advertised by the various protocols such as BGP, ISIS and OSPFv3.

BGP Link-state (BGP-LS) address-family solution for SRv6 is similar to BGP-LS for SR for MPLS dataplane. This draft defines extensions to the BGP-LS to advertise SRv6 Segments along with their functions and other attributes 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.

Copyright Notice

Table of Contents

1. Introduction

SRv6 refers to Segment Routing instantiated on the IPv6 dataplane [RFC8402]. Segment Identifier (SID) is often used as a shorter reference for "SRv6 Segment".

The network programming paradigm [I-D.filsfils-spring-srv6-network-programming] is central to SRv6. It describes how different functions can be bound to their SIDs and how a network program can be expressed as a combination of SIDs.

An SRv6-capable node N maintains a "My SID Table" (refer [I-D.filsfils-spring-srv6-network-programming]). This table contains all the SRv6 segments explicitly instantiated at node N.

The IS-IS [I-D.bashandy-isis-srv6-extensions] and OSPFv3 [I-D.li-ospf-ospfv3-srv6-extensions] link-state routing protocols have been extended to advertise some of these SRv6 SIDs and SRv6-related information. BGP ([I-D.dawra-idr-srv6-vpn]) has been extended to advertise some of these SRv6 SIDs for VPN services. Certain other SRv6 SIDs may be instantiated on a node via other mechanisms for topological or service functionalities.

The advertisement of SR related information along with the topology for the MPLS dataplane instantiation is specified in [I-D.ietf-idr-bgp-ls-segment-routing-ext] and for the BGP Egress Peer Engineering (EPE) is specified in [I-D.ietf-idr-bgpls-segment-routing-epe]. On the similar lines, introducing the SRv6 related information in BGP-LS allows it's consumer applications that require topological visibility to also receive the "My SID Table" from nodes across a domain or even across Autonomous Systems (AS), as required. This allows applications to leverage the SRv6 capabilities for network programming.

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 [RFC7752].

This document describes extensions to BGP-LS to advertise the SRv6 "My SID Table" and other SRv6 information from all the SRv6 capable nodes in the domain when sourced from link-state routing protocols and directly from individual SRv6 capable nodes when sourced from BGP.

2. BGP-LS Extensions for SRv6

BGP-LS[RFC7752] defines the BGP Node, Link and Prefix attributes. All non-VPN link, node, and prefix information SHALL be encoded using AFI 16388 / SAFI 71. VPN link, node, and prefix information SHALL be encoded using AFI 16388 / SAFI 72.

The SRv6 information pertaining to a node is advertised via the BGP-LS Node NLRI and using the BGP-LS Attribute TLVs as follows:

- SRv6 Capabilities of the node is advertised via a new SRv6 Capabilities TLV

- New MSD types introduced for SRv6 are advertised as new sub-TLVs of the Node MSD TLV specified in [I-D.ietf-idr-bgp-ls-segment-routing-msd].

- Algorithm support for SRv6 is advertised via the existing SR Algorithm TLV specified in [I-D.ietf-idr-bgp-ls-segment-routing-ext].

The SRv6 information pertaining to a link is advertised via the BGP-LS Link NLRI and using the BGP-LS Attribute TLVs as follows:

- SRv6 End.X SID of the link state routing adjacency or the BGP EPE Peer Adjacency is advertised via a new SRv6 End.X SID TLV

- SRv6 LAN End.X SID of the link state routing adjacency to a non-DR/DIS router is advertised via a new SRv6 LAN End.X SID TLV

- New MSD types introduced for SRv6 are advertised as new sub-TLVs of the Link MSD TLV specified in [I-D.ietf-idr-bgp-ls-segment-routing-msd].

The SRv6 Locator information of a node is advertised via the BGP-LS Prefix NLRI using the new SRv6 Locator TLV in the BGP-LS Attribute.

The SRv6 SIDs associated with the node from its "My SID Table" are advertised as a newly introduce BGP-LS SRv6 SID NLRI. This enables the BGP-LS encoding to scale to cover a potentially large set of SRv6 SIDs instantiated on a node with the granularity of individual SIDs and without affecting the size and scalability of the BGP-LS updates.

New BGP-LS Attribute TLVs are introduced for the SRv6 SID NLRI as follows:

- The endpoint function of the SRv6 SID is advertised via a new SRv6 Endpoint Function TLV

- The BGP EPE Peer Node and Peer Set SID context is advertised via a new SRv6 BGP EPE Peer Node SID TLV

When the BGP-LS router is advertising topology information that it sources from the underlying link-state routing protocol, then it maps the corresponding SRv6 information from the SRv6 extensions for IS-IS [I-D.bashandy-isis-srv6-extensions] and OSPFv3 [I-D.li-ospf-ospfv3-srv6-extensions] protocols to their BGP-LS TLVs/ sub-TLVs for all SRv6 capable nodes in that routing protocol domain. When the BGP-LS router is advertising topology information from the BGP routing protocol [I-D.ietf-idr-bgpls-segment-routing-epe], then it advertises the SRv6 information from the local node alone (e.g. BGP EPE topology information or in the case of a data center network running BGP as the only routing protocol).

Subsequent sections of this document specify the encoding of the newly defined extensions.

3. SRv6 Node Attributes

SRv6 attributes of a node are advertised using the new BGP-LS Attribute TLVs defined in this section and associated with the BGP-LS Node NLRI.

3.1. SRv6 Capabilities TLV

This BGP-LS Attribute TLV is used to announce the SRv6 capabilities of the node along with the BGP-LS Node NLRI and indicates the SRv6 support by the node. A single instance of this TLV MUST be included in the BGP-LS attribute for each SRv6 capable node. This TLV maps to the SRv6 Capabilities sub-TLV and the SRv6 Capabilities TLV of the IS-IS and OSPFv3 protocol SRv6 extensions respectively.

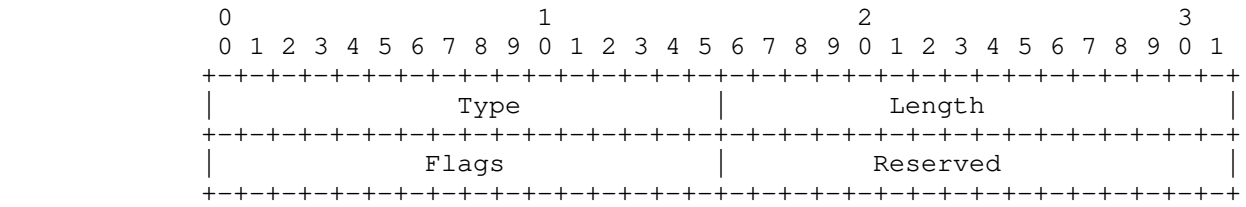


Figure 1: SRv6 Capabilities TLV Format

Where:

- Type: 2 octet field with value TBD, see Section 8.
- Length : 2 octet field with value set to 4.
- Flags: 2 octet field. The following flags are defined:

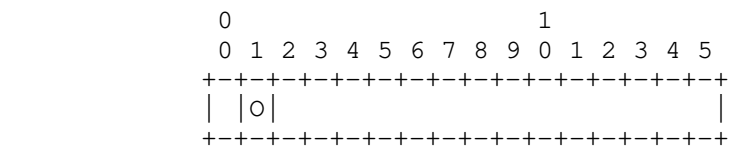


Figure 2: SRv6 Capability TLV Flags Format

  * O-flag: If set, then router is capable of supporting SRH O-bit Flags, as specified in [I-D.ali-spring-srv6-oam].

- Reserved: 2 octet that SHOULD be set to 0 and MUST be ignored on receipt.

3.2. SRv6 Node MSD Types

The Node MSD TLV [I-D.ietf-idr-bgp-ls-segment-routing-msd] of the BGP-LS Attribute of the Node NLRI is also used to advertise the limits and the supported Segment Routing Header (SRH) [I-D.ietf-6man-segment-routing-header] operations supported by the SRv6 capable node. The SRv6 MSD Types specified in [I-D.bashandy-isis-srv6-extensions] are also used with the BGP-LS Node MSD TLV as these codepoints are shared between IS-IS, OSPF and BGP-LS protocols. The description and semantics of these new MSD types for BGP-LS are identical as specified [I-D.bashandy-isis-srv6-extensions] and summarized in the table below:

MSD Type	Description
TBD	Maximum Segments Left
TBD	Maximum End Pop
TBD	Maximum T.Insert
TBD	Maximum T.Encaps
TBD	Maximum End D

Figure 3: SRv6 Node MSD Types

Each MSD type is encoded as a one octet type followed by a one octet value.

4. SRv6 Link Attributes

SRv6 attributes and SIDs associated with a link or adjacency are advertised using the new BGP-LS Attribute TLVs defined in this section and associated with the BGP-LS Link NLRI.

4.1. SRv6 End.X SID TLV

The SRv6 End.X SID TLV is used to advertise the SRv6 End.X SIDs that correspond to a point-to-point or point-to-multipoint link or adjacency of the local node for IS-IS and OSPFv3 protocols. This TLV can also be used to advertise the End.X function SRv6 SID corresponding to the underlying layer-2 member links for a layer-3 bundle interface using L2 Bundle Member Attribute TLV as specified in [I-D.ietf-idr-bgp-ls-segment-routing-ext] .

For the nodes running BGP routing protocol, this TLV is used to advertise the BGP EPE Peer Adjacency SID for SRv6 on the same lines as specified for SR/MPLS in [I-D.ietf-idr-bgpls-segment-routing-epe]. The SRv6 End.X SID for the BGP Peer Adjacency indicates the cross-connect to a specific layer-3 link to the specific BGP session peer (neighbor).

The TLV has the following format:

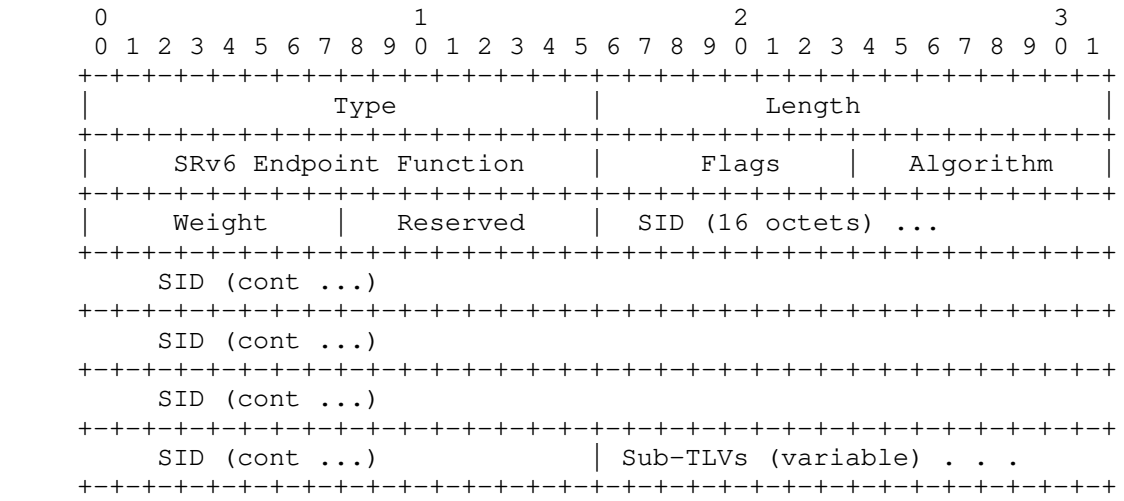


Figure 4: SRv6 End.X TLV Format

Where:

Type: 2 octet field with value TBD, see Section 8.

Length: 2 octet field with the total length of the value portion of the TLV.

Function Code: 2 octet field. The Endpoint Function code point for this SRv6 SID as defined in [I-D.filsfils-spring-srv6-network-programming].

Flags: 1 octet of flags with the following definition:

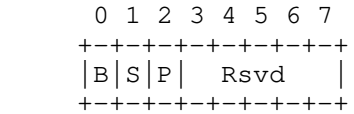


Figure 5: SRv6 End.X SID TLV Flags Format

* B-Flag: Backup Flag. If set, the SID is eligible for protection (e.g. using IPFRR) as described in [RFC8355].

* S-Flag: Set Flag. When set, the S-Flag indicates that the SID refers to a set of adjacencies (and therefore MAY be assigned to other adjacencies as well).

* P-Flag: Persistent Flag: When set, the P-Flag indicates that the SID is persistently allocated, i.e., the value remains consistent across router restart and/or interface flap.

* Rsvd bits: Reserved for future use and MUST be zero when originated and ignored when received.

Algorithm: 1 octet field. Algorithm associated with the SID. Algorithm values are defined in the IGP Algorithm Type registry.

Weight: 1 octet field. The value represents the weight of the SID for the purpose of load balancing. The use of the weight is defined in [RFC8402].

Reserved: 1 octet field that SHOULD be set to 0 and MUST be ignored on receipt.

SID: 16 octet field. This field encodes the advertised SRv6 SID as 128 bit value.

Sub-TLVs : currently none defined. Used to advertise sub-TLVs that provide additional attributes for the given SRv6 End.X SID.

4.2. SRv6 LAN End.X SID TLV

For a LAN interface, normally a node only announces its adjacency to the IS-IS pseudo-node (or the equivalent OSPF Designated Router). The SRv6 LAN End.X SID TLV allows a node to announce SRv6 SID corresponding to functions like END.X for its adjacencies to all other (i.e. non-DIS or non-DR) 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 SRv6 End.X SID TLVs for these neighbor adjacencies.

The IS-IS and OSPFv3 SRv6 LAN End.X SID TLVs have the following format:

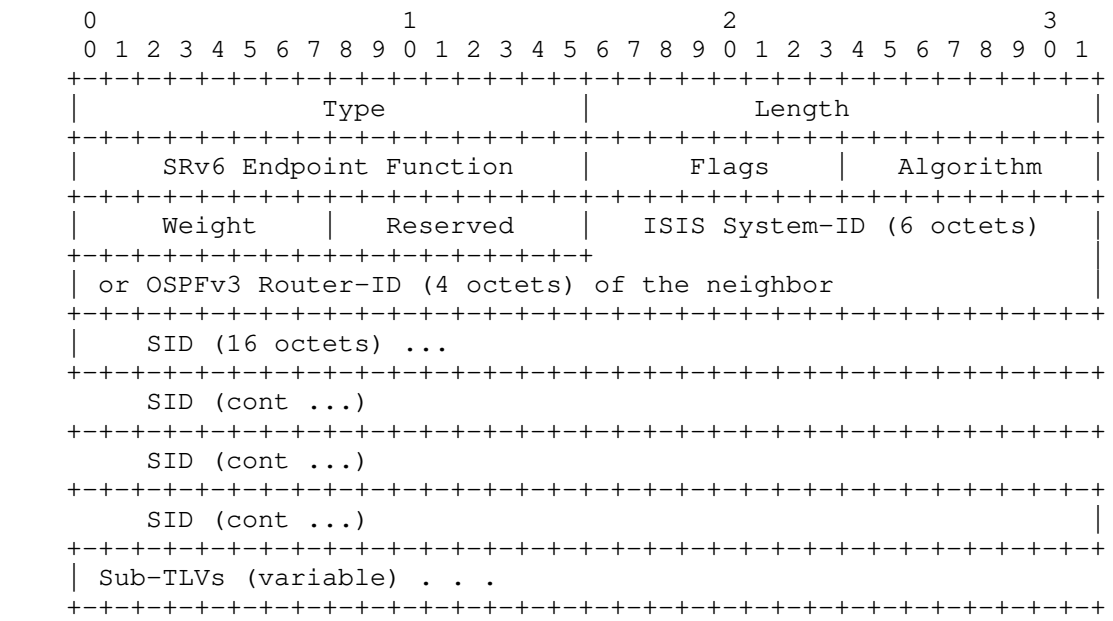


Figure 6: SRv6 LAN End.X SID TLV Format

Where:

- Type: 2 octet field with value TBD in case of IS-IS and TBD in case of OSPFv3, see Section 8.
- Length: 2 octet field with the total length of the value portion of the TLV.
- Function Code: 2 octet field. The Endpoint Function code point for this SRv6 SID as defined in [I-D.filsfils-spring-srv6-network-programming].
- Flags: 1 octet of flags with the following definition:

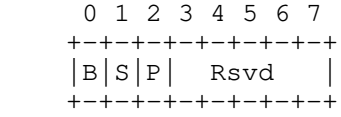


Figure 7: SRv6 LAN End.X SID TLV Flags Format

  * B-Flag: Backup Flag. If set, the SID is eligible for protection (e.g. using IPFRR) as described in [RFC8355].

  * S-Flag: Set Flag. When set, the S-Flag indicates that the SID refers to a set of adjacencies (and therefore MAY be assigned to other adjacencies as well).

  * P-Flag: Persistent Flag: When set, the P-Flag indicates that the SID is persistently allocated, i.e., the value remains consistent across router restart and/or interface flap.

  * Rsvd bits: Reserved for future use and MUST be zero when originated and ignored when received.

- Algorithm: 1 octet field. Algorithm associated with the SID. Algorithm values are defined in the IGP Algorithm Type registry.

- Weight: 1 octet field. The value represents the weight of the SID for the purpose of load balancing. The use of the weight is defined in [RFC8402].

- Reserved: 1 octet field that SHOULD be set to 0 and MUST be ignored on receipt.

- Neighbor ID : 6 octets of ISIS System ID of the neighbor for the ISIS SRv6 LAN End.X SID TLV and 4 octets of OSPFv3 Router-id of the neighbor for the OSPFv3 SRv6 LAN End.X SID TLV.

- SID: 16 octet field. This field encodes the advertised SRv6 SID as 128 bit value.

- Sub-TLVs : currently none defined. Used to advertise sub-TLVs that provide additional attributes for the given SRv6 LAN End.X SID.

4.3. SRv6 Link MSD Types

The Link MSD TLV [I-D.ietf-idr-bgp-ls-segment-routing-msd] of the BGP-LS Attribute of the Link NLRI is also used to advertise the limits and the supported Segment Routing Header (SRH) operations supported on the specific link by the SRv6 capable node. The SRv6 MSD Types specified in [I-D.bashandy-isis-srv6-extensions] are also used with the BGP-LS Link MSD TLV as these codepoints are shared between IS-IS, OSPF and BGP-LS protocols. The description and semantics of these new MSD types for BGP-LS are identical as specified [I-D.bashandy-isis-srv6-extensions] and summarized in the table below:

MSD Type	Description
TBD	Maximum Segments Left
TBD	Maximum End Pop
TBD	Maximum T.Insert
TBD	Maximum T.Encaps
TBD	Maximum End D

Figure 8: SRv6 Link MSD Types

Each MSD type is encoded as a one octet type followed by a one octet value.

5. SRv6 Prefix Attributes

SRv6 attributes with an IPv6 prefix are advertised using the new BGP-LS Attribute TLVs defined in this section and associated with the BGP-LS Prefix NLRI.

5.1. SRv6 Locator TLV

As described in [I-D.filsfils-spring-srv6-network-programming], an SRv6 SID is 128 bits and represented as

LOC:FUNCT

where LOC (the locator portion) is the L most significant bits and FUNCT is the 128-L least significant bits. L is called the locator length and is flexible. A node is provisioned with one or more locators supported by that node. Locators are covering prefixes for the set of SIDs provisioned on that node. These Locators are advertised as BGP-LS Prefix NLRI objects along with the SRv6 Locator TLV in its BGP-LS Attribute.

The IPv6 Prefix matching the Locator MAY be also advertised as a prefix reachability by the underlying routing protocol. In this case, the Prefix NLRI would be also associated with the Prefix Metric TLV that carries the routing metric for this prefix. When the Locator prefix is not being advertised as a prefix reachability, then the Prefix NLRI would have the SRv6 Locator TLV associated with it but no Prefix Metric TLV. In the absence of Prefix Metric TLV, the consumer of the BGP-LS topology information MUST NOT interpret the Locator prefix as a prefix reachability routing advertisement.

The SRv6 Locator TLV has the following format:

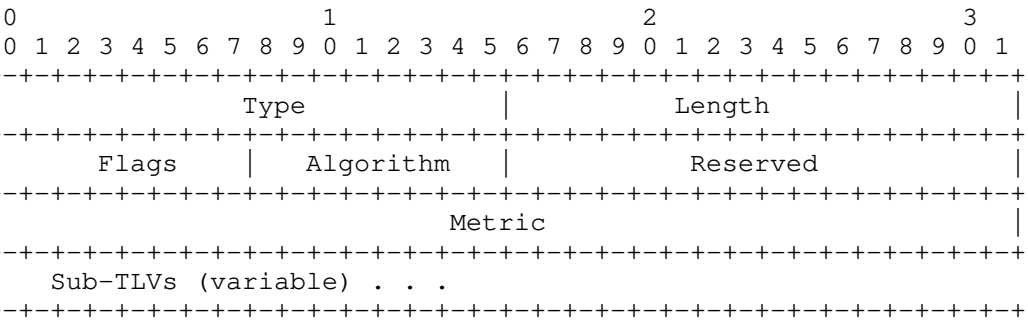


Figure 9: SRv6 Locator TLV Format

Where:

Type: 2 octet field with value TBD, see Section 8.

Length: 2 octet field with the total length of the value portion of the TLV.

Flags: 1 octet of flags with the following definition:

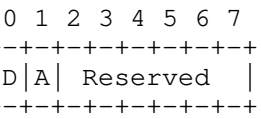


Figure 10: SRv6 Locator TLV Flags Format

* D-Flag: Indicates that the locator has been leaked into the IGP domain when set. IS-IS operations for this are discussed in [I-D.bashandy-isis-srv6-extensions].

* A-Flag: When the Locator is associated with anycast destinations, the A flag SHOULD be set. Otherwise, this bit MUST be clear.

* Reserved bits: Reserved for future use and MUST be zero when originated and ignored when received.

Algorithm: 1 octet field. Algorithm associated with the SID. Algorithm values are defined in the IGP Algorithm Type registry.

Reserved: 2 octet field. The value MUST be zero when originated and ignored when received.

Metric: 4 octet field. The value of the metric for the Locator.

Sub-TLVs : currently none defined. Used to advertise sub-TLVs that provide additional attributes for the given SRv6 Locator.

6. SRv6 SID NLRI

SRv6 SID information is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes [RFC4760]. The "Link-State NLRI" defined in [RFC7752] is extended to carry the SRv6 SID information.

A new "Link-State NLRI Type" is defined for SRv6 SID information as following:

- Link-State NLRI Type: SRv6 SID NLRI (value TBD see IANA Considerations Section 8.1).

The format of this new NLRI type is as shown in the following figure:

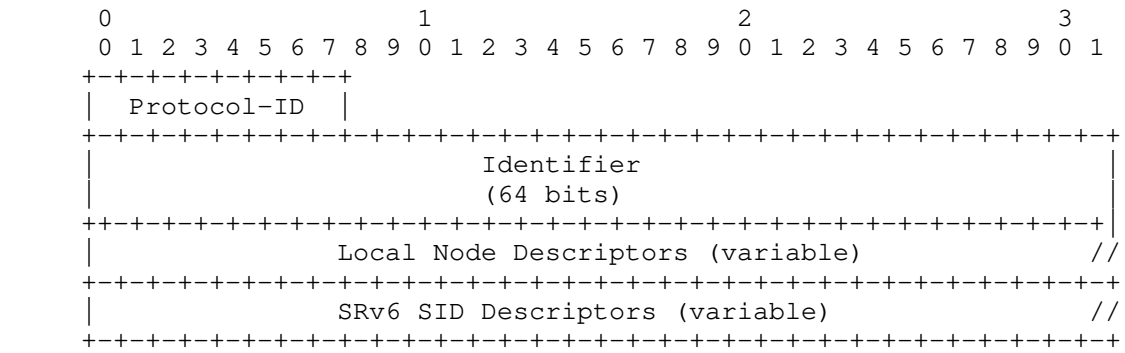


Figure 11: SRv6 SID NLRI Format

Where:

- Protocol-ID: 1 octet field that specifies the protocol component through which BGP-LS learns the SRv6 SIDs of the node. The following Protocol-IDs apply to the SRv6 SID NLRI:

Protocol-ID	NLRI information source protocol
1	IS-IS Level 1
2	IS-IS Level 2
4	Direct
5	Static configuration
6	OSPFv3
7	BGP

Figure 12: Protocol IDs for SRv6 SID NLRI

- Identifier: 8 octet value as defined in [RFC7752].

- Local Node Descriptors TLV: as defined in [RFC7752] for IGPs, local and static configuration and as defined in [I-D.ietf-idr-bgpls-segment-routing-epe] for BGP protocol.

- SRv6 SID Descriptors: MUST include the SRv6 SID Information TLV defined in Section 6.1 and optionally MAY include the Multi-Topology Identifier TLV as defined in [RFC7752].

New TLVs carried in the BGP Link State Attribute defined in [RFC7752] are also defined in order to carry the attributes of a SRv6 SID in Section 7.

6.1. SRv6 SID Information TLV

A SRv6 SID is a 128 bit value [I-D.filsfils-spring-srv6-network-programming] and is encoded using the SRv6 SID Information TLV.

The TLV has the following format:

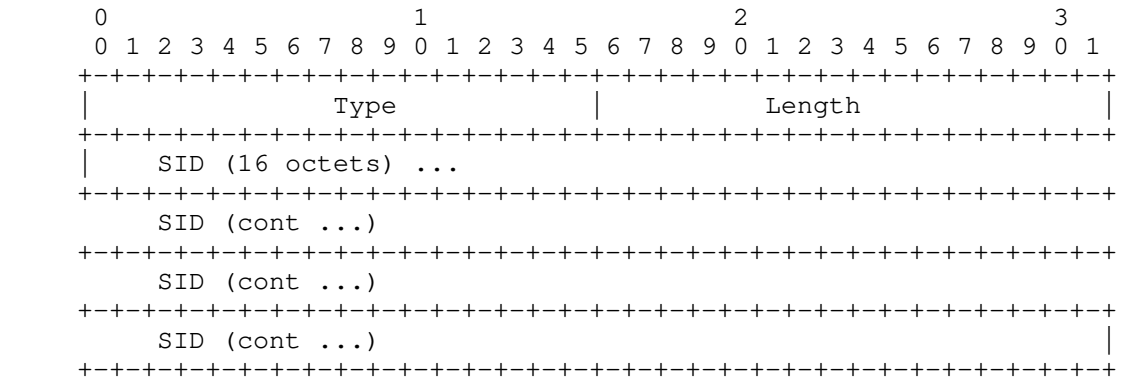


Figure 13: SRv6 SID Information TLV Format

Where:

Type: 2 octet field with value TBD, see Section 8.

Length: 2 octet field with value set to 16.

SID: 16 octet field. This field encodes the advertised SRv6 SID as 128 bit value.

7. SRv6 SID Attributes

This section specifies the new TLVs to be carried in the BGP Link State Attribute associated with the BGP-LS SRv6 SID NLRI.

7.1. SRv6 Endpoint Function TLV

Each SRv6 SID instantiated in the "My SID Table" of an SRv6 capable node has a specific instruction bound to it. A set of well-known functions that can be associated with a SID are defined in [I-D.filsfils-spring-srv6-network-programming].

The SRv6 Endpoint Function TLV is a mandatory TLV that MUST be included in the BGP-LS Attribute associated with the BGP-LS SRv6 SID NLRI. The TLV has the following format:

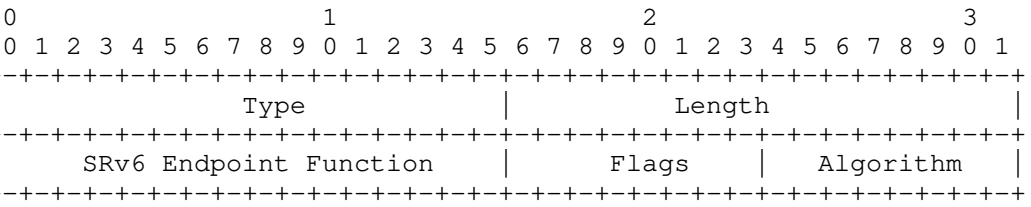


Figure 14: SRv6 Endpoint Function TLV

Where:

Type: 2 octet field with value TBD, see Section 8.

Length: 2 octet field with the value 4.

Function Code: 2 octet field. The Endpoint Function code point for this SRv6 SID as defined in [I-D.filsfils-spring-srv6-network-programming].

Flags: 1 octet of flags with the none defined currently. Reserved for future use and MUST be zero when originated and ignored when received.

Algorithm: 1 octet field. Algorithm associated with the SID. Algorithm values are defined in the IGP Algorithm Type registry.

7.2. SRv6 BGP Peer Node SID TLV

The BGP Peer Node SID and Peer Set SID for SR with MPLS dataplane are specified in [I-D.ietf-idr-bgpls-segment-routing-epe]. The similar Peer Node and Peer Set SID functionality can be realized with SRv6 using the END.X SRv6 SID. The SRv6 BGP Peer Node SID TLV is an optional TLV for use in the BGP-LS Attribute for an SRv6 SID NLRI corresponding to BGP protocol. This TLV MUST be included along with SRv6 End.X SID that is associated with the BGP Peer Node or Peer Set functionality.

The TLV has the following format:

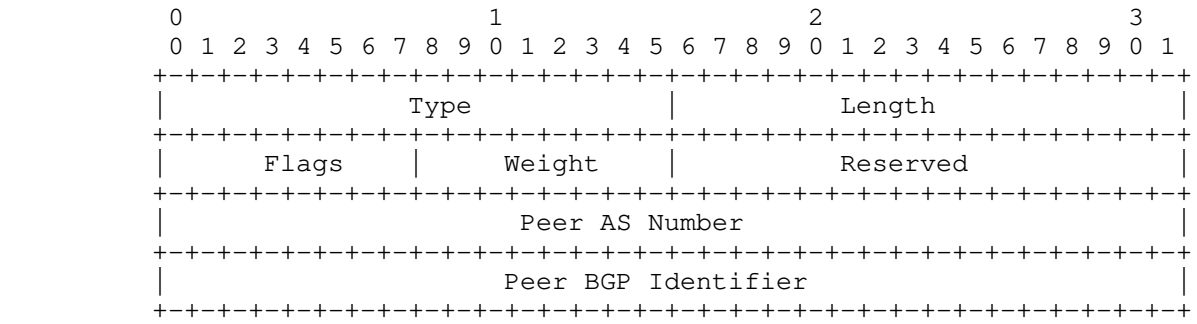


Figure 15: SRv6 BGP Peer Node SID TLV Format

Where:

- Type: 2 octet field with value TBD, see Section 8.
- Length: 2 octet field with the value 12.
- Flags: 1 octet of flags with the following definition:

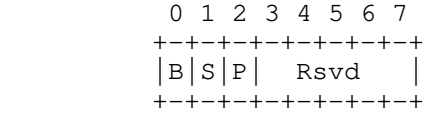


Figure 16: SRv6 BGP Peer End.X SID TLV Flags Format

  * B-Flag: Backup Flag. If set, the SID is eligible for protection (e.g. using IPFRR) as described in [RFC8355].
  * S-Flag: Set Flag. When set, the S-Flag indicates that the SID refers to a set of BGP peering sessions (i.e. BGP Peer Set SID functionality) and therefore MAY be assigned to one or more End.X SIDs associated with BGP peer sessions.
  * P-Flag: Persistent Flag: When set, the P-Flag indicates that the SID is persistently allocated, i.e., the value remains consistent across router restart and/or session flap.
  * Rsvd bits: Reserved for future use and MUST be zero when originated and ignored when received.

- o Weight: 1 octet field. The value represents the weight of the SID for the purpose of load balancing. The use of the weight is defined in [RFC8402].

- o Peer AS Number : 4 octets of BGP AS number of the peer router.

- o Peer BGP Identifier : 4 octets of the BGP Identifier (BGP Router-ID) of the peer router.

For a SRv6 BGP EPE Peer Node SID, one instance of this TLV is associated with the SRv6 SID. For SRv6 BGP EPE Peer Set SID, multiple instances of this TLV (one for each peer in the "peer set") are associated with the SRv6 SID and the S (set/group) flag is SET.

8. IANA Considerations

This document requests assigning code-points from the IANA "Border Gateway Protocol - Link State (BGP-LS) Parameters" registry as described in the sub-sections below.

8.1. BGP-LS NLRI-Types

The following codepoints is suggested (to be assigned by IANA) from within the sub-registry called "BGP-LS NLRI-Types":

Type	NLRI Type	Reference
6	SRv6 SID	this document

Figure 17: SRv6 SID NLRI Type Codepoint

8.2. BGP-LS TLVs

The following TLV codepoints are suggested (to be assigned by IANA) from within the sub-registry called "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs":

TLV Code Point	Description	Value defined in
TBD	SRv6 Capabilities TLV	this document
TBD	SRv6 End.X SID TLV	this document
TBD	IS-IS SRv6 LAN End.X SID TLV	this document
TBD	OSPFv3 SRv6 LAN End.X SID TLV	this document
TBD	SRv6 Locator TLV	this document
TBD	SRv6 SID Information TLV	this document
TBD	SRv6 Endpoint Function TLV	this document
TBD	SRv6 BGP Peer Node SID TLV	this document

Figure 18: SRv6 BGP-LS Attribute TLV Codepoints

9. Manageability Considerations

This section is structured as recommended in[RFC5706]

10. Operational Considerations

10.1. Operations

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

11. 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.

12. Contributors

Arjun Sreekantiah
Individual
US

Les Ginsberg
Cisco Systems
US
Email: ginsberg@cisco.com

Shunwan Zhuang
Huawei
China
Email: zhuangshunwan@huawei.com

13. Acknowledgements

The authors would like to thank Peter Psenak and Arun Babu for their review of this document and their comments.

14. References

14.1. Normative References

[I-D.ali-spring-srv6-oam]
Ali, Z., Filsfils, C., Kumar, N., Pignataro, C., faiqbal@cisco.com, f., Gandhi, R., Leddy, J., Matsushima, S., Raszuk, R., daniel.voyer@bell.ca, d., Dawra, G., Peirens, B., Chen, M., and G. Naik, "Operations, Administration, and Maintenance (OAM) in Segment Routing Networks with IPv6 Data plane (SRv6)", draft-ali-spring-srv6-oam-02 (work in progress), October 2018.

[I-D.bashandy-isis-srv6-extensions]
Psenak, P., Filsfils, C., Bashandy, A., Decraene, B., and Z. Hu, "IS-IS Extensions to Support Routing over IPv6 Dataplane", draft-bashandy-isis-srv6-extensions-05 (work in progress), March 2019.

[I-D.dawra-idr-srv6-vpn]
Dawra, G., Filsfils, C., Dukes, D., Brissette, P., Camarillo, P., Leddy, J., daniel.voyer@bell.ca, d., daniel.bernier@bell.ca, d., Steinberg, D., Raszuk, R., Decraene, B., Matsushima, S., and S. Zhuang, "BGP Signaling for SRv6 based Services.", draft-dawra-idr-srv6-vpn-05 (work in progress), October 2018.

[I-D.filsfils-spring-srv6-network-programming]
Filsfils, C., Camarillo, P., Leddy, J., daniel.voyer@bell.ca, d., Matsushima, S., and Z. Li, "SRv6 Network Programming", draft-filsfils-spring-srv6-network-programming-07 (work in progress), February 2019.

[I-D.ietf-6man-segment-routing-header]
Filsfils, C., Previdi, S., Leddy, J., Matsushima, S., and d. daniel.voyer@bell.ca, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-16 (work in progress), February 2019.

[I-D.ietf-idr-bgp-ls-segment-routing-ext]
Previdi, S., Talaulikar, K., Filsfils, C., Gredler, H., and M. Chen, "BGP Link-State extensions for Segment Routing", draft-ietf-idr-bgp-ls-segment-routing-ext-12 (work in progress), March 2019.

[I-D.ietf-idr-bgp-ls-segment-routing-msd]
Tantsura, J., Chunduri, U., Mirsky, G., Sivabalan, S., and N. Triantafillis, "Signaling MSD (Maximum SID Depth) using Border Gateway Protocol Link-State", draft-ietf-idr-bgp-ls-segment-routing-msd-04 (work in progress), February 2019.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Talaulikar, K., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", draft-ietf-idr-bgpls-segment-routing-epe-17 (work in progress), October 2018.

[I-D.li-ospf-ospfv3-srv6-extensions]
Li, Z., Hu, Z., Cheng, D., Talaulikar, K., and P. Psenak, "OSPFv3 Extensions for SRv6", draft-li-ospf-ospfv3-srv6-extensions-03 (work in progress), March 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>.

[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>.

[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>.

[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>.

14.2. Informative References

[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>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[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>.

[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>.

[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, <https://www.rfc-editor.org/info/rfc6952>.

[RFC8355] Filsfils, C., Ed., Previdi, S., Ed., Decraene, B., and R. Shakir, "Resiliency Use Cases in Source Packet Routing in Networking (SPRING) Networks", RFC 8355, DOI 10.17487/RFC8355, March 2018, <https://www.rfc-editor.org/info/rfc8355>.

Authors' Addresses

Gaurav Dawra (editor)
LinkedIn
USA

Email: gdawra.ietf@gmail.com

Clarence Filsfils
Cisco Systems
Belgium

Email: cfilsfil@cisco.com

Ketan Talaulikar (editor)
Cisco Systems
India

Email: ketant@cisco.com

Mach Chen
Huawei
China

Email: mach.chen@huawei.com

Daniel Bernier
Bell Canada
Canada

Email: daniel.bernier@bell.ca

Jim Uttaro
AT&T
USA

Email: ju1738@att.com

Bruno Decraene
Orange
France

Email: bruno.decraene@orange.com

Hani Elmalky
Ericsson
USA

Email: hani.elmalky@gmail.com

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2019

G. Dawra, Ed.
LinkedIn
C. Filsfils
D. Dukes
P. Brissette
P. Camarilo
Cisco Systems
J. Leddy
Comcast
D. Voyer
D. Bernier
Bell Canada
D. Steinberg
Steinberg Consulting
R. Raszuk
Bloomberg LP
B. Decraene
Orange
S. Matsushima
SoftBank
S. Zhuang
Huawei Technologies
October 22, 2018

BGP Signaling for SRv6 based Services.
draft-dawra-idr-srv6-vpn-05

Abstract

This draft defines procedures and messages for BGP SRv6-based L3VPN and EVPN. It builds on RFC4364 "BGP/MPLS IP Virtual Private Networks (VPNs)" and RFC7432 "BGP MPLS-Based Ethernet VPN" and provides a migration path from MPLS-based VPNs to SRv6 based VPNs.

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/.

Copyright Notice

Table of Contents

1. Introduction

SRv6 refers to Segment Routing instantiated on the IPv6 dataplane [I-D.filsfils-spring-srv6-network-programming][I-D.ietf-6man-segment-routing-header].

SRv6 based BGP services refers to the L3 and L2 overlay services with BGP as control plane and SRv6 as dataplane.

SRv6 SID refers to a SRv6 Segment Identifier as defined in [I-D.filsfils-spring-srv6-network-programming].

SRv6 Service SID refers to an SRv6 SID that MAY be associated with one of the service specific behavior on the advertising PE, such as (but not limited to) in the case of L3VPN service, END.DT (crossconnect to a VRF) or END.DX (crossconnect to a nexthop) functions as defined in[I-D.filsfils-spring-srv6-network-programming].

To provide SRv6 Service service with best-effort connectivity, the egress PE signals an SRv6 Service SID with the VPN route. The ingress PE encapsulates the VPN packet in an outer IPv6 header where the destination address is the SRv6 Service SID provided by the egress PE. The underlay between the PE's only need to support plain IPv6 forwarding [RFC2460].

To provide SRv6 Service service in conjunction with an underlay SLA from the ingress PE to the egress PE, the egress PE colors the overlay VPN route with a color extended community[I-D.ietf-idr-segment-routing-te-policy]. The ingress PE encapsulates the VPN packet in an outer IPv6 header with an SRH that contains the SR policy associated with the related SLA followed by the SRv6 Service SID associated with the route. The underlay nodes whose SRv6 SID's are part of the SRH must support SRv6 data plane.

BGP is used to advertise the reachability of prefixes in a particular VPN from an egress Provider Edge (egress-PE) to ingress Provider Edge (ingress-PE) nodes.

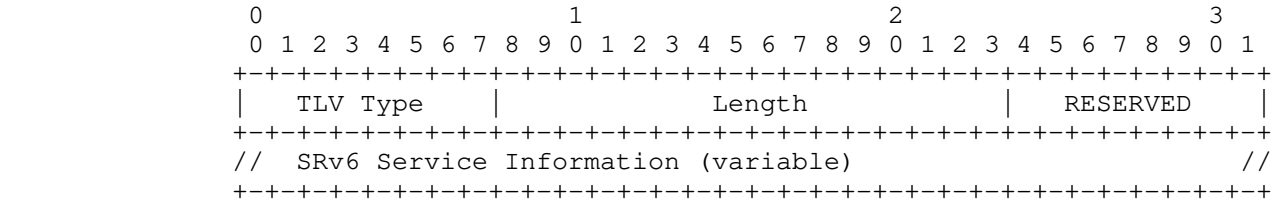
This document describes how existing BGP messages between PEs may carry SRv6 Segment IDs (SIDs) as a means to interconnect PEs and form VPNs.

2. SRv6 Services TLV

The SRv6 Service TLVs are defined as two new TLVs for BGP Prefix SID Attribute [I-D.ietf-idr-bgp-prefix-sid], to achieve signaling of SRv6 Service SID for L3 and L2 services.

BGP Prefix SID Attribute[I-D.ietf-idr-bgp-prefix-sid]is referred as BGP SID Attribute in the rest of the document.

When an egress-PE is capable of SRv6 data-plane, it SHOULD signal SRv6 Service SID TLV within the BGP SID Attribute attached to MP-BGP NLRI defined in [RFC4659][RFC5549][RFC7432]. [RFC4364]

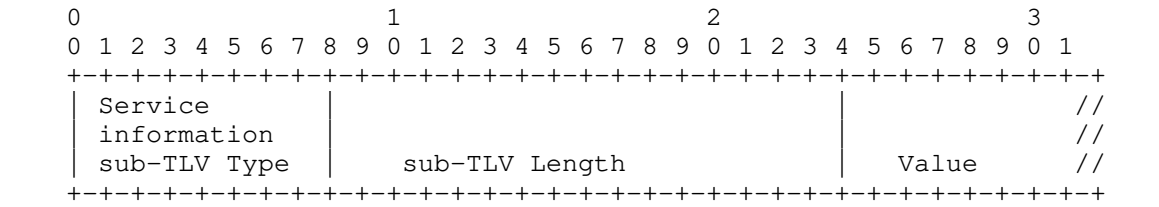


This document defines the following two new TLVs for BGP SID Attribute.

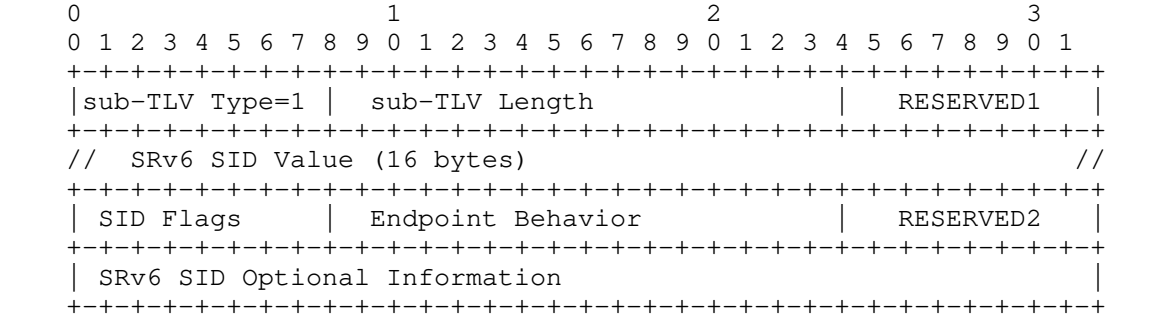
- SRv6 L3 Service TLV. Type code 5 (to be assigned by IANA as described in section 8). This TLV encodes Service SID information for the SRv6 based L3 services. It corresponds to the equivalent functionality provided by an MPLS Label when received with a Layer 3 VPN route [RFC4364]. Some functions which MAY be encoded, but not limited to, are End.DX4, End.DT4, End.DX6, End.DT6, etc.

- SRv6 L2 Service TLV. Type code 6 (to be assigned by IANA as described in section 8). This TLV encodes Service SID information for the SRv6 based L2 services. It corresponds to the equivalent functionality provided by an MPLS Label1 for EVPN Route-Types as defined in [RFC7432]. Some functions which MAY be encoded, but not limited to, are End.DX2, End.DX2V, End.DT2U, End.DT2M etc.

The "SRv6 Service Information" is encoded as an un-ordered list of sub-TLVs ("Type/Length/Value" blocks), as following:



This document defines a sub-TLV Type code to encode a single SRv6 SID value along with its properties as following:



Where:

- Type is 1 (to be assigned by IANA as described in Section 8). As defined to be "SID information sub-TLV".

- Length: 16 bit field. The total length of the value portion of the sub-TLV.

- RESERVED1: 8 bit field. SHOULD be 0 on transmission and MUST be ignored on reception.

- SRv6 SID Value: 128 bit field. Encodes an SRv6 SID as defined in [I-D.filsfils-spring-srv6-network-programming]

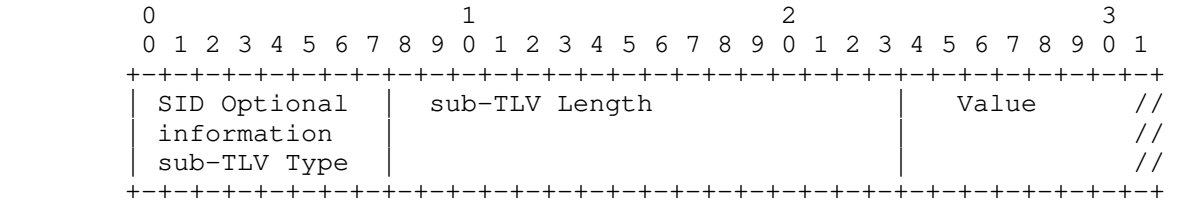
- SID Flags: 8 bit field. Encodes SRv6 SID Flags. Value is opaque to BGP.

- Endpoint Behavior : 16 bit field. Encodes Endpoint behavior. For SRv6 VPN services, this field is always set to (0xFFFF).

- RESERVED2: 8 bit field. SHOULD be 0 on transmission and MUST be ignored on reception.

- SRv6 SID Optional Information. Variable length. Encodes optional properties as described below.

SRv6 SID Optional information is encoded as a list of "SID optional information sub-TLV" blocks. Where each block is encoded as Type/Length/Value triplet.



No Type codes for SID Optional information sub-TLV are defined at this point.

3. BGP based L3 over SRv6

BGP egress nodes (egress-PEs) advertise a set of reachable prefixes. Standard BGP update propagation schemes [RFC4271], which MAY make use of route reflectors [RFC4456], are used to propagate these prefixes. BGP ingress nodes (ingress-PE) receive these advertisements and may add the prefix to the RIB in an appropriate VRF.

Egress-PEs which supports SRv6-VPN advertises a Service SID encoded within SRv6 Service TLV within BGP SID attribute, with the VPN routes. The Service SID thus signaled only has local significance at the egress-PE, where it is allocated or configured on a per-CE or per-VRF basis. In practice, the SID encodes a cross-connect to a specific Address Family table (END.DT) or next-hop/interface (END.DX) as defined in the SRv6 Network Programming Document [I-D.filsfils-spring-srv6-network-programming].

The SRv6 Service SID MAY be routable within the AS of the egress-PE and serves the dual purpose of providing reachability between ingress-PE and egress-PE while also encoding the VPN identifier.

To support SRv6 based L3VPN overlay, a SID is advertised with BGP MPLS L3VPN route update[RFC4364]. SID is encoded in a SRv6 Service SID TLV within the optional transitive BGP SID attribute[I-D.ietf-idr-bgp-prefix-sid]. This attribute serves two purposes; first it indicates that the BGP egress device is reachable via an SRv6 underlay and the BGP ingress device receiving this route MAY choose to encapsulate or insert an SRv6 SRH, second it indicates the value of the SID to include in the SRH encapsulation. For L3VPN, only a single SRv6 Service SID MAY be necessary. A BGP speaker supporting an SRv6 underlay MAY distribute SID per route via the SRv6 Service TLV. If the BGP speaker supports MPLS based L3VPN simultaneously, it MAY also populate the Label values in L3VPN route

NLRI, and allow the BGP ingress device to decide which encapsulation to use. If the BGP speaker does not support MPLS based L3VPN services the MPLS Labels in L3VPN NLRI MUST be set to IMPLICIT-NULL.[RFC7432]

At an ingress-PE, BGP installs the advertised prefix in the correct RIB table, recursive via an SR Policy leveraging the received SRv6 Service SID.

Assuming best-effort connectivity to the egress PE, the SR policy has a path with a SID list made up of a single SID: the SRv6 Service SID received with the related BGP route update.

However, when VPN route is colored with an extended color community C and signaled with Next-Hop N and the ingress PE has a valid SRv6 Policy (N, C) associated with SID list <S1,S2, S3> [I-D.filsfils-spring-segment-routing-policy] then the SR Policy is <S1, S2, S3, SRv6 Service SID>.

Multiple VPN routes MAY resolve recursively on the same SR Policy.

3.1. IPv4 VPN Over SRv6 Core

IPv4 VPN Over IPv6 Core is defined in [RFC5549], the MP_REACH_NLRI is encoded as follows for an SRv6 Core:

- AFI = 1
- SAFI = 128
- Length of Next Hop Network Address = 16 (or 32)
- Network Address of Next Hop = IPv6 address of the egress PE
- NLRI = IPv4-VPN routes
- Label = Implicit-Null

SRv6 Service SID is encoded as part of the SRv6 Service SID TLV defined in Section 2. The function of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the function may likely be End.DX4 or End.DT4.

3.2. IPv6 VPN Over SRv6 Core

IPv6 VPN over IPv6 Core is defined in [RFC4659], the MP_REACH_NLRI is enclosed as follows for an SRv6 Core:

- AFI = 2
- SAFI = 128
- Length of Next Hop Network Address = 16 (or 32)
- Network Address of Next Hop = IPv6 address of the egress PE
- NLRI = IPv6-VPN routes
- Label = Implicit-Null

SRv6 Service SID are encoded as part of the SRv6 Service SID TLV defined in Section 2. The function of the IPv6 SRv6 SID is entirely up to the originator of the advertisement. In practice the function may likely be End.DX6 or End.DT6.

3.3. Global IPv4 over SRv6 Core

IPv4 over IPv6 Core is defined in [RFC5549]. The MP_REACH_NLRI is encoded with:

- AFI = 1
- SAFI = 1
- Length of Next Hop Network Address = 16 (or 32)
- Network Address of Next Hop = IPv6 address of Next Hop
- NLRI = IPv4 routes

SRv6 SID for Global IPv4 routes is encoded as part of the SRv6 Service SID defined in Section 2. The function of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the function may likely be End.DX6 or End.DT6.

3.4. Global IPv6 over SRv6 Core

The MP_REACH_NLRI is encoded with:

- AFI = 2
- SAFI = 1
- Length of Next Hop Network Address = 16 (or 32)
- Network Address of Next Hop = IPv6 address of Next Hop

- NLRI = IPv6 routes

SRv6 SID for Global IPv6 routes is encoded as part of the SRv6 Service SID defined in Section 2. The function of the SRv6 SID is entirely up to the originator of the advertisement. In practice, the function may likely be End.DX6 or End.DT6.

Also, by utilizing the SRv6 Service SID TLV, as defined in Section 2, to encode the Global SID, BGP free core is possible by encapsulating all BGP traffic from edge to edge over SRv6.

4. BGP based Ethernet VPN(EVPN) over SRv6

Ethernet VPN(EVPN), as defined in [RFC7432] provides an extendable method of building an EVPN overlay. It primarily focuses on MPLS based EVPNs but calls out the extensibility to IP based EVPN overlays. It defines 4 route-types which carry prefixes and MPLS Label attributes, the Labels each have specific use for MPLS encapsulation of EVPN traffic. The fifth route-type carrying MPLS label information (and thus encapsulation information) for EVPN is defined in[I-D.ietf-bess-evpn-prefix-advertisement]. The Route Types discussed below are:

- Ethernet Auto-discovery Route
- MAC/IP Advertisement Route
- Inclusive Multicast Ethernet Tag Route
- Ethernet Segment route
- IP prefix route
- Selective Multicast route
- IGMP join sync route
- IGMP leave sync route

To support SRv6 based EVPN overlays a SRv6 Service SID is advertised in route-type 1,2,3 and 5 above. The SRv6 Service SID (or list of those, when applicable) per route-type are advertised in SRv6 Service TLV, as described in section 2. Signaling of SRv6 Service SID serves two purposes; first it indicates that the BGP egress device is reachable via an SRv6 underlay and the BGP ingress device receiving this route MAY choose to encapsulate or insert an SRv6 SRH, second it indicates the value of the SID or SIDs to include in the SRH encapsulation. If the BGP speaker does not support MPLS based EVPN

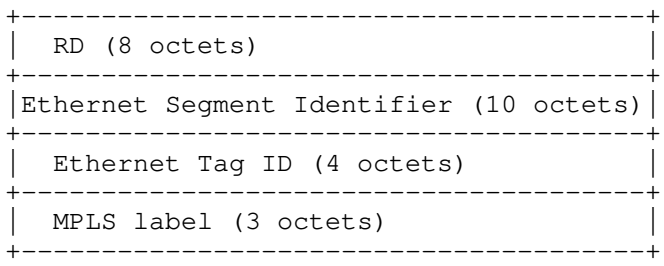
services the MPLS Labels in EVPN route types MUST be set to IMPLICIT-NULL.

4.1. Ethernet Auto-discovery Route over SRv6 Core

Ethernet Auto-discovery (A-D) routes are Type-1 route type defined in [RFC7432]and may be used to achieve split horizon filtering, fast convergence and aliasing. EVPN route type-1 is also used in EVPN-VPWS as well as in EVPN flexible cross-connect; mainly used to advertise point-to-point services id.

Multi-homed PEs MAY advertise an Ethernet auto discovery route per Ethernet segment with the introduced ESI MPLS label extended community defined in [RFC7432].The extended community label is set to implicit-null. PEs may identify other PEs connected to the same Ethernet segment after the EVPN type-4 ES route exchange. All the multi-homed and remote PEs that are part of same EVI may import the auto discovery route.

EVPN Route Type-1 is encoded as follows for SRv6 Core:



For a SRv6 only BGP speaker for an SRv6 Core:

- SRv6 Service SID TLV MAY be advertised with the route.

4.1.1. EVPN Route Type-1(Per ES AD)

Where:

- BGP next-hop: IPv6 address of an egress PE
- Ethernet Tag ID: all FFFF's
- MPLS Label: always set to zero value
- Extended Community: Per ES AD, ESI label extended community

BGP SID Attribute with SRv6 Service TLV MAY be advertised along with the route advertisement and the behavior of the SRv6 Service SID thus signaled, is entirely up to the originator of the advertisement. This is typically used to signal Arg.FE2 SID argument for applicable End.DT2M SIDs.

4.1.2. Prefix Type-1(Per EVI/ES AD)

Where:

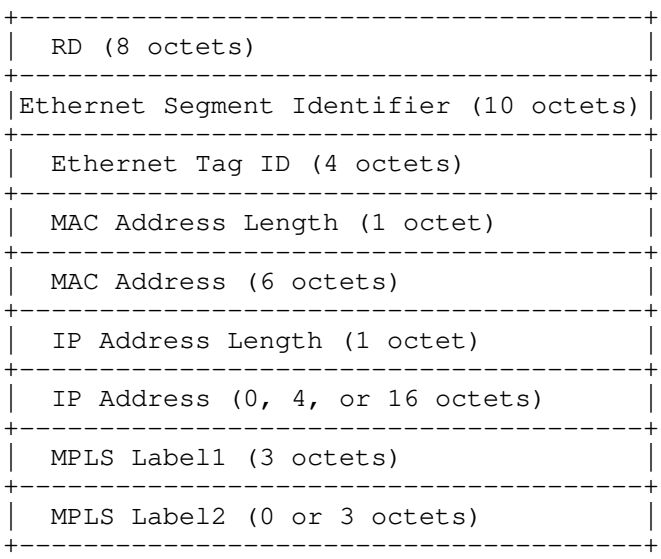
- BGP next-hop: IPv6 address of an egress PE
- Ethernet Tag ID: non-zero for VLAN aware bridging, EVPN VPWS and FXC
- MPLS Label: Implicit-Null

BGP SID Attribute with SRv6 Service TLV MAY be advertised along with the route advertisement and the behavior of the SRv6 Service SID is entirely up to the originator of the advertisement. In practice, the behavior would likely be END.DX2, END.DX2V or END.DT2U.

4.2. MAC/IP Advertisement Route(Type-2) with SRv6 Core

EVPN route type-2 is used to advertise unicast traffic MAC+IP address reachability through MP-BGP to all other PEs in a given EVPN instance.

A MAC/IP Advertisement route type is encoded as follows for SRv6 Core:



where:

- BGP next-hop: IPv6 address of an egress PE
- MPLS Label1: Implicit-null
- MPLS Label2: Implicit-null

BGP SID Attribute with SRv6 Service TLV MAY be advertised. The behavior of the SRv6 Service SID is entirely up to the originator of the advertisement. In practice, the behavior of the SRv6 SID is as follows:

- END.DX2, END.DT2U (Layer 2 portion of the route)
- END.DT6/4 or END.DX6/4 (Layer 3 portion of the route)

Described below are different types of Type-2 advertisements.

- MAC/IP Advertisement Route(Type-2) with MAC Only
  - BGP next-hop: IPv6 address of egress PE
  - MPLS Label1: Implicit-null
  - MPLS Label2: Implicit-null

    * SRv6 Service SID TLV within BGP SID Attribute MAY encode END.DX2 or END.DT2U behavior

- MAC/IP Advertisement Route(Type-2) with MAC+IP

    * BGP next-hop: IPv6 address of egress PE

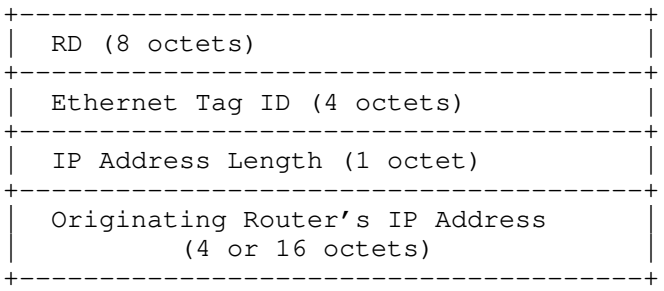
    * MPLS Label1: Implicit-Null

    * MPLS Label2: Implicit-Null

    * SRv6 Service TLV within BGP SID Attribute MAY encode Layer2 END.DX2 or END.DT2U behavior and Layer3 END.DT6/4 or END.DX6/4 behavior

4.3. Inclusive Multicast Ethernet Tag Route with SRv6 Core

EVPN route Type-3 is used to advertise multicast traffic reachability information through MP-BGP to all other PEs in a given EVPN instance.



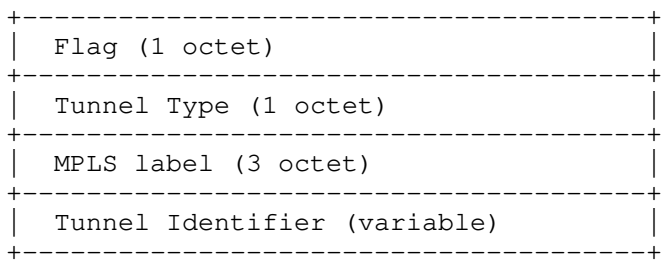
An Inclusive Multicast Ethernet Tag route type specific EVPN NLRI consists of the following [RFC7432] where:

- BGP next-hop: IPv6 address of egress PE

- SRv6 Service TLV MAY encode END.DX2/END.DT2M function.

- BGP Attribute: PMSI Tunnel Attribute[RFC6514] MAY contain MPLS implicit-null label and Tunnel Type would be similar to defined in EVPN Type-6 i.e. Ingress replication route.

The format of PMSI Tunnel Attribute attribute is encoded as follows for an SRv6 Core:



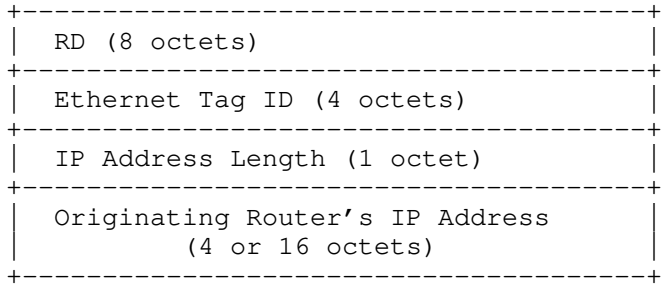
- Flag: zero value defined per [RFC7432]
- Tunnel Type: defined per [RFC6514]
- MPLS label: Implicit-Null
- Tunnel Identifier: IP address of egress PE

SRv6 Service TLV may be encoded as part of BGP SID Attribute. The behavior of the SRv6 Service SID is entirely up to the originator of the advertisement. In practice, the behavior of the SRv6 SID is as follows:

- END.DX2 or END.DT2M function
- The ESI Filtering argument(Arg.FE2) carried along with EVPN Route Type-1 (in SRv6 VPN SID), MAY be merged together with the applicable End.DT2M SID advertised by remote PE by doing a bitwise logical OR to create a single SID on the ingress PE for Split-horizon and other filtering mechanisms. Details of filtering mechanisms are described in[RFC7432]

4.4. Ethernet Segment Route with SRv6 Core

An Ethernet Segment route type specific EVPN NLRI consists of the following defined in [RFC7432]



where:

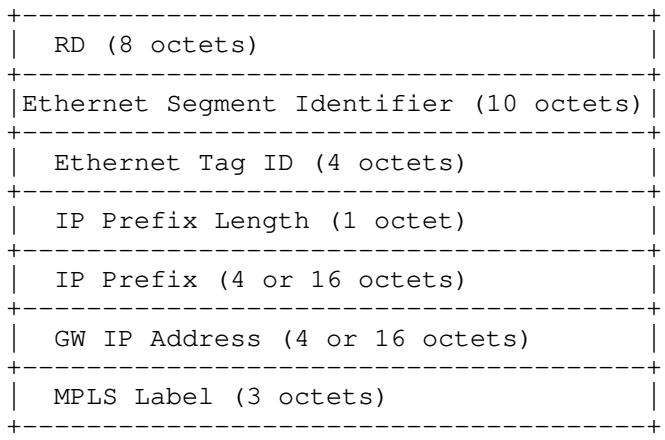
- o BGP next-hop: IPv6 address of egress PE

As opposed to the previous route types, SRv6 Service TLV as part of BGP SID Attribute, is NOT advertised along with the route. The processing of that route has not changed; it remains as described in [RFC7432].

4.5. IP prefix router(Type-5) with SRv6 Core

EVPN route Type-5 is used to advertise IP address reachability through MP-BGP to all other PEs in a given EVPN instance. IP address may include host IP prefix or any specific subnet. EVPN route Type-5 is defined in[I-D.ietf-bess-evpn-prefix-advertisement]

An IP Prefix advertisement is encoded as follows for an SRv6 Core:



- o BGP next-hop: IPv6 address of egress PE
- o MPLS Label: Implicit-Null

BGP SID Attribute with SRv6 Service TLV MAY be advertised. The behavior of the SRv6 Service SID is entirely up to the originator of the advertisement. In practice, the behavior of the SRv6 SID is an End.DT6/4 or End.DX6/4.

4.6. Multicast routes (EVPN Route Type-6, Type-7, Type-8)

These routes do not require any additional SRv6 Service TLV. As per EVPN route-type 4, the BGP nexthop is equal to the IPv6 address of

egress PE. More details may be added in future revisions of this document.

5. Migration from L3 MPLS based Segment Routing to SRv6 Segment Routing

Migration from IPv4 to IPv6 is independent of SRv6 BGP endpoints, and the selection of which route to use (received via the IPv4 or IPv6 session) is a local configurable decision of the ingress-PE, and is outside the scope of this document.

Migration from IPv6 MPLS based underlay to an SRv6 underlay with BGP speakers is achieved with a few simple rules at each BGP speaker.

```
At Egress-PE
  If BGP offers an SRv6 Service service
      Then BGP allocates an SRv6 Service SID for the VPN service
      and adds the BGP SRv6 Service SID TLV while advertising VPN prefixes.
  If BGP offers an MPLS VPN service
      Then BGP allocates an MPLS Label for the VPN service and
      use it in NLRI as normal for MPLS L3 VPNs.
  else MPLS label for VPN service is set to IMPLICIT-NULL.

At Ingress-PE
  *Selection of which encapsulation below (SRv6 Service or MPLS-VPN) is
   defined by local BGP policy
  If BGP supports SRv6 Service service, and
  receives a BGP SID Attribute with an SRv6 Service TLV encoding a SRv6 Service
SID
      Then BGP programs the destination prefix in RIB recursive via
      the related SR Policy.
  If BGP supports MPLS VPN service, and
  the MPLS Label is not Implicit-Null
      Then the MPLS label is used as a VPN label and inserted with the
      prefix into RIB via the BGP Nexthop.
```

6. Implementation Status

The SRv6 Service is available for SRv6 on various Cisco hardware and other software platforms. An end-to-end integration of SRv6 L3VPN, SRv6 Traffic-Engineering and Service Chaining. All of that with data-plane interoperability across different implementations [1]:

- Three Cisco Hardware-forwarding platforms: ASR 1K, ASR 9k and NCS 5500
- Huawei network operating system
- Two Cisco network operating systems: IOS XE and IOS XR

- Barefoot Networks Tofino on OCP Wedge-100BF
- Linux Kernel officially upstreamed in 4.10
- Fd.io

7. Error Handling of BGP SRv6 SID Updates

If the SRv6 Service TLV within the received BGP SID Attribute is malformed, consider the entire BGP SID Attribute as malformed, discard it and not propagate it further to other peers i.e. use the -attribute discard- action specified in [RFC7606] an error MAY be logged for further analysis.

The SRv6 Service TLV is not considered to be malformed in the following cases. The rest of the BGP SID Attribute MUST be processed normally. An error MAY be logged for further analysis.

- The Service Information sub-TLV Type is unrecognized: all unrecognized sub-TLV Types must be stored locally and propagated further to other peers. It is a matter of local implementation whether to use locally any recognized SID Types that may be present in the TLV along with the unrecognized Types.

In addition, the following rules apply for processing NLRIs received with BGP SID Attribute containing SRv6 Service TLV:

- If the TLV is advertised by a CE peer, the receiving PE may discard it before advertising the route to its PE peers.
- If the received NLRI has neither a valid SRv6 Service SID nor a valid MPLS label as specified in [RFC4659][RFC5549][RFC7432] , the NLRI MUST be considered unreachable i.e. apply the -treat as withdraw- action specified in [RFC7606].

8. IANA Considerations

This document defines a new TLV, SRv6 Service TLV, within BGP SID attribute. This document defines the following new TLV Types of BGP SID attribute:

- Type 5: SRv6 Layer3 Service
- Type 6: SRv6 Layer2 Service

and are assigned to SRv6 Layer3 Service TLV and SRv6 Layer2 Service TLV defined in this document.

Further, this document defines a new sub-TLV; namely Service information sub-TLV, within SRv6 Service TLV, as described in Section 2. A new registry "BGP SRv6 Service Information sub-TLV Types" is required and a new Type code point with value 1, is requested in this registry, to denote "SID information sub-TLV".

Further, this document defines new optional sub-TLVs, namely "SID optional information sub-TLV" within Service information sub-TLV, as described in Section 2. New registry for this purpose is required.

9. Security Considerations

This document introduces no new security considerations beyond those already specified in [RFC4271] and [RFC8277].

10. Conclusions

This document proposes extensions to the BGP to allow advertising certain attributes and functionalities related to SRv6.

11. References

11.1. Normative References

[I-D.filsfils-spring-segment-routing-policy]
Filsfils, C., Sivabalan, S., Hegde, S., daniel.voyer@bell.ca, d., Lin, S., bogdanov@google.com, b., Krol, P., Horneffer, M., Steinberg, D., Decraene, B., Litkowski, S., Mattes, P., Ali, Z., Talaulikar, K., Liste, J., Clad, F., and K. Raza, "Segment Routing Policy Architecture", draft-filsfils-spring-segment-routing-policy-06 (work in progress), May 2018.

[I-D.filsfils-spring-srv6-network-programming]
Filsfils, C., Camarillo, P., Leddy, J., daniel.voyer@bell.ca, d., Matsushima, S., and Z. Li, "SRv6 Network Programming", draft-filsfils-spring-srv6-network-programming-05 (work in progress), July 2018.

[I-D.ietf-6man-segment-routing-header]
Filsfils, C., Previdi, S., Leddy, J., Matsushima, S., and d. daniel.voyer@bell.ca, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-14 (work in progress), June 2018.

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

[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>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[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>.

[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, <https://www.rfc-editor.org/info/rfc7606>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

11.2. Informative References

[I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.

[I-D.ietf-idr-bgp-prefix-sid]
Previdi, S., Filsfils, C., Lindem, A., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix SID extensions for BGP", draft-ietf-idr-bgp-prefix-sid-27 (work in progress), June 2018.

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Jain, D., Mattes, P., Rosen, E., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-04 (work in progress), July 2018.

[I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Ginsberg, L., Filsfils, C., Bashandy, A., Gredler, H., Litkowski, S., Decraene, B., and J. Tantsura, "IS-IS Extensions for Segment Routing", draft-ietf-isis-segment-routing-extensions-19 (work in progress), July 2018.

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-15 (work in progress), January 2018.

[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>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, DOI 10.17487/RFC4659, September 2006, <https://www.rfc-editor.org/info/rfc4659>.

[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, <https://www.rfc-editor.org/info/rfc5549>.

11.3. URIs

[1] http://www.segment-routing.net

Appendix A. Acknowledgements

The authors would like to thank Shyam Sethuram for comments and discussion of TLV processing and validation.

Appendix B. Contributors

Bart Peirens
Proximus
Belgium

Email: bart.peirens@proximus.com

Authors' Addresses

Gaurav Dawra (editor)
LinkedIn
USA

Email: gdawra.ietf@gmail.com

Clarence Filsfils
Cisco Systems
Belgium

Email: cfilsfil@cisco.com

Darren Dukes
Cisco Systems
Canada

Email: ddukes@cisco.com

Patrice Brissette
Cisco Systems
Canada

Email: pbrisset@cisco.com

Pablo Camarilo
Cisco Systems
Spain

Email: pcamaril@cisco.com

Jonn Leddy
Comcast
USA

Email: john_leddy@cable.comcast.com

Daniel Voyer
Bell Canada
Canada

Email: daniel.voyer@bell.ca

Daniel Bernier
Bell Canada
Canada

Email: daniel.bernier@bell.ca

Dirk Steinberg
Steinberg Consulting
Germany

Email: dws@steinberg.net

Robert Raszuk
Bloomberg LP
USA

Email: robert@raszuk.net

Bruno Decraene
Orange
France

Email: bruno.decraene@orange.com

Satoru Matsushima
SoftBank
1-9-1,Higashi-Shimbashi,Minato-Ku
Japan 105-7322

Email: satoru.matsushima@g.softbank.co.jp

Shunwan Zhuang
Huawei Technologies
China

Email: zhuangshunwan@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 6, 2021

J. Dong
S. Zhuang
Huawei Technologies
G. Van de Velde
Nokia

November 2, 2020

BGP Extended Community for Identifying the Target Nodes
draft-dong-idr-node-target-ext-comm-03

Abstract

BGP has been used to distribute different types of routing and policy information. In some cases, the information distributed may be only intended for one or a particular group of BGP nodes in the network. Currently BGP does not have a generic mechanism of designating the target nodes of the routing information. This document defines a new type of BGP Extended Community called "Node Target". The mechanism of using the Node Target Extended Community to steer BGP route distribution to particular BGP nodes is specified.

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 May 6, 2021.

Table of Contents

1. Introduction

BGP [RFC4271] has been used to distribute different types of routing and policy information. In some cases, the information distributed may be only intended for one or a particular group receiving BGP nodes in the network. One typical use case is the distribution of BGP Flow Spec [I-D.ietf-idr-rfc5575bis] [I-D.ietf-idr-flow-spec-v6] rules only to a particular group of BGP nodes. Such a targeting mechanism is considered useful that it can save the resources on nodes which do not need that information.

Currently BGP does not have a generic mechanism of designating the set of nodes to which the information is to be distributed. Route Target (RT) as defined in [RFC4364] was designed for the matching of VPN routes into the target VPN Routing and Forwarding tables (VRFs) on PE nodes. Although [I-D.ietf-idr-segment-routing-te-policy] introduces the mechanism of steering the SR policy information to the target head end node based on RT, it is only defined for the SR

Policy Address Family. Although it is possible to reuse RTs to control the distribution of non-VPN information to one or a group of receiving nodes, such mechanism is not applicable when the information to be distributed is VPN-specific and is advertised with a set of RTs for the VRF matching. In that case, the matching of any of the VPN RTs in the Update will result in the information eligible for installation, regardless of whether the RTs representing the target nodes are matched or not. Thus a mechanism which is independent from the control of VPN route to VRF distribution is needed.

Another possible approach is to configure, on each router, a community and the corresponding policies to match the community to determine whether to accept the received routes. Such mechanism relies on manual configuration thus is considered error-prone. It is preferable by some operators that an automatic approach can be provided, which would make the operation much easier.

This document defines a new type of BGP Extended Community called "Node Target". The mechanism of using the Node Target extended community to steer BGP route distribution to particular BGP nodes is also specified.

2. Node Target Extended Communities

This section defines a new BGP Extended Community [RFC4360] called "Node Target Extended Community". It can be a transitive extended community with the high-order octect of the type set to 0x01, or a non-transitive extended community with the high-order octect type set to 0x41. The sub-type of the Node Target Extended Community is TBA.

The format of Node Target Extended Community is shown in Figure 1.

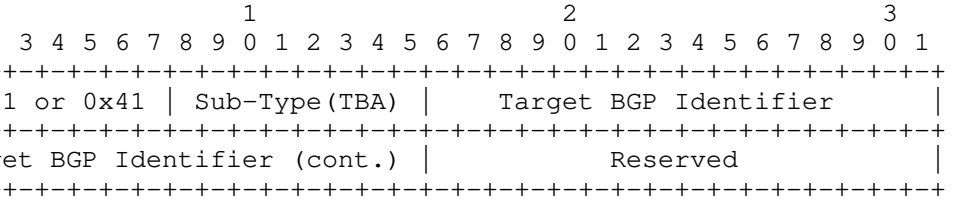


Figure 1. Node Target extended community

Where:

Target BGP Identifier (4 octets): The BGP Identifier of a target node. It is a 4-octet, unsigned, non-zero integer as defined in [RFC6286].

Reserved field (2 octets): Reserved for future use, MUST be set to zero on transmission and ignored on receipt.

One or more Node Target extended communities MAY be carried in an Update message to designate a group of target BGP nodes.

3. Procedures

In this section, the mechanism for intra-domain scenario is described, the mechanism for inter-domain scenario is for further study. The domain here refers to an administrative domain, which may consists of one or multiple ASes managed by a single operator.

When a network controller or BGP speaker plans to advertise some BGP routing or policy information only to one or a group of BGP nodes in the network, it MUST put the BGP Identifier of each target node into the Node Target extended communities, and attach the Node Target extended communities to the routes or policies to be advertised.

When a BGP speaker receives a BGP Update which contains one or more Node Target extended communities, it MUST check the target BGP Identifiers carried in the Node Target extended communities of the Update.

- If the target BGP Identifier in any of the Node Target extended community matches with the local BGP Identifier, this node is one of the target nodes of the Update, the information in the Update is eligible to be kept and installed on this node.
  - If this node is a Route Reflector, and in the Update there is one or more Node Target extended communities which contains non-local BGP Identifiers, information in the Update are eligible be reflected to its peers according to the rules defined in [RFC4456]. The RR may check the BGP Identifiers of its peers to determine the set of peers which are the target nodes of the Update, and only reflect the information in the Update to the matched BGP peers.
  - If this node is an Autonomous System Border Router (ASBR), and the BGP Identifiers of one or more of its EBGP peers match with the Node Target extended communities in the Update, information in the Update is eligible to be advertised to the matched EBGP peers.
- If the target BGP Identifier in any of the Node target extended community does not match with the local BGP Identifier, this node is not the target node of Update, the information in the Update is not eligible to be installed on this node.

* If this node is a Route Reflector, information in the Update is eligible to be reflected to its peers according to the rules defined in [RFC4456]. The RR may check the BGP Identifiers of its peers to determine the set of peers which are the target nodes of the Update, and only reflect the information in the Update to the matched BGP peers.

4. Compatibility Considerations

The Node Target extended community introduced in this document can be deployed incrementally in the network. For BGP speakers which understand the Node Target extended community, it is used to determine whether the nodes are the target nodes of the Update. For BGP speakers which do not understand the Node Target extended community, it will be ignored and the information in the Update will be processed and advertised based on normal BGP procedure. Although this could ensure that the target nodes can always obtain the information needed, this may result in unnecessary state maintained on legacy BGP speakers. And if the information advertised is the Flow Spec rules, the legacy BGP speakers may install unnecessary flowspec rules, this may have impact on traffic which matches such rules, thus may result in unexpected traffic steering or filtering behaviors on such nodes. This may be mitigated by setting appropriate routing policies on the legacy BGP nodes.

5. IANA Considerations

This document requests that IANA assigns one new sub-type for "Node Target Extended Community" from the "Transitive IPv4-Address-Specific Extended Community" registry of the "BGP Eextended Communities" registry.

This document requests that IANA assigns the same sub-type for "Node Target Extended Community" from the "Non-Transitive IPv4-Address-Specific Extended Community" registry of the "BGP Eextended Communities" registry.

6. Security Considerations

This document does not change the security properties of BGP.

7. Contributors

Haibo Wang
Email: rainsword.wang@huawei.com

8. Acknowledgements

The authors would like to thank Zhenbin Li, Ercin Torun, Jeff Haas and Robert Raszuk for the review and discussion of this document.

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>.

[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>.

[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>.

9.2. Informative References

[I-D.ietf-idr-flow-spec-v6] Loibl, C., Raszuk, R., and S. Hares, "Dissemination of Flow Specification Rules for IPv6", draft-ietf-idr-flow-spec-v6-18 (work in progress), November 2020.

[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-27 (work in progress), October 2020.

[I-D.ietf-idr-segment-routing-te-policy] Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-09 (work in progress), May 2020.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[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>.

[RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, DOI 10.17487/RFC6286, June 2011, <https://www.rfc-editor.org/info/rfc6286>.

Authors' Addresses

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

Email: jie.dong@huawei.com

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

Email: zhuangshunwan@huawei.com

Gunter Van de Velde
Nokia
Antwerp
BE

Email: gunter.van_de_velde@nokia.com

IDR and SIDR
Internet-Draft
Intended status: Standards Track
Expires: October 20, 2019

K. Sriram, Ed.
USA NIST
A. Azimov, Ed.
Yandex

April 18, 2019

Methods for Detection and Mitigation of BGP Route Leaks
draft-ietf-idr-route-leak-detection-mitigation-11

Abstract

Problem definition for route leaks and enumeration of types of route leaks are provided in RFC 7908. This document describes a solution for detection and mitigation route leaks which is based on conveying route-leak protection (RLP) information in a Border Gateway Protocol (BGP) community. The RLP information is carried in a new well-known transitive BGP community, called the RLP community. The RLP community helps with detection and mitigation of route leaks at ASes downstream from the leaking AS (in the path of the BGP update). This is an inter-AS (multi-hop) solution mechanism. This solution complements the intra-AS (local AS) route-leak avoidance solution that is described in ietf-idr-bgp-open-policy draft.

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 October 20, 2019.

Table of Contents

1. Introduction

RFC 7908 [RFC7908] provides a definition of the route leak problem, and enumerates several types of route leaks. For this document, the definition that is applied is that a route leak occurs when a route received from a transit provider or a lateral peer is forwarded (against commonly used policy) to another transit provider or a lateral peer. The commonly used policy is that a route received from a transit provider or a lateral peer may be forwarded "down only" to customers.

This document describes a solution for detection and mitigation route leaks which is based on conveying route-leak protection (RLP) information in a Border Gateway Protocol (BGP) community. The RLP information is carried in a new well-known transitive BGP community, called the RLP community. The RLP community helps with detection and mitigation of route leaks at ASes downstream from the leaking AS (in the path of the BGP update). This is an inter-AS (multi-hop) solution mechanism. This solution complements the intra-AS (local

AS) route-leak avoidance solution that is described in [I-D.ietf-idr-bgp-open-policy].

Previously, an optional transitive BGP RLP Attribute was proposed to carry the RLP information (in earlier versions of this document). However, this updated document proposes a well-known transitive BGP community to carry the RLP information, with the intention of promoting faster adoption.

The inter-AS RLP mechanism described here can be incrementally deployed. Early adopters would see significant benefits. If a group of big ISPs deploy RLP, then they would be helping each other by blocking route leaks originated within one's customer cone from propagating into a peer's AS or their customer cone.

2. Mechanisms for Detection and Mitigation of Route Leaks

There are two considerations for route leaks: (1) Prevention of route leaks from a local AS [I-D.ietf-idr-bgp-open-policy], and (2) Detection and mitigation of route leaks in ASes that are downstream from the leaking AS (in the path of BGP update). This document specifies the latter.

2.1. Ascertaining Peering Relationship

There are four possible peering relationships (i.e., roles) an AS can have with a neighbor AS: (1) Provider: transit-provider for all prefixes exchanged, (2) Customer: customer for all prefixes exchanged, (3) Lateral Peer: lateral peer (i.e., non-transit) for all prefixes exchanged, and (4) Complex: different relationships for different sets of prefixes [Luckie]. For the complex case, the peering role types provider, customer, or lateral peer apply for different non-overlapping sets of prefixes.

Operators rely on some form of out-of-band (OOB) (i.e., external to BGP) communication to exchange information about their peering relationship, AS number, interface IP address, etc. If the relationship is complex, the OOB communication also includes the sets of prefixes for which they have different roles. [I-D.ietf-idr-bgp-open-policy] introduces a method of re-confirming the BGP Role during BGP OPEN messaging (except when the role is complex). It defines a new BGP Role capability, which helps in re-confirming the relationship when it is provider, customer, or lateral peer. BGP Role does not replace the OOB communication since it relies on the OOB communication to set the role type in the BGP OPEN message. However, BGP Role provides a means to double check, and if there is a contradiction detected via the BGP Role messages, then a Role Mismatch Notification is sent [I-D.ietf-idr-bgp-open-policy].

When the BGP relationship information has been correctly exchanged including the sets of prefixes with different roles (if complex), then this information SHOULD be used to automatically set the role per-prefix with each peer. For example, if the local AS's role is Provider with a neighbor AS, then the per-prefix role is set to 'Provider' for all prefixes sent to the neighbor, and set to 'Customer' for all prefixes received from the neighbor.

Once the per-prefix roles are set, this information is used in the RLP solution mechanism that is described in this document.

2.2. Route-Leak Protection (RLP) Semantics

The key principle is that, in the event of a route leak, a receiving router in a transit-provider AS (e.g., referring to Figure 1, ISP2 (AS2) router) should be able to detect from the RLP community in the update message that its customer AS (e.g., AS3 in Figure 1) should not have forwarded the update (towards the transit-provider AS). Likewise when the update is received from a lateral peer. This means that at least one of the ASes in the AS path of the update put RLP information in RLP community to indicate that it sent the update to its customer or lateral peer, but forbade any subsequent 'Up' (customer to provider) or 'Lateral' (peer to peer) forwarding.

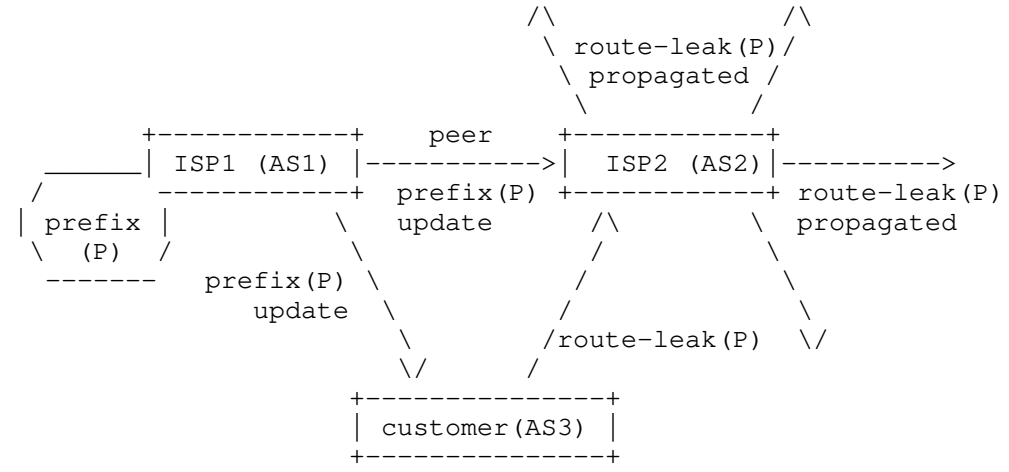


Figure 1: Illustration of the basic notion of a route leak.

The RLP information contained in the RLP community consists of one or two AS numbers (ASNs) and has the following semantics:

1. Down Only (DO) indication: ASN of the most recent RLP-aware AS in the path to assert that it sent the update to a customer or lateral peer;

2. Leak detected (L) indication: ASN of the first RLP-aware AS in the path to assert that it forwarded the route from a customer or lateral peer despite detecting a leak (to avoid unreachability).

If the RLP community is present in an update, it will always contain a single DO. However, L need not be always present. (Note: The bits designated to carry L may be always present along with a DO, except that a default value (all zeros) is carried in L when no AS in the current AS path needed to assert L.) Once an AS asserts L (Leak detected) by inserting its ASN value, it MUST not be changed subsequently as the update propagates. But the ASN value in DO (Down Only) is changeable along the AS path per its definition above.

Design assumption 1: Operators desire to avoid unreachability. So, a design assumption here is that in the absence of an alternative route, an AS may select and forward a route that is detected to be a leak. (Note: This is the reason Leak detected (L) indication is part of the design.)

Design assumption 2: An AS that is RLP-aware (i.e., implements the RLP solution in this document) MUST also implement an intra-AS solution for route leak avoidance in the local AS. The latter solution uses an intra-AS signaling mechanism (see [I-D.ietf-idr-bgp-open-policy], Section 3.7 of [RLP-Discussion]). By doing this, the AS locally prevents the leaking of routes learned from a transit provider or lateral peer to another transit provider or lateral peer. Why this is critical to the overall solution is made clear in slides 7 and 8 of [sriram2].

2.2.1. Format of the RLP Community

The format of the RLP community using a single Large Community is shown in Figure 2.

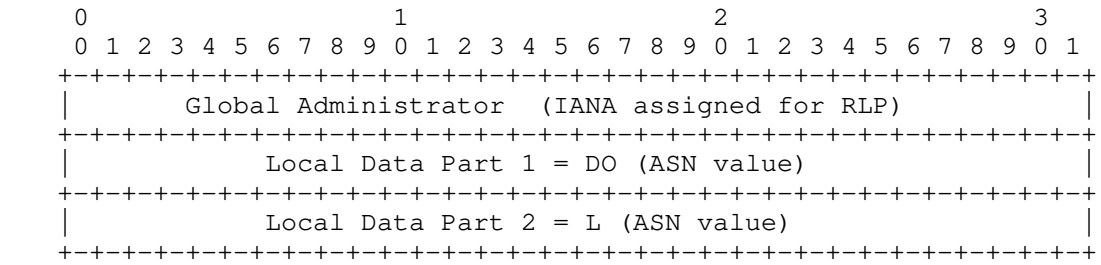


Figure 2: Format of the RLP Community using a Large Community [RFC8092].

2.3. Route Leak Detection Rules and the Ingress Router (Receiver) Actions

A received BGP update is determined to be a route leak if:

1. if L is present in the update;

2. else (L is absent), the update is received from a customer and DO is present;

3. else (L is absent), the update is received from a lateral peer and DO is present that is not the lateral peer's ASN.

Note: Here by "L is present" we mean that its value is not the default value (all zeros) but is a proper ASN. Effectively "L is absent" if its value is the default value.

In steps 2 and 3 above, the ingress router (receiver) MUST add L = local ANS. Doing this prior to the best path selection process is necessary. Also, if the route is selected as best path, then L is already set correctly before the egress router (sender) acts on it.

2.4. Route Selection Policy

Minimum Default Policy: Whenever there is a choice between a customer route and a provider route that are both detected to be leaks (L is present), then lower the LocalPref to X (TBD by operator) for each of them. Then shortest path criterion would typically make the customer route preferred. (Note: This would help mitigate any possibility of persistent oscillation; see slide #7 in [sriram1].)

Generalized Minimum Default Policy: Whenever there is a choice between multiple routes (customer/peer/provider) and each is detected to be a leak (L is present), then lower the LocalPref to X TBD by

operator) for each of them. Then apply shortest path criterion. (Note: Some network operators may find this inadequate; see scenarios #3 and #6 in slides #14 and #16, respectively, in [sriram2]. But they may locally modify their policy while respecting the basic principle.)

2.5. Egress Router (Sender) Actions

After best path selection has been performed, a sender MUST perform the following RLP-related actions on the update to be propagated:

1. When propagating a route originated by the local AS to a customer or lateral peer, add DO = local ASN;

2. Else, when propagating a route that already includes a DO (i.e., was received with a DO) to a customer or lateral peer, replace the DO value with the local ASN.

3. Pseudo Code

[Begin: receiver action for route leak detection]

{Comment: This precedes route selection policy.}

if received route includes L, then save the route in RIB-in as is;

else (L is absent), if route is received from a customer and DO is preset, then add L = local ASN;

else (L is absent), if route is received from a lateral peer and DO is present that is not the lateral peer's ASN, then add L = local ASN

{Comment: "Route does not include L" or "L is absent" only if L is either literally absent or has the default (all zeros) value.}

[End: receiver action for route leak detection]

[Begin: route selection policy]

for each route that includes L, lower the LocalPref to X (TBD); apply best path selection policy*

{*Comment: E.g., best path selection based on LocalPref first and then shortest path.}

[End: route selection policy]

[Begin: sender action]

{Comment: RLP (includes DO and L or just DO) is a *transitive* BGP community and should propagate globally.}

when propagating a route originated by local AS to a customer or lateral peer, add DO = local ASN;

when propagating a route that includes a DO (i.e., was received with a DO) to a customer or lateral peer, replace the DO value with the local ASN;

[End: sender action]

4. Security Considerations

With the use of BGP community, there is often a concern that the community propagates beyond its intended perimeter and causes harm [streibelt]. However, that concern does not apply to the RLP community because it is a transitive community that must propagate as far as the update goes.

The proposed Route-Leak Protection (RLP) information carried in the RLP community can benefit from cryptographic protection to prevent abuse by malicious actors in the AS path. In the future, if there is BGPsec deployment, the RLP information can be encoded in the Flags field in the Secure_Path Segment in BGPsec updates [RFC8205]. So, the cryptographic security mechanisms in BGPsec can also secure the RLP information. The reader is directed to the security considerations provided in [RFC8205].

5. IANA Considerations

IANA is requested to register RLP in the well-known Large Community [RFC8092] registry (need help to clarify this). IANA is requested to allocate a new Global Administrator ID for the RLP community (Large Community) (see Figure 2 in this document). Note that BGP Path Attribute value for Large Community is 32 (IANA allocated) [RFC8092].

6. References

6.1. Normative References

[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>.

6.2. Informative References

[draft-dickson-sidr-route-leak-solns] Dickson, B., "Route Leaks -- Proposed Solutions", IETF Internet Draft (expired), March 2012, <https://tools.ietf.org/html/draft-dickson-sidr-route-leak-solns-01>.

[I-D.ietf-idr-bgp-open-policy] Azimov, A., Bogomazov, E., Bush, R., Patel, K., and K. Sriram, "Route Leak Prevention using Roles in Update and Open messages", draft-ietf-idr-bgp-open-policy-05 (work in progress), February 2019.

[Luckie] Luckie, M., Huffaker, B., Dhamdhere, A., Giotsas, V., and kc. claffy, "AS Relationships, Customer Cones, and Validation", IMC 2013, October 2013, <http://www.caida.org/~amogh/papers/asrank-IMC13.pdf>.

[RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, DOI 10.17487/RFC6811, January 2013, <https://www.rfc-editor.org/info/rfc6811>.

[RFC7454] Durand, J., Pepelnjak, I., and G. Doering, "BGP Operations and Security", BCP 194, RFC 7454, DOI 10.17487/RFC7454, February 2015, <https://www.rfc-editor.org/info/rfc7454>.

[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, <https://www.rfc-editor.org/info/rfc7908>.

[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>.

[RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", RFC 8205, DOI 10.17487/RFC8205, September 2017, <https://www.rfc-editor.org/info/rfc8205>.

[RLP-Discussion]
Sriram (Ed.), K., "Design Discussion of Route Leaks Solution Methods", Work in Progress, draft-sriram-idr-route-leak-solution-discussion-00, July 2018, <https://tools.ietf.org/html/draft-sriram-idr-route-leak-solution-discussion-00>.

[sriram1] Sriram et al., K., "Route Leaks Solution Merger of RLP and eOTC Drafts", Presented at the IDR Working Group Meeting, IETF-102, Montreal, July 2018, <https://datatracker.ietf.org/meeting/102/materials/slides-102-idr-sessb-route-leaks-merged-solution-00>.

[sriram2] Sriram et al., K., "Solution for Route Leaks Using BGP Communities", Authors Team Discussion Slides, October 2018, <https://www.nist.gov/sites/default/files/documents/2018/10/22/rlp_using_bgp_community-v4.pdf>.

[streibelt]
Streibelt et al., F., "BGP Communities: Even more Worms in the Routing Can", ACM IMC, October 2018, <https://archive.psg.com//181101.imc-communities.pdf>.

Acknowledgements

The authors wish to thank John Scudder and Susan Hares for their review and comments.

Contributors

The following people made significant contributions to this document and should be considered co-authors:

Brian Dickson
Independent
Email: brian.peter.dickson@gmail.com

Doug Montgomery
USA National Institute of Standards and Technology
Email: dougm@nist.gov

Keyur Patel
Arrcus
Email: keyur@arrcus.com

Andrei Robachevsky
Internet Society
Email: robachevsky@isoc.org

Eugene Bogomazov
Qrator Labs
Email: eb@qrator.net

Randy Bush
Internet Initiative Japan
Email: randy@psg.com

Authors' Addresses

Kotikalapudi Sriram (editor)
USA National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899
United States of America

Email: ksriram@nist.gov

Alexander Azimov (editor)
Yandex
Moscow
Russia

Email: a.e.azimov@gmail.com

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: March 12, 2021

K. Talaulikar
C. Filsfils
K. Swamy
Cisco Systems
S. Zandi
G. Dawra
LinkedIn
M. Durrani
Equinix

September 8, 2020

BGP Link-State Extensions for BGP-only Fabric
draft-ketant-idr-bgp-ls-bgp-only-fabric-05

Abstract

BGP is used as the only routing protocol in some networks today. In such networks, it is useful to get a detailed view of the nodes and underlying links in the topology along with their attributes similar to one available when using link state routing protocols. Such a view of a BGP-only fabric enables use cases like traffic engineering and forwarding of services along paths other than the BGP best path selection.

This document defines extensions to the BGP Link-state address-family (BGP-LS) and the procedures for advertisement of the topology in a BGP-only network. It also describes a specific use-case for traffic engineering based on Segment Routing.

Table of Contents

1. Introduction

Network operators are going for a BGP-only routing protocol for certain networks like Massively Scaled Data Centers (MSDCs). [RFC7938] describes the requirement, design and operational aspects for use of BGP as the only routing protocol in MSDCs. The underlying link and topology information between BGP routers is hidden or abstracted in this design from the underlay routing for improving scalability and stability in a large scale network. On the flip side, there is no detailed topology view similar to one available in form of the Traffic Engineering (TE) Database (TED) when running link state routing protocols like OSPF [RFC2328] with extensions specified in [RFC3630].

BGP Link-State (BGP-LS)[RFC7752] enables advertisement of a link state topology via BGP that can be consumed by a controller or in general any software component to get a complete topology view of the network. BGP-LS extensions for advertisement of a BGP topology for the Egress Peer Engineering (EPE) use-case [I-D.ietf-spring-segment-routing-central-epe] are specified in [I-D.ietf-idr-bgpls-segment-routing-epe]. This document leverages the BGP-LS TLVs defined for BGP-LS EPE and other BGP-LS documents and specifies the procedures for advertising the underlying topology in a more generic BGP-only fabric use-case.

This document specifies the operations and procedures when using the design involving BGP use for hop-by-hop routing between directly connected network nodes (refer [RFC7938] for details).

1.1. 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.

2. BGP Routing in the Fabric

This document does not change base BGP routing protocol operations in the fabric that provides routing using the BGP best path selection process [RFC4271] .

The applicability of this specification is limited to those deployments where BGP is used as hop-by-hop routing protocol between directly connected nodes in the fabric. While a data-center design [RFC7938] is used as a reference, the topology advertisement and its

use for computation may also apply to other networks with BGP-only fabric or to BGP-only portions of a larger network topology.

BGP hop-by-hop routing can be setup using EBGP single-hop sessions over individual links between directly connected routers using their link addresses for peering as described in [RFC7938]. In such a design, the neighbors' link addresses may be provisioned for peering and the EBGP session operating directly over the link performs the monitoring of the neighbor on that link. A variation of this design would be that the EBGP session is setup between directly connected routers using their loopback sessions. The mechanisms for discovery of the neighbor's link addresses and their monitoring on a per link basis are outside the scope of this document. [I-D.xu-idr-neighbor-autodiscovery] describes one such mechanism and the same may be also realized by other means.

Though this document uses the EBGP based design as a reference, it does not preclude other alternate designs using IBGP.

3. Topology Collection Mechanism

BGP-LS [RFC7752] has been defined to allow BGP to convey topology information in the form of Link-State objects - node, link and prefix. The properties of each of these objects are encoded using BGP-LS Attribute TLVs. Applications need a topological view and visibility even for networks where BGP is the only routing protocol. In such networks, each BGP router advertises its local information which includes its node, links and prefix attributes via BGP-LS.

Figure 1 describes a typical deployment scenario. Every BGP router in the network is enabled for BGP-LS and forms BGP-LS sessions with one or more centralized BGP-LS speakers over which it sends its local topology information. Each BGP router MAY also receive the topology information from all other BGP routers via these centralized BGP-LS speakers. This way, any BGP router (as also the centralized BGP-LS speakers) MAY obtain aggregated Link-State information for the entire BGP network. An external component (e.g. a controller) can obtain this information from the centralized BGP-LS speakers or directly by doing BGP-LS peering to the BGP routers. An internal software component on any of the BGP routers (e.g. TE module) can also receive the entire BGP network topology information from its local BGP process.

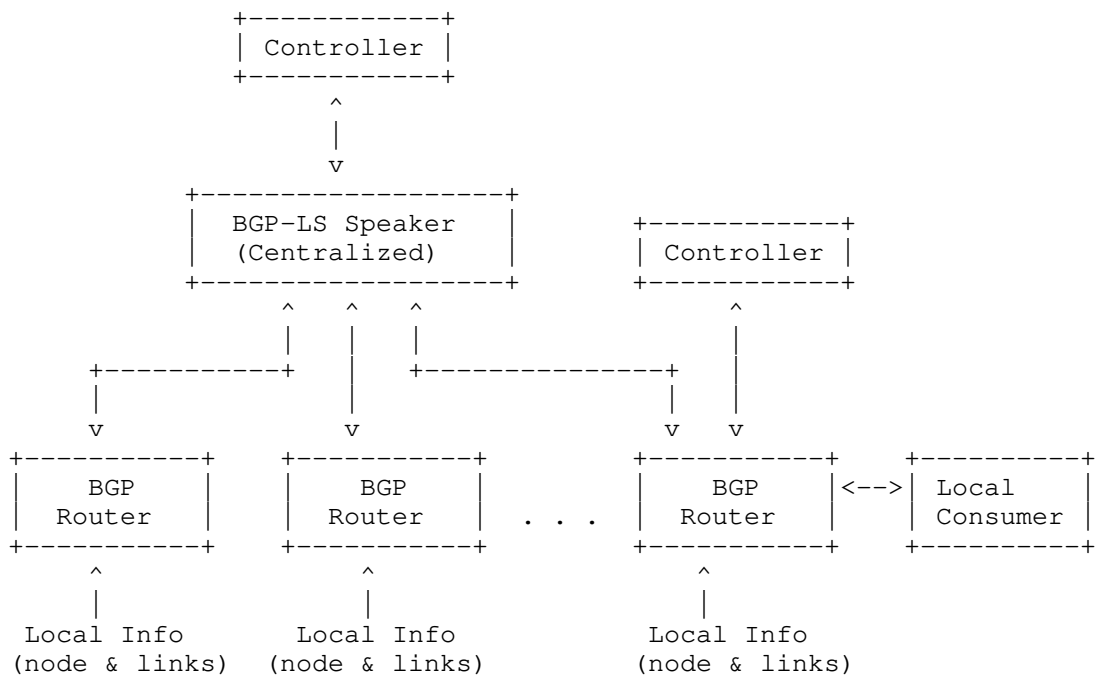


Figure 1: Link State info collection

3.1. Peering Models

The peering model described above relies on the base BGP IPv4 or IPv6 routing underlay (e.g. as described in [RFC7938]) or any other mechanism for reachability for the BGP-LS session establishment with the centralized BGP speakers. A variation of this model would be to setup reachability to the centralized BGP speakers (or controller) over the out of band management network, where available, and for each BGP router in the fabric use the same for the BGP-LS session establishment with the centralized BGP speakers. This variation removes the dependency between the topology learning via BGP-LS from the base best effort reachability over the BGP routing in the fabric.

Another alternate design would be to enable BGP-LS as well on the hop by hop EBGP sessions in the underlay as described in [RFC7938]. This approach results in the topology information being flooded via BGP-LS hop-by-hop along the BGP routers in the network. Other peering designs for BGP-LS sessions may also be possible and they are not precluded by this document.

4. Advertising BGP-only Network Topology

This section specifies the BGP-LS TLVs and sub-TLVs and their use for advertising the topology of a BGP-only network in the form of BGP-LS Node, Link and Prefix NLRIs.

BGP-LS [RFC7752] defines the BGP-LS NLRI types (i.e. Node NLRI, Link NLRI and Prefix NLRI) along with their corresponding BGP-LS Attribute (i.e. Node Attribute, Link Attribute or Prefix Attribute) and the TLVs that map to the respective NLRI and Attribute for each type.

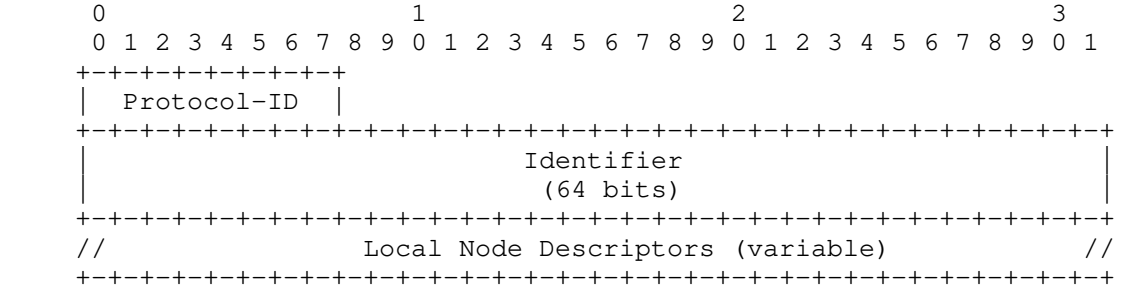
[I-D.ietf-idr-bgpls-segment-routing-epe] specifies the BGP Protocol ID to be used for signaling BGP EPE information and the same is used for advertising of BGP topology.

[I-D.ietf-idr-te-lsp-distribution] defines the BGP-LS NLRI that can be used to advertise the RSVP-TE or Segment Routing (SR) policies instantiated on a BGP Router head-end along with their corresponding BGP-LS Attribute TLVs to advertise their properties and state.

The following sub-sections specify the use of these encodings by a router running BGP protocol.

4.1. Node Advertisements

[RFC7752] defines Node NLRI Type and the Node Descriptor TLVs as follows:



[I-D.ietf-idr-bgpls-segment-routing-epe] introduces additional Node Descriptor TLVs for BGP protocol that are required to be used.

The following Node Descriptors TLVs MUST appear in the Node NLRI as Local Node Descriptors:

- BGP Router-ID, which contains the BGP Identifier of the originating BGP router

- Autonomous System Number, which contains the advertising router ASN.

The BGP-LS Attribute associated with the Node NLRI MAY include the following TLVs that are defined in respective documents to signal the router properties and capabilities (Section 5.1 defines the procedures for their advertisements):

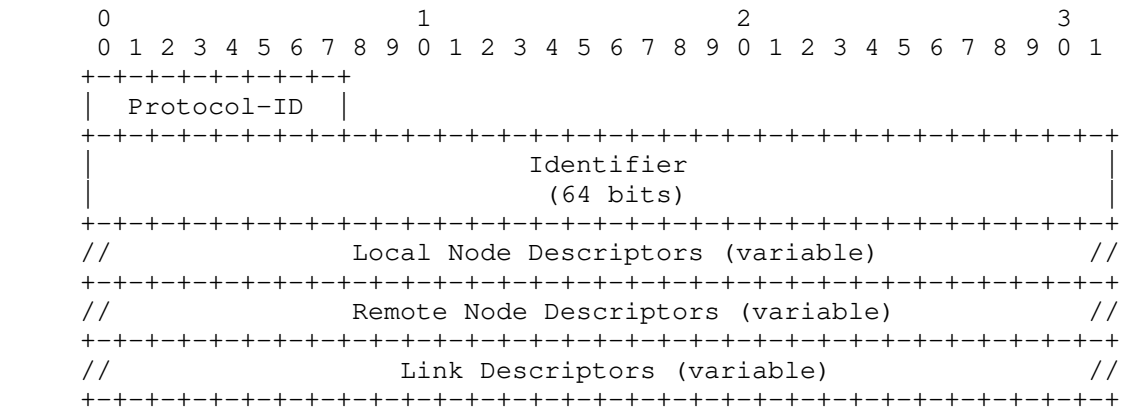
TLV Code Point	Description	Reference Document
1026	Node Name	[RFC7752]
1028	IPv4 TE Router-ID	[RFC7752]
1029	IPv6 TE Router-ID	[RFC7752]
1161	SID/Label	[I-D.ietf-idr-bgp-ls-segment-routing-ext]
1034	SRGB & Capabilities	[I-D.ietf-idr-bgp-ls-segment-routing-ext]
1035	SR Algorithm	[I-D.ietf-idr-bgp-ls-segment-routing-ext]
1036	SR Local Block	[I-D.ietf-idr-bgp-ls-segment-routing-ext]
266	Node MSD	[RFC8814]
TBD	Flex Algorithm Definition	[I-D.ietf-idr-bgp-ls-flex-algo]

Table 1: Node Attribute TLVs

The above list of TLVs is not exhaustive but indicative as of the time of writing of this document.

4.2. Link Advertisements

[RFC7752] defines Link NLRI Type and its Node and Link Descriptor TLVs as follows:



The following Node Descriptors TLVs MUST appear in the Link NLRI as Local Node Descriptors:

- BGP Router-ID, which contains the BGP Identifier of the originating BGP router
- Autonomous System Number, which contains the advertising router ASN.

The following Node Descriptors TLVs MUST appear in the Link NLRI as Remote Node Descriptors:

- BGP Router-ID, which contains the BGP Identifier of the peer BGP router
- Autonomous System Number, which contains the peer ASN.

The following Link Descriptors TLVs MUST appear in the Link NLRI as Link Descriptors:

- Link Local/Remote Identifiers containing the 4-octet Link Local Identifier followed by the 4-octet Link Remote Identifier. The value 0 MUST be used for the Link Remote Identifier when the value is unknown.

In addition, the following Link Descriptors TLVs SHOULD appear in the Link NLRI as Link Descriptors based on the address family used for setting up the BGP Peering or the addresses configured on the links:

- IPv4 Interface Address contains the address of the local interface through which the BGP session is established using IPv4 address.

- IPv6 Interface Address contains the address of the local interface through which the BGP session is established using IPv6 address.

- IPv4 Neighbor Address contains the IPv4 address of the peer interface used by the BGP session establishment using IPv4 address.

- IPv6 Neighbor Address contains the IPv6 address of the peer interface used by the BGP session establishment using IPv6 address.

The BGP-LS Attribute associated with the Link NLRI MAY include the following TLVs that are defined in respective documents to signal the router's local links' properties and capabilities (Section 5.2 defines the procedures for their advertisements) :

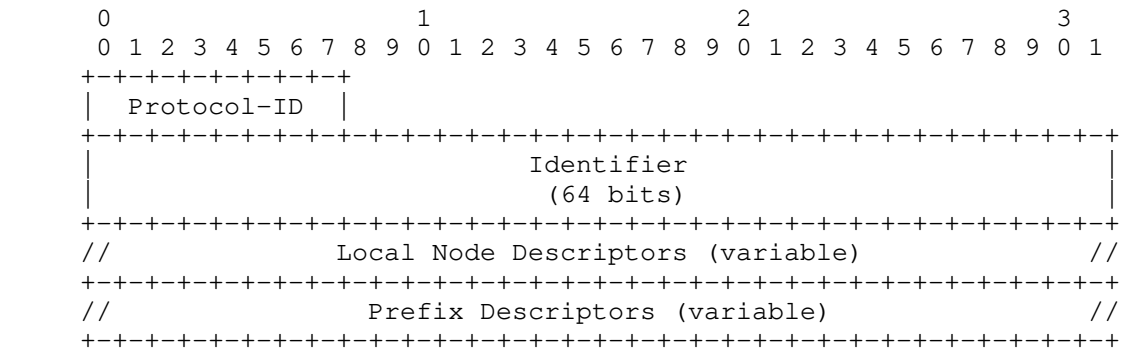
TLV Code Point	Description	Reference Document
1088	Administrative group (color)	[RFC7752]
1173	Extended Administrative group (color)	[I-D.ietf-idr-eag-distribution]
1089	Maximum link bandwidth	[RFC7752]
1092	TE Default Metric	[RFC7752]
1096	SRLG	[RFC7752]
1098	Link Name	[RFC7752]
267	Link MSD	[RFC8814]
1172	L2 Bundle Member	[I-D.ietf-idr-bgp-ls-segment-routing-ext]
1104	Unidirectional link delay	[RFC8571]
1105	Min/Max Unidirectional link delay	[RFC8571]
1106	Min/Max Unidirectional link delay	[RFC8571]
1107	Unidirectional packet loss	[RFC8571]
1101	EPE Peer Node SID	[I-D.ietf-idr-bgpls-segment-routing-epe]
1102	EPE Peer Adj SID	[I-D.ietf-idr-bgpls-segment-routing-epe]
1103	EPE Peer Set SID	[I-D.ietf-idr-bgpls-segment-routing-epe]

Table 2: Link Attribute TLVs

The above list of TLVs is not exhaustive but indicative as of the time of writing of this document.

4.3. Prefix Advertisements

[RFC7752] defines Prefix NLRI Type and its Node and Prefix Descriptor TLVs as follows:

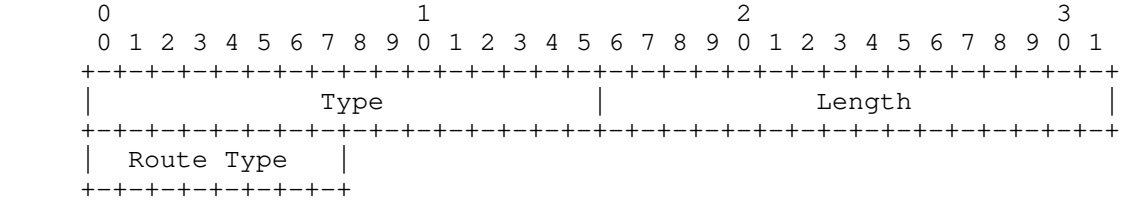


The following Node Descriptors TLVs MUST appear in the Prefix NLRI as Local Node Descriptors:

- BGP Router-ID, which contains the BGP Identifier of the originating BGP router
- Autonomous System Number, which contains the advertising router ASN.

The Prefix Descriptor MUST contain the IP Reachability information TLV to identify the prefix.

This document defines a new BGP Route Type TLV that MUST be included in the Prefix Descriptor when the BGP node advertises the Prefix NLRI. The format of this TLV is as follows:



Where:

Type: 2 octet field with value TBD, see Section 7.

Length: 2 octet field with value set to 1.

Route Type: one octet with the following values defined:

Value	Type	Description
1	Local	Local interface prefix e.g. Loopback
2	Attached	Directly attached node's prefix e.g host
3	External BGP	Prefix learnt via EBGP
4	Internal BGP	Prefix learnt via IBGP
5	Redistributed	Prefix redistributed into BGP

Figure 2: BGP Route Types

The BGP-LS Attribute associated with the Prefix NLRI MAY include the following TLVs that are defined in respective documents to signal the router's own prefix properties and capabilities (Section 5.3 defines the procedures for their advertisements):

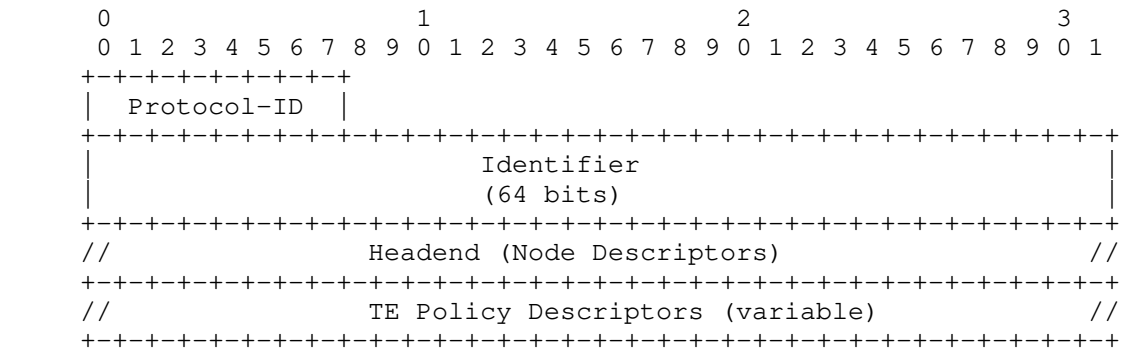
TLV Code Point	Description	Reference Document
1155	Prefix Metric	[RFC7752]
1158	Prefix SID	[I-D.ietf-idr-bgp-ls-segment-routing-ext]

Table 3: Prefix Attribute TLVs

The above list of TLVs is not exhaustive but indicative as of the time of writing of this document.

4.4. TE Policy Advertisements

[I-D.ietf-idr-te-lsp-distribution] defines TE Policy NLRI Type and its Headend Node and TE Policy Descriptor TLVs as follows:



The Node Descriptors TLVs are the same as specified in Section 4.1. The semantics for the TE Policy Descriptor TLVs and the TLVs associated with the BGP-LS Attribute are used as specified in [I-D.ietf-idr-te-lsp-distribution].

5. Procedures

In a network where BGP is the only routing protocol, the BGP-LS session is used to advertise the necessary information about the local node properties, its local links' properties and where necessary the prefix's owned by the node. TE Policies, that are instantiated on the local node (i.e. when it is the head-end for the policy), along with their properties are also advertised via the BGP-LS session. This information, once collected across all BGP routers in the network, provides a complete topology view of the network. Many of these attributes are not part of the base BGP protocol operations and are either configured or provided by other components on the router. BGP-LS performs the role of collecting this information and propagating it across the BGP network.

The following sections describe the procedures for the propagation of the BGP-LS NLRIs on a BGP router into the BGP-LS session. These procedures for propagation of BGP topology information via BGP-LS SHOULD be applied only in deployments and use-cases where necessary and SHOULD NOT be applied in every BGP deployment when BGP-LS is enabled. Implementations MAY provide a configuration option to enable these procedures in required deployments.

5.1. Advertisement of Router's Node Attributes

Advertisement of the Node NLRI via BGP-LS by each BGP router in a BGP-only network enables the discovery of all the router nodes in the topology. The Node NLRI MUST be generated by a BGP router only for

itself and even when there are no attributes to be advertised along with it.

The Node attributes defined currently related to Segment Routing (SR) [RFC8402] have been described in Table 1 and are to be advertised when SR is enabled. This includes:

- All SR enabled routers support the default SR algorithm 0 and MUST advertise it in the SR Algorithm TLV. Other algorithms (including Flexible Algorithm [I-D.ietf-lsr-flex-algo]) SHOULD be advertised when supported.

- The Segment Routing Global Block (SRGB) provisioned on the router which is used by BGP Prefix SIDs [RFC8669] and other SR control plane protocols on the router MUST be advertised. The value for Flags field in the TLV is not defined for BGP protocol and MUST be set to 0 by the originator and ignored by receivers.

- The Segment Routing Local Block (SRLB) provisioned on the router which MAY be used by BGP EPE SIDs [I-D.ietf-idr-bgpls-segment-routing-epe] SHOULD be advertised. The value for Flags field in the TLV is not defined for BGP protocol and MUST be set to 0 by the originator and ignored by receivers.

- The Node level MSD provides the Node's capabilities for SR SID operations and SHOULD be advertised.

- When the router supports SR Flexible Algorithms and is provisioned with the Flexible Algorithm Definition (FAD), then it MUST advertise the same.

The Node Name Attribute SHOULD be advertised when available.

This document introduces some of the TE concepts into BGP-only networks. Provisioning of TE Router-ID with a unique address normally associated with a loopback interface on the router enables TE use-cases for both IPv4 and IPv6 SHOULD be supported. The BGP Router-ID along with the ASN also provides the capability for uniquely identifying a BGP router in the network.

Other Node Attributes applicable to a BGP Router may also be included and this document does not describe the exhaustive list.

5.2. Advertisement of Router's Local Links Attributes

Each BGP router in a BGP-only network also advertises its local links using the Link NLRIs thru BGP-LS. The Link NLRI for a given link between two BGP routers is advertised as uni-directional logical "half-link" and its link descriptors allow the correlation between the two NLRIs "half-links" originated by the peering routers to describe the bi-directional logical link and its attributes on both routers.

The discovery of all the links and their local and remote identifiers in a BGP-only network relies on the design that uses EBGP sessions over each interconnecting link using the link IP addresses (refer [RFC7938]). In this case, a Link NLRI MUST be generated by a BGP router for each of its local link regardless of whether it has any link attributes to be advertised for it.

When doing EBGP multi-hop sessions between directly connected BGP routers, the underlying link information would need to learn by some discovery protocol or provisioning entity. The mechanisms to learn the underlying link information for BGP-LS advertisements are outside the scope of this document. However, to provide a true link topology picture, the advertisement of underlying links is RECOMMENDED for most use-cases instead of a single EBGP peering representation of a link between the routers using their loopback addresses.

The Link NLRI represents an adjacency between BGP routers and its association with the underlying Layer 3 link. When the underlying Layer 3 link or the BGP session on top of it goes down, the Link NLRI MUST be withdrawn by the BGP router. The monitoring of links, detecting of their failures and notification to BGP may be performed using mechanisms like BFD. This enables faster detection of failures and verification of the underlying links.

Advertisement of the Link NLRIs via BGP-LS by each BGP router in a BGP-only network enables the discovery of all the active links in the topology.

TE attributes for links have been traditionally associated with Link State Routing protocols. However, with the ability to discover the link topology via BGP-LS as specified in this document, the TE attributes and their principles can also be applied to a network running BGP alone. The TE attributes for a link have been described in Table 2 and MAY be advertised when TE use-cases are enabled. This includes:

- The maximum bandwidth of a link is its protocol independent attribute and SHOULD be advertised.

- TE concepts of Administrative Groups (also known as affinities) and Shared Risk Link Groups (SRLGs) MAY be provisioned locally on links and then MUST be advertised.

- The BGP base protocol does not operate with link metrics, however, a TE metric concept can be introduced in a BGP only network as well for TE use-cases. Implementations MAY provide the ability to provision TE metric value for a link for BGP use including a different default value for it. The TE metric attribute SHOULD be advertised for each link when configured and its default value is taken as 100. When not advertised for a link, implementations who intend to use the TE metric MUST assume the value to be 100.

- The delay and loss TE metrics for links are measured via MPLS Performance Monitoring [RFC6374] and their measurement mechanism over a link are independent of the routing protocol. The same mechanism MAY be enabled in BGP-only networks and their values advertised via BGP-LS.

The Link attributes defined currently related to the Segment Routing feature BGP EPE [I-D.ietf-idr-bgpls-segment-routing-epe] have been described in Table 2 and are to be advertised when SR use-cases are enabled. This includes:

- The BGP Peering SIDs provide a functionality similar to Adjacency-SID (refer [RFC8402]) in BGP-only networks. Implementations SHOULD allocate the BGP Peer-Adjacency SID for all its links and the BGP Peer-Node SID for all its peer routers. Implementations MAY allocate the BGP Peer-Set SID based on local configuration.

- The Link level MSD provides the per link capabilities for SR SID operations and SHOULD be advertised when the router links have differing capabilities.

The use of Layer 3 bundle links which comprise of multiple layer 2 member links are often used in BGP networks. When BGP session is configured over such a layer 3 link, the link attributes of the underlying layer 2 links MAY be advertised individually using the L2 Bundle Member TLV. The applicable attributes for the L2 links are described in [I-D.ietf-idr-bgp-ls-segment-routing-ext].

The Link Name Attribute MAY be advertised when available.

Other Link Attributes applicable to a BGP Router may also be included and this document does not describe the exhaustive list.

5.3. Advertisement of Router's Prefix Attributes

Advertisement of the Prefix NLRI via BGP-LS may be required only in specific use-cases. Since the base BGP protocol along with its extensions already signals Prefix reachability via different NLRIs, there is no necessity to duplicate the information via BGP-LS session. However, for specific use-cases related to SR Traffic Engineering (SR-TE), it is required for each router to advertise it's Prefix SID(s) (refer [RFC8402]) that can be used to direct traffic via specific BGP routers. Advertising such BGP Prefix SID for every BGP router provides this key attribute via BGP-LS and avoids the requirement for the consumer of the topology information (e.g. a controller or local TE process) to tap into other BGP NLRI information.

Advertisement of the Prefix NLRI via BGP-LS MUST be done for its locally configured prefixes (e.g. its loopback interface address) and when BGP is advertising the BGP Prefix SID ([RFC8669]) for it. The advertisement of the Prefix NLRI via BGP-LS for other prefixes learnt by the router MAY be done based on the specific use-case requirement and the BGP Route Type as described in Figure 2 indicates the type of route being advertised.

The Prefix attributes defined currently related to SR [RFC8402] have been described in Table 3 and MAY be advertised when SR is enabled. This includes:

- The Prefix SID TLV is included with the SID advertised as the index to be consistent with the Label-Index TLV of BGP Prefix SID attribute. The default algorithm is MUST be set to 0 by the originator except in the case where a local prefix is associated with a specific SR Algorithm. The flags are defined as the most significant 8 bits of the 16 bit field defined for Label-Index TLV in [RFC8669].

- For certain SR-TE uses, the Prefix Metric value MAY be included and it is set based on the SR-TE computation based on the link-state topology learnt via BGP-LS.

Other Prefix Attributes applicable may also be included and this document does not describe the exhaustive list.

5.4. Advertisement of Router's TE Policy Attributes

TE Policies that are setup using RSVP-TE or SR-TE mechanisms MAY be instantiated on a BGP router. One use-case that results in such SR Policy instantiation on a BGP router is described later in this document in Section 6.2. Advertising such TE Policies instantiated

for every BGP router as head-end via BGP-LS provides the consumer of the topology information (e.g. a controller or local TE process) a policy view of the BGP fabric as well.

The procedures for advertisement of the TE Policy NLRI via BGP-LS MUST be done only for its locally instantiated TE Policies and as specified in [I-D.ietf-idr-te-lsp-distribution]). Implementation MAY provide configuration options to control the specific set of TE Policies that are to be advertised from the local node.

6. Usage of BGP Topology

This section describes some of the use-cases for the building of the BGP topology information as specified in this document and leveraging it for enabling new functionality.

6.1. Topology View for Monitoring

The BGP-LS advertisement of the BGP topology as specified in this document provides a live topology view of the BGP network for an application or controller that is monitoring the network. The topology view is from the BGP protocol perspective and includes the underlying links as well that aids in network monitoring as well as diagnostics use-cases. BGP-LS is the de-facto protocol for northbound propagation of network topology related information for most IGP networks and extending this capability for BGP-only networks allows existing controllers and applications to consume the information with some incremental BGP protocol awareness.

6.2. SR-TE in BGP Networks

The SR-TE use-case for BGP builds on top of functionality specified in [RFC8669] and also described in [RFC8670].The BGP SR Prefix SID signaled, provides the basic connectivity between all BGP routers using their loopback addresses. This provides the basic best-effort paths in the network using the base BGP decision process that is unchanged. BGP and other overlay routes and services recurse on top of these loopback addresses of the egress nodes and the forwarding is done via the BGP SR Prefix SID labels in the underlay. While this version of the document focuses on the examples with MPLS dataplane instantiation for SR, the same is applicable for the IPv6 dataplane instantiation (SRv6) as well.

SR-TE for BGP provides underlay paths through the network for the overlay routes and services with specific SLA requirements and use-cases like path disjointness, low latency paths, inclusion or exclusion and other TE considerations.

[I-D.ietf-spring-segment-routing-policy] specifies the SR-TE architecture and the SR Policy construct. [I-D.ietf-idr-segment-routing-te-policy] describes the extensions to BGP for signaling of SR Policies from a controller to the SR-TE headend BGP router. BGP-LS has been extended to allow signaling of the SR Policies from SR-TE head-end to controllers via [I-D.ietf-idr-te-lsp-distribution] which allows the controllers to learn the state of SR Policies instantiated on routers in the network. This document completes the missing piece that is related to getting the BGP topology information from all the routers to a controller as well the local SRTE process on each router for their path computation requirements.

The signaling of SR Polices from controller to SR-TE headend and reporting of the state back to the controller can also be done using PCEP ([RFC8664], [RFC8281], [RFC8231]). However, the BGP topology learning via BGP-LS which is specified in this document is also required for the deployments that uses PCEP in the BGP-only network.

The topology collected via BGP-LS in a BGP-only fabric in a Segment Routing deployment comprise of:

- The properties of every BGP router node and the Prefix SIDs to reach that node.

- The properties of all the links between the BGP routers and the Peer-Adjacency-SIDs (and other EPE SIDs) corresponding to them that allow directing traffic over specific links and/or to specific neighbors.

- The properties and state of the SR Policies instantiatied on each of the BGP routers along with their end points, their properties and most importantly the Binding SID to steer traffic into the SR Policies.

This topology information allows a computation node to build SR Policies for services over the BGP fabric for a given traffic engineering objective at any given node.

The topology of the BGP fabric is advertised to a centralized controller or application for use-cases that need a centralized computation of SR Policy which can then be signaled to the SR-TE head-end node via PCEP or BGP-SRTE. The topology may also be distributed to any node in the BGP fabric to be used by its local SR-TE process to perform path computation for its own SR Policies for use-cases that are addressed by local computation.

A high level summary of the key topology information advertised via BGP-LS by BGP routers can be used for TE computations as follows

- The BGP SR Prefix SIDs and the BGP EPE Peering Adjacency SIDs provide the equivalent of the IGP Prefix and Adjacency SIDs and can be used to direct traffic to a specific BGP router and over a specific BGP peer session or link respectively. Traffic for the BGP SR Prefix SIDs follow the path computed by the BGP decision process.

- The TE metric can be used to tailor the choice of specific paths in the network for SR-TE.

- The TE administrative group (also known as affinities) and SRLG attributes can be configured over links to enable computation of paths with inclusion and exclusion of specific links or paths that are mutually disjoint.

- The enabling of link delay and loss measurements and their advertisements can help monitoring the link quality and carve out paths based on latency and other SLA requirements.

- The signaling of the Node and Link MSD allows controllers to instantiate SR Policies based on the capability of the routers.

This section attempts to highlight and describe at a high level some of the possible SR-TE solutions and use-cases in a BGP-only network. The actual SR-TE computation and algorithms are outside the scope of this document.

7. IANA Considerations

IANA maintains a registry called "Border Gateway Protocol - Link State (BGP-LS) Parameters" with a sub-registry called "Node Anchor, Link Descriptor and Link Attribute TLVs".

The following TLV codepoints are suggested (to be assigned by IANA):

TLV Code Point	Description	Value defined in
TBD	BGP Route Type TLV	this document

8. Manageability Considerations

This section is structured as recommended in [RFC5706].

8.1. Operational Considerations

8.1.1. Operations

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

9. 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.

10. Acknowledgements

The authors would like to thank Bruno Decraene for his review and comments on this document.

11. References

11.1. Normative References

[I-D.ietf-idr-bgp-ls-flex-algo]
Talaulikar, K., Psenak, P., Zandi, S., and G. Dawra, "Flexible Algorithm Definition Advertisement with BGP Link-State", draft-ietf-idr-bgp-ls-flex-algo-04 (work in progress), July 2020.

[I-D.ietf-idr-bgp-ls-segment-routing-ext]
Previdi, S., Talaulikar, K., Filsfils, C., Gredler, H., and M. Chen, "BGP Link-State extensions for Segment Routing", draft-ietf-idr-bgp-ls-segment-routing-ext-16 (work in progress), June 2019.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Talaulikar, K., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", draft-ietf-idr-bgpls-segment-routing-epe-19 (work in progress), May 2019.

[I-D.ietf-idr-eag-distribution]
Wang, Z., WU, Q., Tantsura, J., and K. Talaulikar, "Distribution of Traffic Engineering Extended Admin Groups using BGP-LS", draft-ietf-idr-eag-distribution-12 (work in progress), May 2020.

[I-D.ietf-idr-te-lsp-distribution]
Previdi, S., Talaulikar, K., Dong, J., Chen, M., Gredler, H., and J. Tantsura, "Distribution of Traffic Engineering (TE) Policies and State using BGP-LS", draft-ietf-idr-te-lsp-distribution-13 (work in progress), April 2020.

[I-D.ietf-lsr-flex-algo]
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-algo-10 (work in progress), August 2020.

[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>.

[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>.

[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>.

[RFC8571] Ginsberg, L., Ed., Previdi, S., Wu, Q., Tantsura, J., and C. Filsfils, "BGP - Link State (BGP-LS) Advertisement of IGP Traffic Engineering Performance Metric Extensions", RFC 8571, DOI 10.17487/RFC8571, March 2019, <https://www.rfc-editor.org/info/rfc8571>.

[RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <https://www.rfc-editor.org/info/rfc8669>.

[RFC8814] Tantsura, J., Chunduri, U., Talaulikar, K., Mirsky, G., and N. Triantafillis, "Signaling Maximum SID Depth (MSD) Using the Border Gateway Protocol - Link State", RFC 8814, DOI 10.17487/RFC8814, August 2020, <https://www.rfc-editor.org/info/rfc8814>.

11.2. Informative References

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Rosen, E., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-09 (work in progress), May 2020.

[I-D.ietf-spring-segment-routing-central-epe]
Filsfils, C., Previdi, S., Dawra, G., Aries, E., and D. Afanasiev, "Segment Routing Centralized BGP Egress Peer Engineering", draft-ietf-spring-segment-routing-central-epe-10 (work in progress), December 2017.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-08 (work in progress), July 2020.

[I-D.xu-idr-neighbor-autodiscovery]
Xu, X., Talaulikar, K., Bi, K., Tantsura, J., and N. Triantafillis, "BGP Neighbor Discovery", draft-xu-idr-neighbor-autodiscovery-12 (work in progress), November 2019.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <https://www.rfc-editor.org/info/rfc3630>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://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, <https://www.rfc-editor.org/info/rfc5706>.

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, DOI 10.17487/RFC6374, September 2011, <https://www.rfc-editor.org/info/rfc6374>.

[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, <https://www.rfc-editor.org/info/rfc6952>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <https://www.rfc-editor.org/info/rfc8231>.

[RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <https://www.rfc-editor.org/info/rfc8281>.

[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>.

[RFC8664] Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Extensions for Segment Routing", RFC 8664, DOI 10.17487/RFC8664, December 2019, <https://www.rfc-editor.org/info/rfc8664>.

[RFC8670] Filsfils, C., Ed., Previdi, S., Dawra, G., Aries, E., and P. Lapukhov, "BGP Prefix Segment in Large-Scale Data Centers", RFC 8670, DOI 10.17487/RFC8670, December 2019, <https://www.rfc-editor.org/info/rfc8670>.

Authors' Addresses

Ketan Talaulikar
Cisco Systems
Pune 411057
India

Email: ketant@cisco.com

Clarence Filsfils
Cisco Systems
Brussels
Belgium

Email: cfilsfil@cisco.com

Krishna Swamy
Cisco Systems
San Jose
USA

Email: kriswamy@cisco.com

Shawn Zandi
LinkedIn
USA

Email: szandi@linkedin.com

Gaurav Dawra
LinkedIn
USA

Email: gdawra.ietf@gmail.com

Muhammad Durrani
Equinix
USA

Email: mdurrani@equinix.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: February 15, 2019

A. Wang
China Telecom
H. Chen
Huawei Technologies

August 14, 2018

BGP-LS Extension for Inter-AS Topology Retrieval
draft-wang-idr-bgpls-inter-as-topology-ext-02

Abstract

This document describes the process to build BGP-LS key parameters in Native IP multi-domain scenario and defines some new inter-AS TE related TLVs for BGP-LS to let SDN controller retrieve the network topology automatically under various environments.

Such process and extension can expand the usage of BGP-LS protocol to multi- domain, enable the network operator to collect the connection relationship between different AS domains and then calculate the overall network topology automatically based on the information provided by BGP-LS protocol.

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 February 15, 2019.

Copyright Notice

Table of Contents

1. Introduction

BGP-LS [RFC7752] describes the methodology that using BGP protocol to transfer the Link-State information. Such method can enable SDN controller to collect the underlay network topology automatically, but normally it can only get the information within one IGP domain. If the operator has more than one IGP domain, and these domains interconnect with each other, there is no general TLV within current BGP- LS to transfer the interconnect information.

Draft [I-D.ietf-idr-bgpls-segment-routing-epe] defines some extensions for exporting BGP peering node topology information (including its peers, interfaces and peering ASs) in a way that is exploitable in order to compute efficient BGP Peering Engineering policies and strategies. Such information can also be used to calculate the interconnection topology among different IGP domains, but it requires the border routers to run BGP-LS protocol to collect this information and report them to the PCE/SDN controller, which restricts the deployment flexibility of BGP-LS protocol.

This draft analysizes the situations that the PCE/SDN controller needs to get about the inter-connected information between different AS domains, defines new TLVs to extend the BGP-LS protocol to

transfer the key information related to the interconnect TE topology. After that, the SDN controller can then deduce the multi-domain topology automatically based on the information from BGP-LS protocol.

2. Conventions used in this document

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. Inter-AS Domain Scenarios.

Fig.1 illustrates the multi-domain scenarios that this draft discussed. Normally, SDN Controller can get the topology of IGP A and IGP B individually via the BGP-LS protocol, but it can’t get the topology connection information between these two IGP domains because there is generally no IGP protocol run on the connected links.

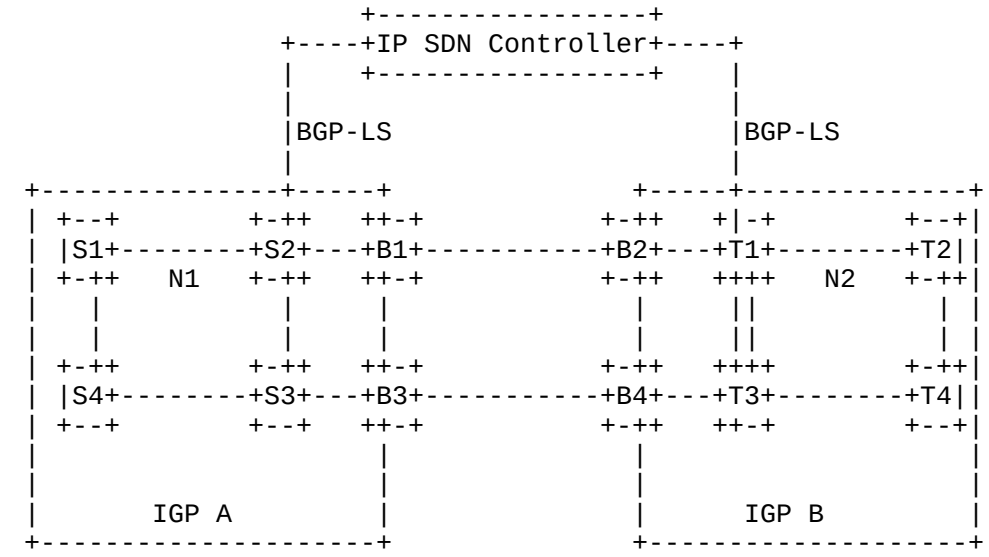


Figure 1: Inter-AS Domain Scenarios

3.1. IS-IS/OSPF Inter-AS Native IP Scenario

When the IGP A or IGP B runs native IS-IS/OSPF protocol, the operator often redistributes the IPv4/IPv6 prefixes of interconnect links into IS-IS/OSPF protocol to ensure the inter-domain connectivity.

If the IGP runs IS-IS protocol, the redistributed link information will be carried in IP External Reachability Information TLV within

the Level 2 PDU type that defined in [RFC1195], every router within the IGP domain can deduce the redistributed router from the IS-IS LSDB.

If the IGP runs OSPF protocol,[RFC2328]defines the type 5 external LSA to transfer the external IPv4 routes; [I-D.ietf-ospf-ospfv3-lsa-extend] defines the "External-Prefix TLV" to transfer the external IPv6 routes; these LSAs have also the advertising router information that initiates the redistribute activity. Every router within IGP domain can also deduce the redistributed router from the OSPF LSDB.

For prefix information that associated with each router, BGP-LS [RFC7752] defines the Prefix NLRI which is illustrated below:

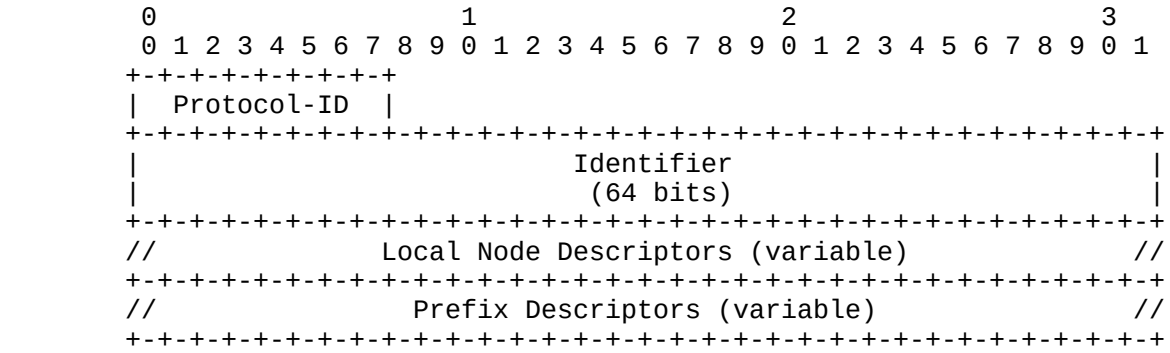


Figure 2: The IPv4/IPv6 Topology Prefix NLRI Format

For these redistributed inter-domain links, their prefix information should be included in the "Prefix Descriptor", and the associated redistributed router information should be included in the "Local Node Descriptors".

When such information is reported via the BGP-LS protocol, the PCE/ SDN controller can construct the underlay inter-domain topology according to procedure described in section 5

3.2. IS-IS/OSPF Inter-AS TE Scenario

[RFC5316] and [RFC5392] define the IS-IS and OSPF extensions respectively to deal with the requirements for inter-AS traffic engineering. They define some new sub-TLVs(Remote AS Number、IPv4 Remote ASBR ID、IPv6 Remote ASBR ID) which are associated with the inter-AS TE link TLVs to report the TE topology between different domains.

These TLVs are flooded within the IGP domain automatically. If the PCE/SDN controller can know these information via one of the interior router that runs BGP-LS protocol, the PCE/SDN controller can rebuild the inter-AS TE topology correctly.

4. Inter-AS TE related TLVs

This draft proposes to add three new TLVs that is included within the inter-AS TE link NLRI to transfer the information via BGP-LS, which are required to build the inter-AS related topology by the PCE/SDN controller.

The following Link Descriptor TLVs are added into the Link NLRI in BGP-LS protocol :

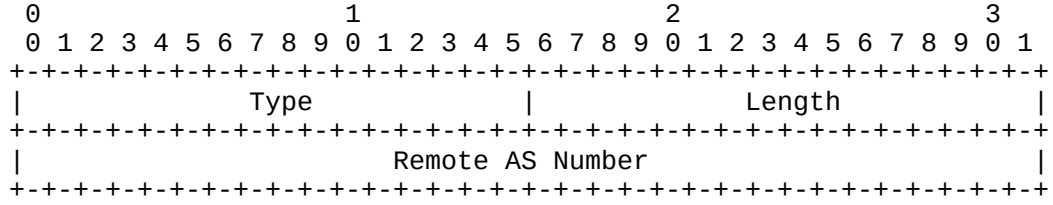
TLV Code Point	Description	IS-IS/OSPF TLV /Sub-TLV	Reference (RFC/Section)
TBD	Remote AS Number	24/21	[RFC5316]/3.3.1
			[RFC5392]/3.3.1
TBD	IPv4 Remote ASBR ID	25/22	[RFC5316]/3.3.2
			[RFC5392]/3.3.2
TBD	IPv6 Remote ASBR ID	26/24	[RFC5316]/3.3.3
			[RFC5392]/3.3.3

Detail encoding of these TLVs are synchronized with the corresponding parts in [RFC5316] and [RFC5392], which keeps the BGP-LS protocol is agnostic to the underly protocol.

4.1. Remote AS Number TLV

A new TLV, the remote AS number TLV, is defined for inclusion in the link descriptor when advertising inter-AS links. The remote AS number TLV specifies the AS number of the neighboring AS to which the advertised link connects.

The remote AS number TLV is TLV type TBD (see Section 7) and is 4 octets in length. The format is as follows:

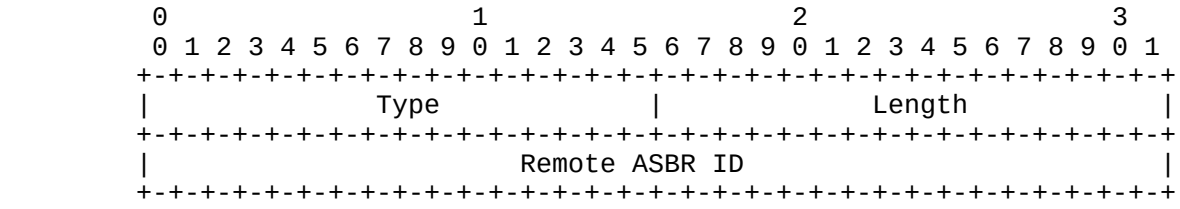


The Remote AS number field has 4 octets. When only 2 octets are used for the AS number, as in current deployments, the left (high-order) 2 octets MUST be set to 0. The remote AS number TLV MUST be included when a router advertises an inter-AS TE link.

4.2. IPv4 Remote ASBR ID

A new TLV, which is referred to as the IPv4 remote ASBR ID TLV, is defined for inclusion in the link descriptor when advertising inter-AS links. The IPv4 remote ASBR ID TLV specifies the IPv4 identifier of the remote ASBR to which the advertised inter-AS link connects. This could be any stable and routable IPv4 address of the remote ASBR. Use of the TE Router ID as specified in the Traffic Engineering router ID TLV [RFC5305] is RECOMMENDED.

The IPv4 remote ASBR ID TLV is TLV type TBD (see Section 7) and is 4 octets in length. The format of the IPv4 remote ASBR ID sub-TLV is as follows:

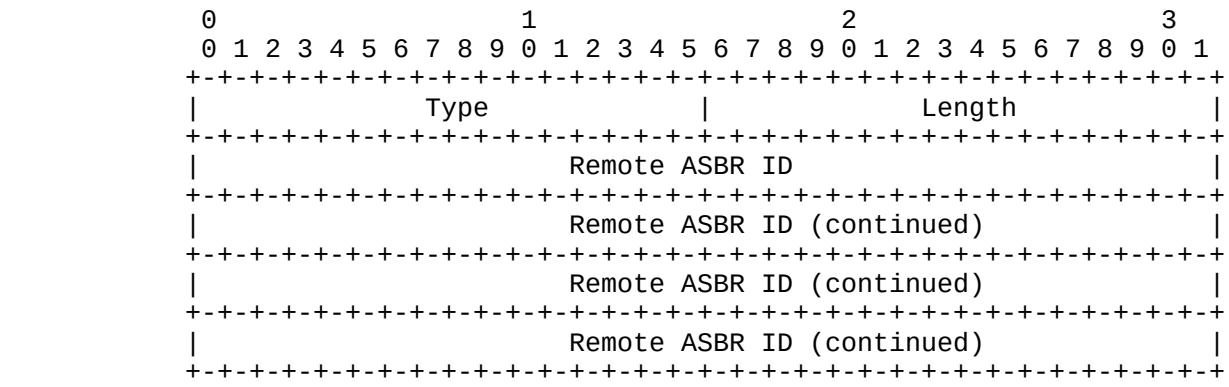


The IPv4 remote ASBR ID TLV MUST be included if the neighboring ASBR has an IPv4 address. If the neighboring ASBR does not have an IPv4 address (not even an IPv4 TE Router ID), the IPv6 remote ASBR ID TLV MUST be included instead. An IPv4 remote ASBR ID TLV and IPv6 remote ASBR ID TLV MAY both be present in an extended IS reachability TLV.

4.3. IPv6 Remote ASBR ID

A new TLV, which is referred to as the IPv6 remote ASBR ID TLV, is defined for inclusion in the inter-AS reachability TLV when advertising inter-AS links. The IPv6 remote ASBR ID TLV specifies the IPv6 identifier of the remote ASBR to which the advertised inter-AS link connects. This could be any stable and routable IPv6 address of the remote ASBR. Use of the TE Router ID as specified in the IPv6 Traffic Engineering router ID TLV [RFC6119] is RECOMMENDED.

The IPv6 remote ASBR ID TLV is TLV type TBD (see Section 7) and is 16 octets in length. The format of the IPv6 remote ASBR ID TLV is as follows:



The IPv6 remote ASBR ID TLV MUST be included if the neighboring ASBR has an IPv6 address. If the neighboring ASBR does not have an IPv6 address, the IPv4 remote ASBR ID TLV MUST be included instead. An IPv4 remote ASBR ID TLV and IPv6 remote ASBR ID TLV MAY both be present in an extended IS reachability TLV.

5. Topology Reconstruction.

When SDN Controller gets such information from BGP-LS protocol, it should compares the proximity of the redistributed prefixes. If they are under the same network scope, then it should find the corresponding associated router information, build the link between these two border routers.

After iterating the above procedures for all of the redistributed prefixes, the SDN controller can then retrieve the connection topology between different domains automatically.

6. Security Considerations

It is common for one operator to occupy several IGP domains that composited by its backbone network and several MAN(Metrio-Area-Network)s/IDCs. When they do traffic engineering from end to end that spans MAN-backbone-IDC, they need to know the inter-as topology via the process described in this draft. Then it is naturally to redistribute the interconnection prefixes in Native IP scenario.

If these IGP domains belong to different operators, it is uncommon do inter-as traffic engineering under one PCE/SDN controller, then it is unnecessary to get the inter-as topology. But redistributing the interconnection prefixes will do no harm to their networks, because the redistributed interconnection link prefixes belongs to both of them, they are also the interfaces addresses on the border routers. .

7. IANA Considerations

TBD.

8. Acknowledgement

The author would like to thank Acee Lindem, Ketan Talaulikar, Jie Dong, Jeff Tantsura and Dhruv Dhody for their valuable comments and suggestions.

9. Normative References

[I-D.ietf-idr-bgp-ls-segment-routing-ext]
Previdi, S., Talaulikar, K., Filsfils, C., Gredler, H., and M. Chen, "BGP Link-State extensions for Segment Routing", draft-ietf-idr-bgp-ls-segment-routing-ext-08 (work in progress), May 2018.

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", draft-ietf-idr-bgpls-segment-routing-epe-15 (work in progress), March 2018.

[I-D.ietf-ospf-ospfv3-lsa-extend]
Lindem, A., Roy, A., Goethals, D., Vallem, V., and F. Baker, "OSPFv3 LSA Extendibility", draft-ietf-ospf-ospfv3-lsa-extend-23 (work in progress), January 2018.

[I-D.ietf-teas-native-ip-scenarios]
Wang, A., Huang, X., Qou, C., Li, Z., Huang, L., and P. Mi, "CCDR Scenario, Simulation and Suggestion", draft-ietf-teas-native-ip-scenarios-01 (work in progress), June 2018.

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <https://www.rfc-editor.org/info/rfc1195>.

[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>.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, DOI 10.17487/RFC5316, December 2008, <https://www.rfc-editor.org/info/rfc5316>.

[RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, DOI 10.17487/RFC5392, January 2009, <https://www.rfc-editor.org/info/rfc5392>.

[RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, DOI 10.17487/RFC6119, February 2011, <https://www.rfc-editor.org/info/rfc6119>.

[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>.

[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, <https://www.rfc-editor.org/info/rfc7794>.

[RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <https://www.rfc-editor.org/info/rfc8362>.

Authors' Addresses

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangaj.bri@chinatelecom.cn

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimo.chen@huawei.com