

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 19 August 2022

A. Lindem
Cisco Systems
K. Patel
Arrcus, Inc
S. Zandi
LinkedIn
J. Haas
Juniper Networks, Inc
X. Xu
Capitalonline
15 February 2022

BGP Logical Link Discovery Protocol (LLDP) Peer Discovery
draft-acee-idr-lldp-peer-discovery-11

Abstract

Link Layer Discovery Protocol (LLDP) or IEEE Std 802.1AB is implemented in networking equipment from many vendors. It is natural for IETF protocols to avail this protocol for simple discovery tasks. This document describes how BGP would use LLDP to discover directly connected and 2-hop peers when peering is based on loopback addresses.

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 19 August 2022.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
1.1. Requirements Notation	3
1.1.1. Requirements Language	3
2. LLDP Extensions	3
2.1. LLDP IETF Organizationally Specific TLV Format	3
2.2. BGP Config OS-TLV Format	4
2.2.1. BGP Config OS-TLV - Peering Address Sub-TLV	4
2.2.2. BGP Config OS-TLV - BGP Local AS Sub-TLV	5
2.2.3. BGP Config OS-TLV - BGP Identifier Sub-TLV	6
2.2.4. BGP Config OS-TLV - Session Group-ID Sub-TLV	7
2.2.5. BGP Config OS-TLV - BGP Session Capabilities Sub-TLV	7
2.2.6. BGP Config OS-TLV - Key Chain Sub-TLV	8
2.2.7. BGP Config OS-TLV - Local Address Sub-TLV	9
2.2.8. BGP Config OS-TLV - BGP State Version Sub-TLV	10
3. BGP LLDP Peer Discovery Operations	11
3.1. Advertising BGP Speaker	11
3.2. Receiving BGP Speaker	12
3.3. Updating or Deleting Auto-Discovery Parameters	13
4. LLDP Authentication/Encryption	13
5. Security Considerations	14
6. IANA Considerations	14
6.1. IANA Assigned LLDP Subtype	14
6.2. BGP Config LLDP OS-TLV Sub-TLVs	15
7. Contributors	16
8. References	16
8.1. Normative References	16
8.2. Informative References	16
Appendix A. Acknowledgments	17
Authors' Addresses	17

1. Introduction

Link Layer Discovery Protocol (LLDP) [LLDP] or IEEE Std 802.1AB is implemented in networking equipment from many vendors. It is natural for IETF protocols to avail this protocol for simple discovery tasks. This document describes how BGP [RFC4271] would use LLDP to discover directly connected and 2-hop peers when peering is based on loopback addresses.

1.1. Requirements Notation

1.1.1. 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. LLDP Extensions

2.1. LLDP IETF Organizationally Specific TLV Format

The format of the LLDP IETF Organizationally Specific TLV (OS-TLV) is defined in [LLDP]. It is shown below for completeness.

0																1																2																3																															
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1																																																
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																															
Type (127)																Length																OUI (3 Octets) 00-00-5E																																															
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																															
OUI Continued																Subtype																Value																																															
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+																															
... (Up to 507 Octets)																																																																															

Type	IETF Organizationally Specific TLV type value, 127.
------	---

Length The length of the remainder of the TLV.

OUI	IETF Organizationally unique identifier for the organization's OUI. For IANA, this is value is 00-00-5E as specified in [IEEE-802-IANA].
-----	--

Subtype IETF specific subtype

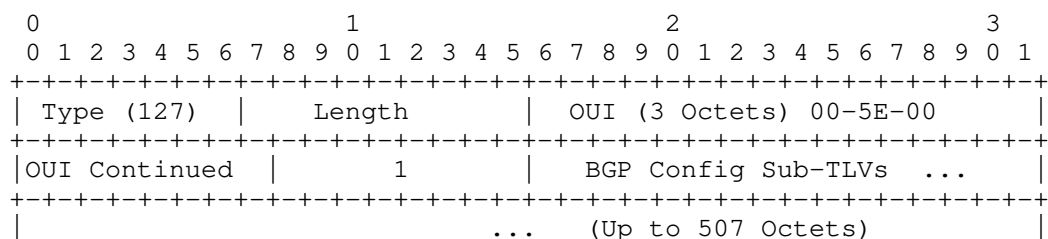
Value	Value for organizationally specific TLV. The Length of the value is 4 octets less than the TLV length.
-------	--

Figure 1: LLDP IETF Organizationally Specific TLV

The OUI for IANA was allocated in section 1.4 of [RFC7042]. This document requests creation of a registry for IETF specific sub-types for LLDP IETF Organizationally Specific TLVs.

2.2. BGP Config OS-TLV Format

The BGP Config IETF Organizationally Specific TLV (OS-TLV) will be used to advertise BGP configuration information. The configuration information will be composed of Sub-TLVs. Since the length is limited to 507 octets, multiple BGP Config OS-TLVs could be included in a single LLDP advertisement.



Length The length of the BGP TLV.

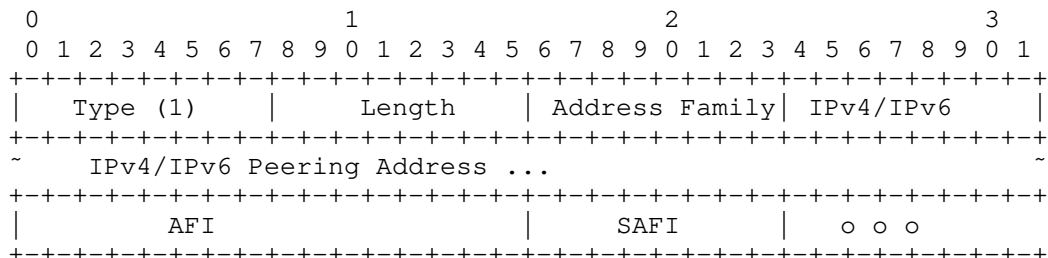
Subtype IETF specific subtype for BGP Config OS-TLV. The value shall be 1.

Value BGP Config Sub-TLVs each with a 1 byte Type and Length. The Length will include solely the value portion of the TLV and not the Type and Length fields themselves.

2.2.1. BGP Config OS-TLV - Peering Address Sub-TLV

The BGP OS-TLV Peering Address Sub-TLV will be used to advertise the local IP addresses used for BGP sessions and the associated address families specified by AFI/SAFI tuples. The AFI/SAFI tuple, 0/0, indicates to use the associated peering address for all locally configured address families without an explicit peering address specification. As always, the address families supported for a given BGP session will be determined during capabilities negotiation [RFC4760]. It is RECOMMENDED that the wildcard AFI/SAFI be used in deployments with fairly homogenous address family usage.

The format of the BGP Peering Address Sub-TLV is shown below.



Type The Sub-TLV Type value shall be 1.

Length The Sub-TLV length in octets will be 4 for IPv4 or 16 for IPv6 plus 3 times the number of AFI/SAFI tuples.

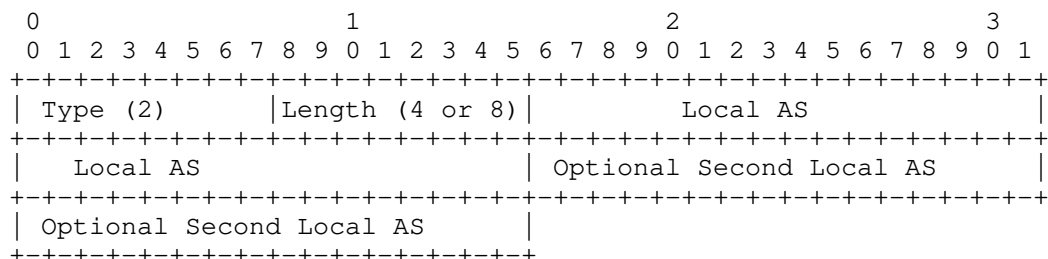
Address Family IANA Address family (1 for IPv4 or 2 for IPv6)

Peering Address An IPv4 address (4 octets) or an IPv6 address (16 octets)

AFI/SAFI Pairs One or more AFI/SAFI tuples for BGP session using this peering address. The AFI/SAFI tuple, 0/0, is a wildcard indicating to attempt negotiation for all AFI/SAFIs.

2.2.2. BGP Config OS-TLV - BGP Local AS Sub-TLV

The BGP Config OS-TLV Local AS Sub-TLV will be used to advertise the 4-octet local Autonomous System (AS) number(s). For AS transitions, a second local AS number may be specified. The format of the BGP Local AS Sub-TLV is shown below.



Type The Sub-TLV Type value shall be 2.

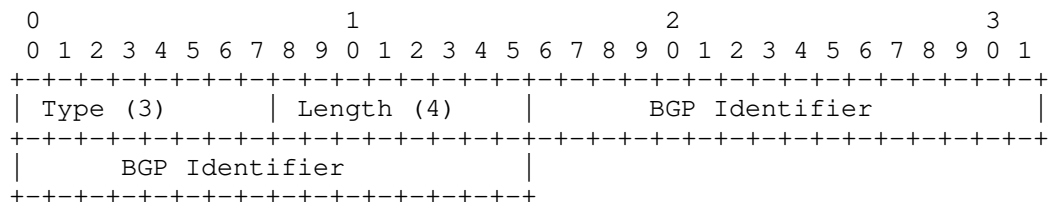
Length The Sub-TLV Length will be 4 or 8 octets.

Local AS Local Autonomous System (AS)

Second Local AS Local Autonomous System (AS)

2.2.3. BGP Config OS-TLV - BGP Identifier Sub-TLV

The BGP Config OS-TLV BGP Identifier Sub-TLV will be used to advertise the 4-octet local BGP Identifier. The BGP Identifier is used for debugging purposes and possibly to reduce the likelihood of BGP connection collisions. The format of the BGP Identifier Sub-TLV is shown below.



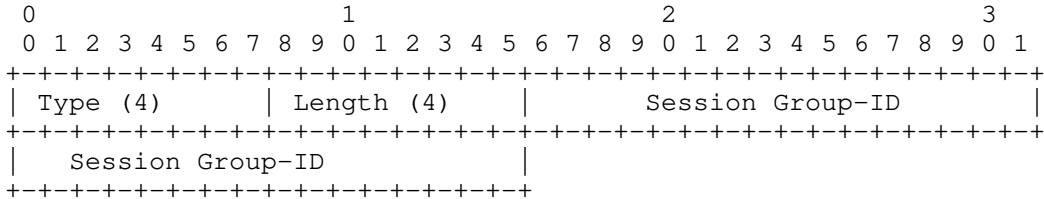
Type The Sub-TLV Type value shall be 3.

Length The Sub-TLV Length will be 4 octets.

BGP Identifier Local BGP Identifier (aka, BGP Router ID)

2.2.4. BGP Config OS-TLV - Session Group-ID Sub-TLV

The BGP Config OS-TLV Session Group-ID Sub-TLV is an opaque 4-octet value that is used to represent a category of BGP session that is supported on the interface. The format of the Session Group-ID Sub-TLV is shown below.



- Type

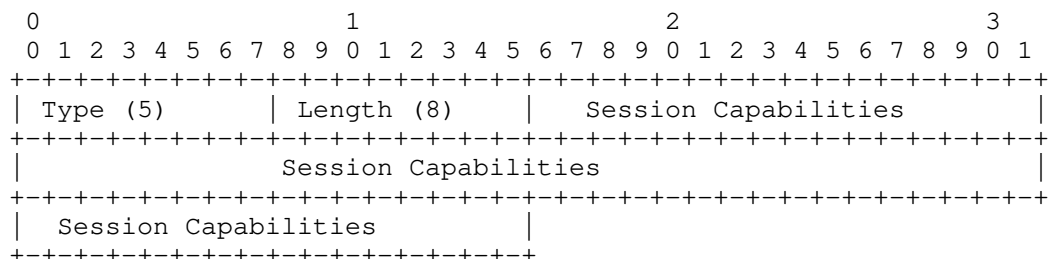
The Sub-TLV Type value shall be 4.
- Length

The Sub-TLV Length will be 4 octets.
- Session Group-ID

The session group-id used to indicate a class or category of BGP session supported on the interface.

2.2.5. BGP Config OS-TLV - BGP Session Capabilities Sub-TLV

The BGP Config OS-TLV Session Capabilities Sub-TLV will be used to advertise an 8-octet Session Capabilities field. The session capabilities are represented as bit flags identifying the supported BGP session capabilities. The format of the BGP Session Capabilities Sub-TLV is shown below.



Type The Sub-TLV Type value shall be 5.

Length The Sub-TLV Length will be 8 octets.

Session Capabilities Bit fields identify BGP session capabilities

The BGP Session Capabilities is an 8-octet bit field. The most significant bit is the first bit (Bit 1) of the Session Capabilities. The following bits are defined:

Bit 1: This bit indicates support for TCP MD5 authentication [TCP-MD5].

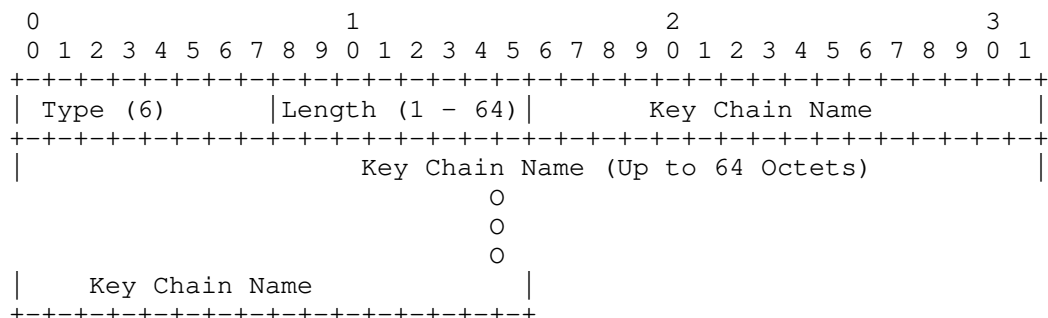
Bit 2: This bit indicates support for TCP-AO authentication [TCP-AO].

Bit 3: This bit indicates support for Generalized TTL Security Mechanism (GTSM) [GTSM] with a configured TTL range of 254-255.

TCP MD5 authentication is described in [RFC2385]. The TCP Authentication Option (TCP-AO) is described in [RFC5925]. The Generalized TTL Security Mechanism (GTSM) is described in [RFC5082]. If both TCP MD5 authentication and TCP-AO authentication are specified and TCP-AO is supported, it will take precedence.

2.2.6. BGP Config OS-TLV - Key Chain Sub-TLV

The BGP Config OS-TLV Key Chain Sub-TLV is a string specifying the name for the key chain used for session authentication. Key chains [RFC8177] are a commonly used for protocol authentication and encryption key specification. Given the limited length of all BGP configuration information, the key chain name will be limited to 64 characters and will not include a trailing string delimiter. The format of the Session Group-ID Sub-TLV is shown below.



Type The Sub-TLV Type value shall be 6.

Length The Sub-TLV Length will be 1 - 64 octets.

Key Chain Name The name of a key chain to be used for
MD5 or TCP-AO authentication.

2.2.7. BGP Config OS-TLV - Local Address Sub-TLV

The BGP OS-TLV Local Address Sub-TLV will be used to advertise a local IP addresses used for BGP next-hops. Advertising a local interface address is useful when the address family is different from the advertised BGP peering address.

The format of the BGP Local Interface Address Sub-TLV is shown below.

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|  Type (7)   |   Length   | Address Family | IPv4/IPv6   |
+-----+-----+-----+-----+-----+-----+-----+-----+
~   IPv4/IPv6 Local Address ...                               ~
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Type The Sub-TLV Type value shall be 7.

Length The Sub-TLV length in octets will be 4 for IPv4 or 16
 for IPv6 plus 3 times the number of AFI/SAFI tuples.

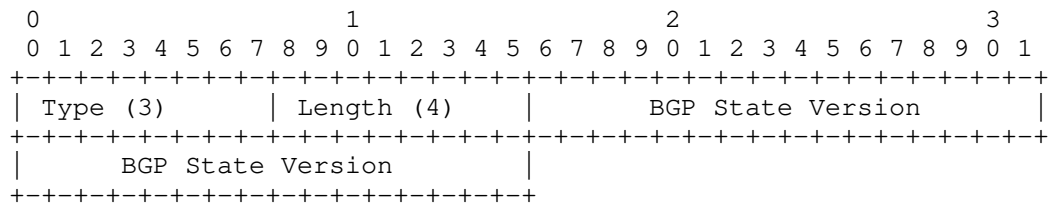
Address Family IANA Address family (1 for IPv4 or 2 for IPv6)

Local Address An IPv4 address (4 octets) or an IPv6 address (16 octets)

2.2.8. BGP Config OS-TLV - BGP State Version Sub-TLV

The BGP OS-TLV Version Sub-TLV will be used to advertise a monotonically increasing version. This version will indicate if any local BGP state that may impact BGP session establishment has changed. Changes can range from anything as obvious a change in local peering address to more indirect changes such as the modification of the key-chain being advertised.

The format of the BGP State Version Sub-TLV is shown below.



Type The Sub-TLV Type value shall be 8.

Length The Sub-TLV Length will be 4 octets.

BGP State Version BGP State Version - Monotonically increasing version number indicating if any local state that may effect BGP session establishment has changed.

3. BGP LLDP Peer Discovery Operations

The simple use case is to just use the peer address advertised in the LLDP Packet Data Unit (PDU) to establish a 1-hop BGP peer session. This can be used in data centers using BGP as described in [RFC7938]. The use case where a loopback address or other local address is advertised as the peering address is also supported. However, reachability to a peering address other than the interface address is beyond the scope of this document.

3.1. Advertising BGP Speaker

A BGP speaker MAY advertise its BGP peering address in an LLDP PDU for a link using the BGP Local Address Sub-TLV of the BGP-OS TLV. This can be an IPv4 or IPv6 local address associated with the LLDP link for 1-hop peering. For 2-hop peering, it could be a loopback address or any other address that is local to the node but not the LLDP link. As noted above, reachability to the loopback address is beyond the scope of this document.

A BGP speaker MAY advertise its local AS number using the BGP Local AS Sub-TLV of the BGP-OS TLV. During AS transitions, a second local AS number may be included in the Local AS Sub-TLV. The local BGP identifier may also be advertised using the BGP Identifier Sub-TLV of the BGP-OS TLV. While not specifically required for session establishment, the values may be used for validation, troubleshooting, and connection collision avoidance. A BGP speaker may also announce a Session Group-ID indicating the class or category of

session(s) supported and/or mapping to a set of session parameters. Additionally, a BGP speaker MAY also announce relevant capabilities using BGP Session Capabilities Sub-TLV of the BGP-OS TLV.

If TCP MD5 authentication [RFC2385] or TCP Authentication Option (TCP-AO) [RFC5925] is to be used on the session, the Key Chain Sub-TLV of the BGP-OS TLV MAY be used to specify the key chain name.

3.2. Receiving BGP Speaker

A BGP speaker configured for LLDP peer discovery WILL attempt to establish BGP sessions using the address in the BGP Local Address Sub-TLV of BGP-OS TLV format. If the peering address is directly accessible over the link on which the LLDP PDU is received, the BGP speaker will attempt to establish a 1-hop BGP session with the peer.

If the received BGP Peering Address is not directly accessible over the link, the peer must be reachable for the session to be established and the mechanisms for establishing reachability are beyond the scope of this specification. If the BGP speaker receives the same BGP peering address in LLDP PDUs received on multiple links, it will not establish multiple sessions. Rather, a single 2-hop session will be established.

When the deployment of address families is fairly homogenous across the deployment, the wildcard AFI/SAFI can be utilized to simplify LLDP advertisement. When there is variance in the address families supported, usage of the wildcard could result in session establishment delay due to capabilities negotiation [RFC5492].

A BGP speaker MAY receive a remote neighbor's local AS number(s) in an LLDP PDU in the BGP Local AS Sub-TLV of the BGP-OS TLV. A BGP speaker MAY use the received local AS number(s) to perform validation checking of the AS received in the OPEN message. A BGP speaker MAY receive a remote neighbor's BGP Identifier in the BGP Identifier Sub-TLV of the BGP-OS TLV. This can be used to avoid connection collisions by delaying session establishment if the remote BGP Identifier is greater than the receiving speaker's BGP Identifier.

A BGP speaker MAY receive a Session Group-ID Sub-TLV in the LLDP BGP-OS TLV. This Session Group-ID may be used for validation and/or mapping the session to a particular set of session parameters. For example, the Session Group-ID could be mapped to a spine, leaf, or Top-of-Rack (ToR) session in a data center deployment and can be used to detect cabling problems when an unexpected Session Group-ID is received.

Additionally, A BGP speaker MAY receive a remote neighbor's capabilities in LLDP in the BGP Session Capabilities Sub-TLV of the BGP-OS TLV. A BGP speaker MAY use the received capabilities to ensure appropriate local neighbor configuration in order to facilitate session establishment.

If TCP MD5 authentication [RFC2385]. or TCP Authentication Option (TCP-AO) [RFC5925] is to be used on the session as determined either via the Session Capabilities Sub-TLV, Session Group-ID, or local policy, the key chain name in the Key Chain Sub-TLV of the BGP-OS TLV MAY be used to identify the correct key chain [RFC8177].

The BGP State Version associated with the LLDP peer SHOULD be retained to determine whether anything impacting BGP session establishment has changed. When session establishment fails, this can be used to avoid back-off on attempting to establish a BGP session when nothing has changed on the peer or locally.

3.3. Updating or Deleting Auto-Discovery Parameters

A BGP speaker MAY change or delete any BGP LLDP auto-discovery parameter by simply updating or removing the corresponding Sub-TLV previously advertised in the BGP-OS TLV. Additionally, the BGP State Version Sub-TLV should be advertised with the version incremented from the previous version. The BGP speaker(s) receiving the advertisement will update or delete the changed or deleted auto-discovery parameters. However, there will be no change to existing BGP sessions with the advertising BGP Speaker. Changes to existing BGP sessions are the purview of the BGP protocol and are beyond the scope of this document.

Since LLDP information is cumulative, reception of an LLDP PDU without the BGP-OS TLV indicates that BGP LLDP auto-discovery has been disabled for the BGP speaker and all parameters learnt during BGP LLDP auto-discovery SHOULD be deleted. As above, changes to existing BGP sessions are beyond the scope of this document.

4. LLDP Authentication/Encryption

The IEEE 802.1AE [MACsec] standard can be used for encryption and/or authentication to provide privacy and integrity. MACsec utilizes the Galois/Counter Mode Advanced Encryption Standard (AES-GCM) for authenticated encryption and Galois Message Authentication Code (GMAC) if only authentication, but not encryption is required.

The MACsec Key Agreement (MKA) is included as part of the IEEE 802.1X-20200 Port-Based Network Access Control Standard [MKA]. The purpose of MKA is to provide a method for discovering MACsec peers and negotiating the security keys needed to secure the link.

5. Security Considerations

This security considerations for BGP [RFC4271] apply equally to this extension.

Additionally, BGP peering address discovery should only be done on trusted links (e.g., in a data center network) since LLDP packets are not authenticated or encrypted [LLDP].

LLDP Authentication and/or encryption can provided as described in section Section 4.

6. IANA Considerations

6.1. IANA Assigned LLDP Subtype

IANA is requested to create a registry for IANA assigned subtypes in the IETF Organizationally Specific TLV assigned to IANA (OUI of 000-00-53 [RFC7042]). Assignment is requested for 1 for the BGP Config OS-TLV.

Range	Assignment Policy
0	Reserved (not to be assigned)
1	BGP Configuration
2-127	Unassigned (IETF Review)
128-254	Reserved (Not to be assigned now)
255	Reserved (not to be assigned)

Figure 2: IANA LLDP IETF Organizationally Specific TLV Sub-Types

- * Types in the range 2-127 are to be assigned subject to IETF Review. New values are assigned only through RFCs that have been shepherded through the IESG as AD-Sponsored or IETF WG Documents [RFC5226].

- * Types in the range 128-254 are reserved and not to be assigned at this time. Before any assignments can be made in this range, there MUST be a Standards Track RFC that specifies IANA Considerations that covers the range being assigned.

6.2. BGP Config LLDP OS-TLV Sub-TLVs

IANA is requested to create a registry for Sub-TLVs of the BGP Config LLDP OS-TLV. Assignment is requested for 1 for the BGP Peering Address Sub-TLV. Assignment is also requested for 2 for the Local AS Sub-TLV. Additionally, assignment is requested for 3 for the BGP Identifier Sub-TLV, 4 for the BGP Session Group-ID, 5 for the Session Capabilities Sub-TLV, and 6 for the Key Chain Name.

Range	Assignment Policy
0	Reserved (not to be assigned)
1	Peering Address
2	Local AS
3	BGP Identifier
4	Session Group-ID
5	Session Capabilities
6	Key Chain Name
7	Local Address
8	BGP State Version
9-127	Unassigned (IETF Review)
128-254	Reserved (Not to be assigned now)
255	Reserved (not to be assigned)

Figure 3: LLDP BGP Config OS-TLV Types

- * Types in the range 9-127 are to be assigned subject to IETF Review. New values are assigned only through RFCs that have been shepherded through the IESG as AD-Sponsored or IETF WG Documents [RFC5226].

- * Types in the range 128-254 are reserved and not to be assigned at this time. Before any assignments can be made in this range, there MUST be a Standards Track RFC that specifies IANA Considerations that covers the range being assigned.

7. Contributors

Contributors' Addresses

8. References

8.1. Normative References

- [LLDP] IEEE, "IEEE Standard for Local and metropolitan area networks-- Station and Media Access Control Connectivity Discovery Corrigendum 2: Technical and Editorial Corrections", IEEE 802.1AB-2009/Cor 2-2015, DOI 10.1109/ieeestd.2015.7056401, 9 March 2015, <<https://doi.org/10.1109/ieeestd.2015.7056401>>.
- [MACsec] IEEE, "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Security", IEEE Standard 802.1AE-2018, 27 September 2018.
- [MKA] IEEE, "IEEE Standard for Local and metropolitan area networks - Port Based Network Access Control", IEEE Standard 802.1X-2020, 30 January 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>>.
- [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>>.

8.2. Informative References

- [RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, DOI 10.17487/RFC2385, August 1998, <<https://www.rfc-editor.org/info/rfc2385>>.

- [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>>.
- [RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, DOI 10.17487/RFC5082, October 2007, <<https://www.rfc-editor.org/info/rfc5082>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <<https://www.rfc-editor.org/info/rfc5226>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<https://www.rfc-editor.org/info/rfc5492>>.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<https://www.rfc-editor.org/info/rfc5925>>.
- [RFC7042] Eastlake 3rd, D. and J. Abley, "IANA Considerations and IETF Protocol and Documentation Usage for IEEE 802 Parameters", BCP 141, RFC 7042, DOI 10.17487/RFC7042, October 2013, <<https://www.rfc-editor.org/info/rfc7042>>.
- [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>>.
- [RFC8177] Lindem, A., Ed., Qu, Y., Yeung, D., Chen, I., and J. Zhang, "YANG Data Model for Key Chains", RFC 8177, DOI 10.17487/RFC8177, June 2017, <<https://www.rfc-editor.org/info/rfc8177>>.

Appendix A. Acknowledgments

Thanks to Sujay Gupta and Paul Congdon for review and comments.

The RFC text was produced using Marshall Rose's xml2rfc tool.

Authors' Addresses

Acee Lindem
Cisco Systems
301 Midenhall Way
Cary, NC 27513
United States of America

Email: acee@cisco.com

Keyur Patel
Arrcus, Inc

Email: keyur@arrcus.com

Shawn Zandi
LinkedIn
222 2nd Street
San Francisco, CA 94105
United States of America

Email: szandi@linkedin.com

Jeff Haas
Juniper Networks, Inc
1133 Innovation, Inc.
Sunnyvale, CA 94089
United States of America

Email: jhaas@juniper.net

Xiaohu Xu
Capitalonline

Email: xiaohu.xu@capitalonline.net

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/>.

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

This Internet-Draft will expire on April 25, 2019.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
2. SRv6 Services TLV	4
3. BGP based L3 over SRv6	6
3.1. IPv4 VPN Over SRv6 Core	7
3.2. IPv6 VPN Over SRv6 Core	7
3.3. Global IPv4 over SRv6 Core	8
3.4. Global IPv6 over SRv6 Core	8
4. BGP based Ethernet VPN(EVPN) over SRv6	9
4.1. Ethernet Auto-discovery Route over SRv6 Core	10
4.1.1. EVPN Route Type-1(Per ES AD)	10
4.1.2. Prefix Type-1(Per EVI/ES AD)	11
4.2. MAC/IP Advertisement Route(Type-2) with SRv6 Core	11
4.3. Inclusive Multicast Ethernet Tag Route with SRv6 Core	13
4.4. Ethernet Segment Route with SRv6 Core	14
4.5. IP prefix router(Type-5) with SRv6 Core	15
4.6. Multicast routes (EVPN Route Type-6, Type-7, Type-8)	15
5. Migration from L3 MPLS based Segment Routing to SRv6 Segment Routing	16
6. Implementation Status	16
7. Error Handling of BGP SRv6 SID Updates	17
8. IANA Considerations	17
9. Security Considerations	18
10. Conclusions	18
11. References	18

11.1.	Normative References	18
11.2.	Informative References	19
11.3.	URIs	20
Appendix A.	Acknowledgements	20
Appendix B.	Contributors	21
Authors' Addresses	21

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]

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
      +-----+-----+-----+-----+-----+-----+-----+-----+
      | TLV Type | Length | RESERVED |
      +-----+-----+-----+-----+-----+-----+-----+-----+
      // SRv6 Service Information (variable) //
      +-----+-----+-----+-----+-----+-----+-----+-----+

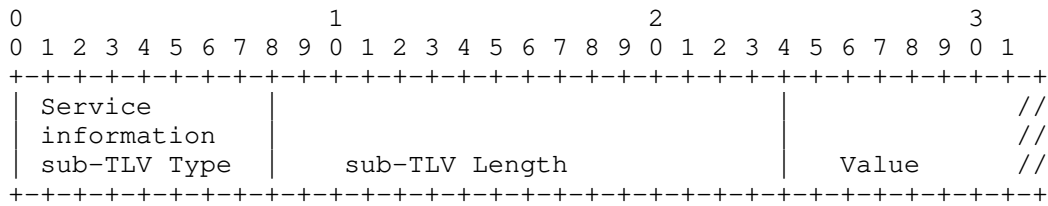
```

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

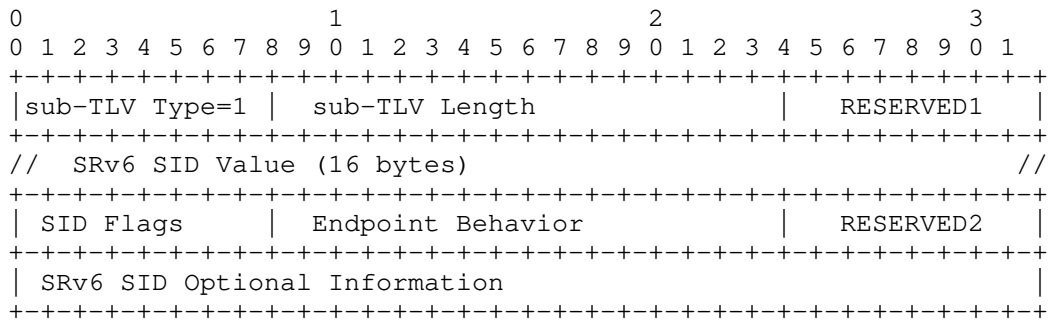
- SRv6 L3 Service TLV. Type code 5 (to be assigned by IANA as described in section 8). This TLV encodes Service SID information for the SRv6 based L3 services. It corresponds to the equivalent functionality provided by an MPLS Label when received with a Layer 3 VPN route [RFC4364]. Some functions which 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:

- o Type is 1 (to be assigned by IANA as described in Section 8). As defined to be "SID information sub-TLV".
- o Length: 16 bit field. The total length of the value portion of the sub-TLV.
- o RESERVED1: 8 bit field. SHOULD be 0 on transmission and MUST be ignored on reception.
- o SRv6 SID Value: 128 bit field. Encodes an SRv6 SID as defined in [I-D.filsfils-spring-srv6-network-programming]
- o SID Flags: 8 bit field. Encodes SRv6 SID Flags. Value is opaque to BGP.
- o Endpoint Behavior : 16 bit field. Encodes Endpoint behavior. For SRv6 VPN services, this field is always set to (0xFFFF).
- o RESERVED2: 8 bit field. SHOULD be 0 on transmission and MUST be ignored on reception.
- o 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.

0									1									2									3								
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1				
SID Optional									sub-TLV Length									Value									//								
information																											//								
sub-TLV Type																											//								

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:

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

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

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

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

- o 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:

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

```

+-----+
|  RD (8 octets)  |
+-----+
|Ethernet Segment Identifier (10 octets)|
+-----+
|  Ethernet Tag ID (4 octets)  |
+-----+
|  MPLS label (3 octets)  |
+-----+

```

For a SRv6 only BGP speaker for an SRv6 Core:

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

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

Where:

- o BGP next-hop: IPv6 address of an egress PE
- o Ethernet Tag ID: all FFFF's
- o MPLS Label: always set to zero value
- o 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:

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

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
MAC Address Length (1 octet)
MAC Address (6 octets)
IP Address Length (1 octet)
IP Address (0, 4, or 16 octets)
MPLS Label1 (3 octets)
MPLS Label2 (0 or 3 octets)

where:

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

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

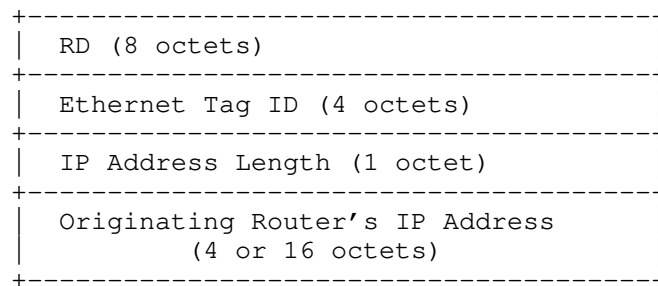
Described below are different types of Type-2 advertisements.

- o 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
- o 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:

- o BGP next-hop: IPv6 address of egress PE
- o SRv6 Service TLV MAY encode END.DX2/END.DT2M function.
- o 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 (1 octet)
Tunnel Type (1 octet)
MPLS label (3 octet)
Tunnel Identifier (variable)

- o Flag: zero value defined per [RFC7432]
- o Tunnel Type: defined per [RFC6514]
- o MPLS label: Implicit-Null
- o 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:

- o END.DX2 or END.DT2M function
- o 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]

RD (8 octets)
Ethernet Tag ID (4 octets)
IP Address Length (1 octet)
Originating Router's IP Address (4 or 16 octets)

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:

RD (8 octets)
Ethernet Segment Identifier (10 octets)
Ethernet Tag ID (4 octets)
IP Prefix Length (1 octet)
IP Prefix (4 or 16 octets)
GW IP Address (4 or 16 octets)
MPLS Label (3 octets)

- 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]:

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

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

- o 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:

- o If the TLV is advertised by a CE peer, the receiving PE may discard it before advertising the route to its PE peers.
- o 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:

- o Type 5: SRv6 Layer3 Service
- o 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

Jonh 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
Updates: 6790 (if approved)
Intended status: Standards Track
Expires: July 30, 2017

B. Decraene
Orange
K. Kompella
Juniper Networks, Inc.
W. Henderickx
Nokia
January 26, 2017

BGP Next-Hop dependant capabilities
draft-decraene-idr-next-hop-capability-03

Abstract

RFC 5492 defines capabilities advertisement for the BGP peer. In addition, it is useful to be able to advertise BGP Next-Hop dependant capabilities, in particular for forwarding plane features. RFC 5492 is not applicable because the BGP peer may be different from the BGP Next-Hop, in particular when BGP Route Reflection is used. This document defines a mechanism to advertise such BGP Next Hop dependant Capabilities.

This document defines a new BGP non-transitive attribute to carry Next-Hop Capabilities. This attribute is deleted or possibly modified when the BGP Next Hop is changed.

This document also defines a Next-Hop capability to advertise the ability to handle the MPLS Entropy Label defined in RFC 6790. It updates RFC 6790 with regard to this BGP signaling.

Requirements Language

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

Status of This Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on July 30, 2017.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
2. BGP Next-Hop dependant Capabilities Attribute	3
2.1. Encoding	3
2.2. Attribute Operation	4
2.3. Capability Code Operation	5
2.4. Attribute Error Handling	5
3. Entropy Label Next-Hop dependant Capability	6
3.1. Entropy Label Next-Hop Capability error handling	7
4. IANA Considerations	7
4.1. Next-Hop Capabilities Attribute	7
4.2. Next-Hop Capability registry	7
5. Security Considerations	8
6. Acknowledgement	8
7. References	8
7.1. Normative References	8
7.2. Informative References	9
Authors' Addresses	9

1. Introduction

[RFC5492] defines capabilities advertisement for the BGP peer. In addition, it is useful to be able to advertise BGP Next-Hop dependant capabilities, in particular for forwarding plane features. RFC 5492 is not applicable because the BGP peer may be different from the BGP Next-Hop, in particular when BGP Route Reflection is used. This

document defines a mechanism to advertise such BGP Next Hop Capabilities.

This document defines a new BGP non-transitive attribute to carry Next-Hop Capabilities. When the BGP Next Hop is changed, this attribute is deleted or possibly modified to take into account the capabilities of the new BGP Next-Hop. Hence it allows advertising capabilities which are dependent of the BGP Next-Hop.

This attribute advertises the capabilities of the BGP Next-Hop for the NLRI advertised in the same BGP update. A BGP Next-Hop may advertise different capabilities for different set of NLRI.

This document also defines a first application to advertise the capability to handle the MPLS Entropy Label defined in [RFC6790]. Note that RFC 6790 had originally defined a BGP attribute for this but it has been latter deprecated in [RFC7447].

2. BGP Next-Hop dependant Capabilities Attribute

2.1. Encoding

The BGP Next-Hop dependant Capabilities Attribute is an optional, non-transitive BGP Attribute, of value TBD1. The attribute consists of a set of Next-Hop Capabilities.

The inclusion of a Next-Hop Capability "X" in a BGP UPDATE message, indicates that the BGP Next-Hop, encoded in either the NEXT_HOP attribute defined in [RFC4271] or the Network Address of Next Hop field of the MP_REACH_NLRI attribute defined in [RFC4760], supports the capability "X" for the NLRI advertised in this BGP UPDATE.

This document does not make a distinction between these two Next-Hop fields and uses the term 'BGP Next-Hop' to refer to whichever one is used in a given BGP UPDATE message.

A Next-Hop Capability is a triple (Capability Code, Capability Length, Capability Value) aka a TLV:

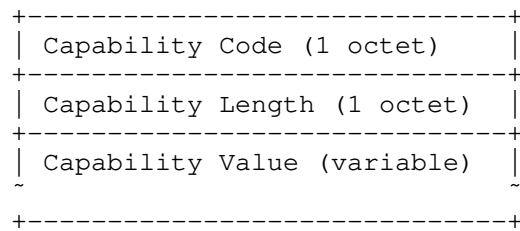


Figure 1: BGP Next-Hop Capability

Capability Code: a one-octet unsigned binary integer which indicates the type of "Next-Hop Capability" advertised and unambiguously identifies an individual capability.

Capability Length: a one-octet unsigned binary integer which indicates the length, in octets, of the Capability Value field. A length of 0 indicates that no Capability Value Field is present.

Capability Value: a variable-length field from 0 to 255 octets. It is interpreted according to the value of the Capability Code.

BGP speakers SHOULD NOT include more than one instance of a Next-Hop capability with the same Capability Code, Capability Length, and Capability Value. Note, however, that processing of multiple instances of such capability does not require special handling, as additional instances do not change the meaning of the announced capability; thus, a BGP speaker MUST be prepared to accept such multiple instances.

BGP speakers MAY include more than one instance of a capability (as identified by the Capability Code) with non-zero Capability Length field, but with different Capability Value and either the same or different Capability Length. Processing of these capability instances is specific to the Capability Code and MUST be described in the document introducing the new capability.

2.2. Attribute Operation

The BGP Next-Hop dependant Capabilities attribute being non-transitive, as per [RFC4271], a BGP speaker which does not understand it will quietly ignore it and not pass it along to other BGP peers.

A BGP speaker that understands the BGP Next-Hop dependant Capabilities Attribute and does not change the BGP Next-Hop, SHOULD NOT change the BGP Next-Hop dependant Capabilities Attribute and SHOULD pass the attribute unchanged along to other BGP peers.

A BGP speaker that understands the BGP Next-Hop dependant Capabilities Attribute and changes the BGP Next-Hop, MUST remove the received BGP Next-Hop dependant Capabilities Attribute before propagating the BGP UPDATE to other BGP peers. It MAY attach a new BGP Next-Hop dependant Capabilities attribute describing the capabilities of the new BGP Next-Hop for these NLRIs.

2.3. Capability Code Operation

A BGP speaker receiving a BGP Next-Hop Capability Code that it supports behave as defined in the document defining this Capability Code. A BGP speaker receiving a BGP Next-Hop Capability Code that it does not support MUST ignore this BGP Next-Hop Capability Code. In particular, this MUST NOT be handled as an error. In both cases, the BGP speaker MUST examine the remaining BGP Next-Hop Capability Code(s) that may be present in the BGP Next-Hop Capabilities Attribute.

The BGP Next-Hop Capability Code MUST reflect the capability of the router indicated in the BGP Next-Hop, for the NLRI advertised in the BGP UPDATE. If a BGP speaker sets the BGP Next-Hop to an address of a different router (e.g. R), it MUST NOT advertise BGP Next-Hop Capabilities not supported by this router R for these NLRI.

The presence of a Next-Hop Capability SHOULD NOT influence route selection or route preference of a route, unless tunneling is used to reach the BGP Next-Hop or the selected route has been learnt from EBGP (i.e. the Next-Hop is in a different AS). Indeed, it is in general impossible for a node to know that all BGP routers of the Autonomous System (AS) will understand a given Next-Hop Capability; and having different routers, within an AS, use a different preference for a route, may result in forwarding loops if tunnelling is not used to reach the BGP Next-Hop.

An implementations MAY allow, by configuration, removing this attribute or specific Next-Hop capabilities when advertising the routes over EBGP.

2.4. Attribute Error Handling

A BGP Next-Hop dependant Capabilities Attribute is considered malformed if the length of the Attribute is not equal to the sum of all (BGP Next-Hop dependant Capability Length +2) of the capabilities carried in this attribute. Note that "2" is the length of the fields "Type" and "Length" of each BGP Next Hop dependant Capability, as the capability length only account for the length of the Value field.

A document that specifies a new Next-Hop Capability SHOULD provide specifics regarding what constitutes an error for that Next-Hop Capability.

A BGP UPDATE message with a malformed BGP Next-Hop dependant Capabilities Attribute SHALL be handled using the approach of "attribute discard" defined in [RFC7606].

Unknown Next-Hop Capabilities Codes MUST NOT be considered as an error. They MUST be silently ignored.

If a Next-Hop dependant Capability is malformed, this Next-Hop Capability Type MUST be ignored. Others Next-Hop Capabilities MUST be processed as usual.

3. Entropy Label Next-Hop dependant Capability

The Entropy Label Next-Hop Capability has type code 1 and a length of 0 or 1 octet.

The inclusion of the "Entropy Label" Next-Hop Capability indicates that the BGP Next-Hop can be sent packets, for all routes indicated in the NLRI, with a MPLS entropy label (ELI, EL) added immediately after the label stack advertised with the NLRI.

On the receiving side, suppose BGP speaker S has determined that packet P is to be forwarded according to BGP route R, where R is a route of one of the labeled address families. And suppose that L is the label stack embedded in the NLRI of route R. Then to forward packet P according to route R, S either replaces P's top label with L, or else pushes L onto the MPLS label stack. If the EL-Capability is advertised in the BGP UPDATE advertising this route R, S knows that it may safely place the ELI and an EL on the label stack immediately beneath L.

A BGP speaker S that sends an UPDATE with the BGP Next-Hop "NH" MAY include the Entropy Label Next-Hop Capability only if the NLRI are labelled and for all the NLRI in the BGP UPDATE, either of the following is true:

- o Egress case: NH is the egress of the LSP advertised with the NLRI and its capable of handling the ELI during the lookup of the MPLS top label.
- o Transit LSR case: NH is a transit LSR for the LSP advertised with the NLRI (i.e. NH swaps one of the label advertised in the NLRI) and next downstream BGP Next-Hop(s) has(have) advertised the Entropy Label Next-Hop Capability (or a similar capability

signalled by protocol P if the route is redistributed, by NH, from protocol P to BGP).

3.1. Entropy Label Next-Hop Capability error handling

If the Entropy Label Next-Hop Capability is present more than once, it MUST be considered as received once with a length of 0.

If the Entropy Label Next-Hop Capability is received with a length other than 0 or 1, it is not considered malformed, but its semantics are exactly the same as if it had a length of 1. In other words, additional octets MUST be ignored. This is to allow for graceful future extension.

4. IANA Considerations

4.1. Next-Hop Capabilities Attribute

IANA is requested to allocate a new Path Attribute, called "Next-Hop Capabilities", type Code TBD1, from the "BGP Path Attributes" registry.

4.2. Next-Hop Capability registry

The IANA is requested to create and maintain a registry entitled "Next-Hop Capabilities".

The registration policies [RFC5226] for this registry are:

1-63	IETF Review
64-127	First Come First Served
128-250	Standards Action
251-254	Experimental Use
255	Reserved

IANA is requested to make the following initial assignments:

Registry Name: Next-Hop Capability.

Value	Meaning	Reference
0	Reserved (not to be allocated)	This document
1	Entropy Label	This document
2-250	Unassigned	
251-254	Experimental	This document
255	Reserved (for futur registry extension)	This document

5. Security Considerations

This document does not introduce new security vulnerabilities in BGP. Specifically, an operator who is relying on the information carried in BGP must have a transitive trust relationship back to the source of the information. Specifying the mechanism(s) to provide such a relationship is beyond the scope of this document. Please refer to the Security Considerations section of [RFC4271] for security mechanisms applicable to BGP.

6. Acknowledgement

The Entropy Label Next-Hop Capability defined in this document is based on the ELC BGP attribute defined in section 5.2 of [RFC6790].

The authors wish to thank John Scudder for the discussions on this topic and Eric Rosen for his in-depth review of this document.

The authors wish to thank Jie Dond for his review and comments.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<http://www.rfc-editor.org/info/rfc4760>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <<http://www.rfc-editor.org/info/rfc5226>>.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <<http://www.rfc-editor.org/info/rfc6790>>.

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

7.2. Informative References

- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.
- [RFC7447] Scudder, J. and K. Kompella, "Deprecation of BGP Entropy Label Capability Attribute", RFC 7447, DOI 10.17487/RFC7447, February 2015, <<http://www.rfc-editor.org/info/rfc7447>>.

Authors' Addresses

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Kireeti Kompella
Juniper Networks, Inc.
1194 N. Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: kireeti.kompella@gmail.com

Wim Henderickx
Nokia
Copernicuslaan 50
Antwerp 2018, CA 95134
Belgium

Email: wim.henderickx@nokia.com

IDR
Internet-Draft
Updates: 4271 (if approved)
Intended status: Standards Track
Expires: September 14, 2017

S. Hares
Huawei
March 13, 2017

Decpreate Atomic Aggregate
draft-hares-depcreate-atomic-aggregate-00.txt

Abstract

This document deprecates the support for the BGP well-know discretionary attribute `ATOMIC_AGGREGATE` specified in RFC4271. It proposes the changes to RFC4271 to remove its support.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on September 14, 2017.

Copyright Notice

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

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

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

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

Table of Contents

1. Introduction	2
2. Changes to Section 4.3	2
3. Changes to Section 5 - Path Attributes	3
4. Changes to Section 9	3
4.1. Changes to section 9.1.4	3
4.2. Section 9.2 Changes	4
5. Operational Considerations	4
6. Error Handling	4
7. IANA Considerations	4
8. Security Considerations	4
9. Normative References	5
Appendix A. Acknowledgements	5
Author's Address	5

1. Introduction

The `ATOMICAggregate` well-known discretionary attribute is specified in [RFC4271] in section 5.1.6. This document specifies the changes to RFC4271 in order to remove the `ATOMICAggregate` attribute.

2. Changes to Section 4.3

delete the following text:

f) ATOMIC_AGGREGATE (Type Code 6)

ATOMIC_AGGREGATE is a well-known discretionary attribute of length 0.

Usage of this attribute is defined in 5.1.6.

3. Changes to Section 5 - Path Attributes

1: Section 5.0 should have the following changes (p. 24)

Old:

attribute	EBGP	IBGP
ORIGIN	mandatory	mandatory
AS_PATH	mandatory	mandatory
NEXT_HOP	mandatory	mandatory
MULTI_EXIT_DISC	discretionary	discretionary
LOCAL_PREF	see Section 5.1.5	required
ATOMIC_AGGREGATE	see Section 5.1.6 and 9.1.4	
AGGREGATOR	discretionary	discretionary

New:

attribute	EBGP	IBGP
ORIGIN	mandatory	mandatory
AS_PATH	mandatory	mandatory
NEXT_HOP	mandatory	mandatory
MULTI_EXIT_DISC	discretionary	discretionary
LOCAL_PREF	see Section 5.1.5	required
AGGREGATOR	discretionary	discretionary

2: Delete Section 5.1.6

4. Changes to Section 9

4.1. Changes to section 9.1.4

3: Changes to section 9.1.4

Old:

If a BGP speaker chooses to aggregate, then it SHOULD either include all ASes used to form the aggregate in an AS_SET, or add the ATOMIC_AGGREGATE attribute to the route.

New

If a BGP speaker chooses to aggregate, then it SHOULD either include all ASes used to form the aggregate in an AS_SET.

delete the following text:

"In particular, a route that carries the ATOMIC_AGGREGATE attribute MUST NOT be de-aggregated."

4.2. Section 9.2 Changes

Text to delete:

ATOMIC_AGGREGATE:
If at least one of the routes to be aggregated has
ATOMIC_AGGREGATE path attribute, then the aggregated route
SHALL have this attribute as well.

5. Operational Considerations

Input needed here.

6. Error Handling

An ATOMIC_AGGREGATE attribute received should be silently ignored.

7. IANA Considerations

IANA Is asked to deprecate the BGP Attribute: Atomic_Aggregate with this document as reference.

8. Security Considerations

Deprecating a BGP attribute does not change the BGP messages sent on over a secure transport.

Users of this mechanism should be aware that unless a transport that provides integrity (such as TCP-AO [RFC5925]) is used for the BGP session in question, BGP Attributes can be forged. This could become an attack vector.

Unless a transport that provides confidentiality (such as IPSec [RFC4303]) is used, BGP attributes Communication messages could be snooped by an attacker allowing access to BGP attributes. These issues are common to any BGP message but may be of greater interest in the context of this proposal since a BGP Attribute is being deleted.

9. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, DOI 10.17487/RFC4303, December 2005, <<http://www.rfc-editor.org/info/rfc4303>>.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <<http://www.rfc-editor.org/info/rfc5925>>.

Appendix A. Acknowledgements

The author would like to gratefully acknowledge the IDR WG discussion

Author's Address

Susan Hares
Huawei
7453 Hickory Hill
Saline, MI 48176
USA

Email: shares@ndzh.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 14, 2017

S. Hares
Huawei
March 13, 2017

BGP Registries by IDR and other BGP WGs
draft-hares-idr-bgp-registries-01.txt

Abstract

The BGP Registries at IANA were set up as one of the earliest IANA registries. Over time, the registries have become denoted as requiring "standards action", "early allocation", "FCFS (first-come, first served)", "vendor specific", and "IETF review". This draft proposes that certain BGP registries that are labelled "standards action", "early allocation", or "IETF Review" add to these registration actions a "Expert Review". It also proposes that the chairs of BGP Protocol related WG groups be part of the review team. The intent is that these chairs will be responsible for bringing questionable allocations to their workings attention.

The BGP relate working groups are currently the IDR, BESS, SIDROPS, and GROW, but other working groups like SPRING might be added.

Status of This Memo

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

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

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

This Internet-Draft will expire on September 14, 2017.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
2. BGP Registries to Change Registration Process on	2
3. Security Considerations	4
4. IANA Considerations	5
5. Acknowledgements	5
6. Normative References	5
Author's Address	7

1. Introduction

During 2016, several BGP attributes were squatted upon causing operational problems during the early deployment of large communities [RFC8092]. Due these problems, [RFC8093] deprecated the use of 6 attribute numbers.

To avoid this problem in the future, it is helpful to increase pace of the early-allocations process and to coordinate the review of key BGP registries. This document proposes to augment existing registration processes for BGP registries with Expert review.

This draft proposes that certain BGP registries that are labelled "standards action", "early allocation", or "IETF Review" add to these registration actions a "Expert Review". It also recommends that the chairs of BGP Protocol related WG groups be part of the review team.

2. BGP Registries to Change Registration Process on

This document proposes the that IETF BGP registries in Table 1 below require their current registration policy plus Expert Review. It recommends that the chairs of the BGP related working groups (e.g. IDR, Bess, SIDROPS, GROW) be a part of this review team. The IESG can define which working groups are BGP working groups, but it is important to get the chairs of the Working Groups that originate or maintain the drafts in Table 1 as part of the review team.

If no BGP WG groups remain, the IESG may select designated experts to fulfill this role.

ER = Expert Review

Table 1 - Registries with changes

BGP registry	Registration	reference	Add ER
Message Types	Standards Action	RFC4271	yes
BGP Path Attributes	Standards Action	RFC4271	yes
BGP Error (notification) codes	Standards Action	RFC4271 RFC7313	yes
BGP Error Subcodes	Standards Action	RFC4271	yes
Open Message Error subcodes	Standards Action	RFC4271	yes
Update Message Error subcodes	Standards Action	RFC4271	yes
BGP Finite State Machine Error subcodes	Standards Action	RFC6608	yes
BGP Cease NOTIFICATION message subcodes	Standards Action or Early Allocation	RFC4486	yes
BGP Route Refresh Message Error subcodes	Standards Action (1-127 range)	RFC7313	yes
BGP Outbound Route Filtering (ORF) Types	Standards Action	RFC5291	yes
BGP Open Optional Parameter types	IETF Review	RFC5492	yes
BGP Tunnel Encapsulation Attribute Sub-TLVs	Standards Action	RFC5512	yes
BGP AIGP Attribute	Standards Action	RFC7311	Yes

BGP Tunnel Encapsulation Attribute Sub-TLVs	Standards Action	RFC5512	yes
BGP AIGP Attribute	Standards Action	RFC7311	Yes
Route Refresh Subcdes	Standards Action (1-127)	RFC7313	Yes
P-Multicast Service Interface Tunnel (PMSI) Tunnel Types	IETF Review	RFC7385	Yes
P-Multicast Service Interface Tunnel (PMSI) Attribute Flags	Standards Action	RFC7385	Yes
BGP MCAST-VPN Route Types	Standards Action	RFC7441	Yes

The registries in Table 2 have Expert Review. This document requests that IANA increase their designated expert pool by adding to the pool the chairs in BGP related Working Groups (E.g. IDR, BESS, SIDROPS, GROW).

ER = Expert Review

Table 2 - Registries with Expert Review

BGP registry	Registration	reference	Add ER
BGP Layer 2 Encapsulation Types	Expert Review (0-127)	RFC6624	yes
BGP Layer 2 TLV Types	Expert Review	RFC6624	yes

3. Security Considerations

Administrative process - Not applicable.

4. IANA Considerations

For all of the BGP registries or portions of BGP Registries listed in table 1 append "Designated reviewers" to the registration process.

This document requests the IESG nominate the chairs of the current BGP related working groups which manage the following base protocols that established the registries:

[RFC4271],

[RFC4486],

[RFC5291],

[RFC5492],

[RFC5512],

[RFC6608],

[RFC6624],

[RFC7311],

[RFC7313],

[RFC7385],

[RFC7441],

5. Acknowledgements

The authors would like to thank Alavaro Retana, John Scudder, Jeff Haas, Job Snijders, and members of the IDR and Grow working groups for the active discussion at IETF 97 and post-IETF 97 that inspired this draft.

6. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.

- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4486] Chen, E. and V. Gillet, "Subcodes for BGP Cease Notification Message", RFC 4486, DOI 10.17487/RFC4486, April 2006, <<http://www.rfc-editor.org/info/rfc4486>>.
- [RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <<http://www.rfc-editor.org/info/rfc5291>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <<http://www.rfc-editor.org/info/rfc5512>>.
- [RFC6608] Dong, J., Chen, M., and A. Suryanarayana, "Subcodes for BGP Finite State Machine Error", RFC 6608, DOI 10.17487/RFC6608, May 2012, <<http://www.rfc-editor.org/info/rfc6608>>.
- [RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, DOI 10.17487/RFC6624, May 2012, <<http://www.rfc-editor.org/info/rfc6624>>.
- [RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <<http://www.rfc-editor.org/info/rfc7311>>.
- [RFC7313] Patel, K., Chen, E., and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", RFC 7313, DOI 10.17487/RFC7313, July 2014, <<http://www.rfc-editor.org/info/rfc7313>>.
- [RFC7385] Andersson, L. and G. Swallow, "IANA Registry for P-Multicast Service Interface (PMSI) Tunnel Type Code Points", RFC 7385, DOI 10.17487/RFC7385, October 2014, <<http://www.rfc-editor.org/info/rfc7385>>.

- [RFC7441] Wijnands, IJ., Rosen, E., and U. Joorde, "Encoding Multipoint LDP (mLDP) Forwarding Equivalence Classes (FECs) in the NLRI of BGP MCAST-VPN Routes", RFC 7441, DOI 10.17487/RFC7441, January 2015, <<http://www.rfc-editor.org/info/rfc7441>>.
- [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, <<http://www.rfc-editor.org/info/rfc8092>>.
- [RFC8093] Snijders, J., "Deprecation of BGP Path Attribute Values 30, 31, 129, 241, 242, and 243", RFC 8093, DOI 10.17487/RFC8093, February 2017, <<http://www.rfc-editor.org/info/rfc8093>>.

Author's Address

Susan Hares
Huawei
7453 Hickory Hill
Saline, MI 48176
USA

Email: shares@ndzh.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 7, 2017

K. Patel
A. Vyavaharkar
N. Fazlollahi
Cisco Systems
A. Przygienda
Juniper Networks
March 06, 2017

Extension to BGP's Route Refresh Message
draft-idr-bgp-route-refresh-options-02

Abstract

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

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

Status of This Memo

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

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

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

This Internet-Draft will expire on September 7, 2017.

Copyright Notice

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

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

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

Table of Contents

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

15.2. Information References	13
Appendix A. Sequence Number Binary Arithmetic	14
Authors' Addresses	15

1. Introduction

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

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

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

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

1.1. Use Case Examples

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

i.e., propagation of the refreshes in a way similar to RT Constrain routes is NOT intended.

Several possible use cases are discernible today:

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

2. Requirements Language

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

3. Route Refresh Options Capability

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

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

4. Route Refresh Sub-Types

[RFC7313] defines route refresh BGP message sub-types that utilize the "Reserved" field of the Route Refresh message originally defined in [RFC2918]. Currently, there are three sub-types defined and this draft proposes three additional sub-types which will be used to

indicate a Route Refresh message that includes options before any ORF field of the Route Refresh message as well as BoRR and EoRR Route Refresh messages with options.

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

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

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

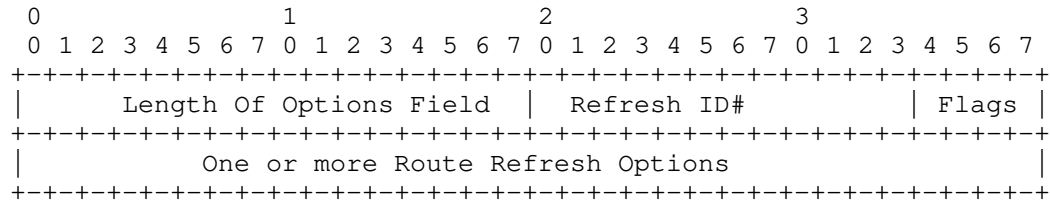
0								1								2								3							
0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
A F I								Res.								S A F I															
Total Option Length																Refresh ID#								Flags							
One or more Route Refresh Options																															

5. Route Refresh Option format

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

As described in the previous section, the options will be added to the Route Refresh message before the ORF field of the message.

Outbound Route Filtering is described in [RFC5291]. The options will assume the following format



6. Route Refresh Option Length

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

7. Route Refresh ID

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

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

LID = MAX(lowest requested Refresh ID# without BoRR,
lowest received BoRR without EoRR)

and

HID = highest requested Refresh ID#

the requesting speaker MUST only use values V where $V \geq \text{LID}$ and $V \geq \text{HID}$ as defined by the relation given in Appendix A. Beside that, $\text{HID} \geq \text{LID}$ MUST hold by the same algebra.

If no such number V exists, LID must catch up to HID, i.e. no further requests can be issued. To use a 3 bit example in Appendix A, if LID was 1 and HID was 4, we cannot progress to unsigned 5 since $1 \geq 5$. When LID progresses to unsigned 2 however, we have $5 \geq 2$ and $5 \geq 4$ and we can choose a V.

Value of 0 MUST NEVER be used as Refresh ID and is considered an "invalid" ID.

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

8. Route Refresh Option Flags

This draft defines several route refresh option flags:

- o 'O'-bit specifies whether the receiving BGP speaker MUST logically OR the attached options or logically AND them (in case of the bit being clear). When the flag is clear, the router on the receiving end SHOULD logically AND the options and only refresh routes that match all received options. If the option flag is set, the router SHOULD select routes that match using a logical OR of the options. In any case the set of routes sent between the according BoRR and EoRR MUST contain at least the logically requested set.
- o 'C' bit indicates that the receiving BGP speaker MUST clear immediately all the received Route Refresh Requests with Options, either pending or being processed. EoRRs MUST NOT be sent. The Refresh ID# on the request MUST be set as the (in unsigned terms) next possible number L for which $\text{LID} \geq L$ and $\text{HID} \geq L$ per Appendix A or in other words we "wrap around the sequence number space" on reset. The C flag MUST NOT be set on BoRR or EoRR messages and CAN be used only with refresh requests.
- o by 'S' bit indicate a refresh is being spontaneously originated by the BGP speaker which received requests and has them pending. The receiving BGP speaker MUST immediately clear all their pending Route Refresh requests with the sending peer. The Refresh ID# on the request MUST be set as the the largest unsigned number L for which $\text{LID} \geq L$ and $\text{HID} \geq L$. When this flag is set, the receiving BGP speaker MUST use this sequence number for its next request. To use example from Appendix A, if the peer received LID 4 and HID 5 (i.e. it didn't send BoRR for 4 yet but received request for 5

already) it will reset the sequence number to 1 by those rules.
 Now, if there is a request with 6 in flight, it will be seen as 1
 >: 6 when arriving.

The precise format is indicated below

```

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

```

C Clear pending requests and reset Refresh ID# space.

O Use logical OR of attached options

S Synchronize sequence numbers

R Reserved bit

9. Route Refresh Options

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

```

    0                               1                               2                               3
    0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7 0 1 2 3 4 5 6 7
  +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
  |   Type   |           Length           |           Value           |
  +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
  |           Value (cont'd).           |
  +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

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

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

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

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

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

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

10. Operation

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

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

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

If a BGP speaker receives the route refresh message with the message subtype set to BoRR with options as described above, then it needs to

process all the included options and MUST mark all matching routes as stale as described in [RFC7313].

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

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

Observe that overlapping refreshes with different options are possible and in such case the according BoRR and EoRR messages are associated by using their Refresh ID#. The BGP speaker responding to the route refresh requests MAY perform the refreshes in parallel. In case of concurrent refreshes overlapping same routes, the responding speaker MUST ensure that the sent advertisements will result in deletion of the omitted routes at the time all EoRRs have been received by the remote speaker or it MUST explicitly advertise withdrawals to correct any anomalies.

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

A BGP speaker may spontaneously originate a refresh to one or more of its peers depending on operator intervention, or due to a policy or configuration change, etc. In such a case, the speaker MUST refresh the entire Adj-RIB-Out. The speaker MUST also send BoRR/EoRR with the options field with the 'S' flag set and a sequence number which lies outside the range of the sequence numbers that are currently in use with the receiving BGP speaker.

11. Error Handling

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

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

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

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

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

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

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

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

EoRR with known Refresh ID# but without preceding BoRR - analogous to EoRR with unknown Refresh ID#. Observe that this can be caused by the peer expiring last resort timer and reusing the ID# for another

request before the EoRR is received. This should be extremely unlikely given the size of the refresh ID space.

12. IANA Considerations

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

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

+74 Route Refresh Options Capability

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

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

13. Security Considerations

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

14. Acknowledgements

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

15. References

15.1. Normative References

- [RFC1982] Elz, R. and R. Bush, "Serial Number Arithmetic", RFC 1982, DOI 10.17487/RFC1982, August 1996, <<http://www.rfc-editor.org/info/rfc1982>>.
- [RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, DOI 10.17487/RFC2918, September 2000, <<http://www.rfc-editor.org/info/rfc2918>>.

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

15.2. Information References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<http://www.rfc-editor.org/info/rfc7752>>.

[Wikipedia]

Wikipedia, "https://en.wikipedia.org/wiki/Serial_number_arithmetic", 2016.

Appendix A. Sequence Number Binary Arithmetic

The only reasonably reference to a cleaner than [RFC1982] sequence number solution is given in [Wikipedia]. It basically converts the problem into two complement's arithmetic. Assuming a straight two complement's subtractions on the bit-width of the sequence number the according $>$: and $=$: relations are defined as:

U_1, U_2 are 12-bits aligned unsigned version number

D_f is $(U_1 - U_2)$ interpreted as two complement signed 12-bits

D_b is $(U_2 - U_1)$ interpreted as two complement signed 12-bits

$U_1 >: U_2$ IIF $D_f > 0$ AND $D_b < 0$

$U_1 =: U_2$ IIF $D_f = 0$

The $>$: relationship is symmetric but not transitive. Observe that this leaves the case of the numbers having maximum two complement distance, e.g. (0 and 0x800) undefined in our 12-bits case since D_f and D_b are both -0x7ff.

A simple example of the relationship in case of 3-bit arithmetic follows as table indicating D_f/D_b values and then the relationship of U_1 to U_2 :

U_2 / U_1	0	1	2	3	4	5	6	7
0	+/+	+/-	+/-	+/-	-/-	-/+	-/+	-/+
1	-/+	+/+	+/-	+/-	+/-	-/-	-/+	-/+
2	-/+	-/+	+/+	+/-	+/-	+/-	-/-	-/+
3	-/+	-/+	-/+	+/+	+/-	+/-	+/-	-/-
4	-/-	-/+	-/+	-/+	+/+	+/-	+/-	+/-
5	+/-	-/-	-/+	-/+	-/+	+/+	+/-	+/-
6	+/-	+/-	-/-	-/+	-/+	-/+	+/+	+/-
7	+/-	+/-	+/-	-/-	-/+	-/+	-/+	+/+

U_2 / U_1	0	1	2	3	4	5	6	7
0	=	>	>	>	?	<	<	<
1	<	=	>	>	>	?	<	<
2	<	<	=	>	>	>	?	<
3	<	<	<	=	>	>	>	?
4	?	<	<	<	=	>	>	>
5	>	?	<	<	<	=	>	>
6	>	>	?	<	<	<	=	>
7	>	>	>	?	<	<	<	=

Authors' Addresses

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

Email: keyupate@cisco.com

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

Email: avyavaha@cisco.com

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

Email: nifazlol@cisco.com

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

Email: prz@juniper.net

IDR
Internet-Draft
Updates: 4684 (if approved)
Intended status: Standards Track
Expires: September 14, 2017

D. Duan
J. Heitz
Cisco
K. Patel
Arrcus
J. Hass
Juniper Networks
March 13, 2017

Persistent Route Oscillation in BGP Constrained Route Distribution
draft-idr-bgp-rt-oscillation-01

Abstract

RFC4684 defines Multi-Protocol BGP (MP-BGP) procedures that allow BGP speakers to exchange Route Target reachability information (RT-Constrain) to restrict the propagation of Virtual Private Network (VPN) routes. In network scenarios where hierarchical route reflection (RR) is used, the existing RT-Constrain mechanism may result in persistent route oscillations within RRs. This document describes the problem and proposes a solution.

Status of This Memo

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

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

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

This Internet-Draft will expire on September 14, 2017.

Copyright Notice

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

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

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

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

Table of Contents

1. Introduction	2
2. Requirements Language	3
3. Problem Statement - Persistent Route Oscillations	3
4. Solution	5
5. IANA Considerations	5
6. Security Considerations	5
7. Acknowledgements	5
8. Normative References	5
Authors' Addresses	6

1. Introduction

[RFC4684] defines Multi-Protocol BGP (MP-BGP) procedures that allow BGP speakers to exchange Route Target reachability information to restrict the propagation of Virtual Private Network (VPN) routes.

[RFC4684] section 3.2 defines a route advertisement rule for Route Target membership information. When advertising an RT membership NLRI to a non-client peer, if the best path as selected by the path selection procedure described in Section 9.1 of [RFC4271] is a route received from a non-client peer, and if there is an alternative path to the same destination from a client peer, then the attributes of the client path are advertised to the peer. [RFC4684] does not clarify which path to choose in case there are multiple client paths to the same destination.

In network scenarios where hierarchical route reflection (RR) is used, and multiple such client paths exist, persistent route oscillations might be formed based on which client path attributes are advertised to the non-client peers.

2. Requirements Language

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

3. Problem Statement - Persistent Route Oscillations

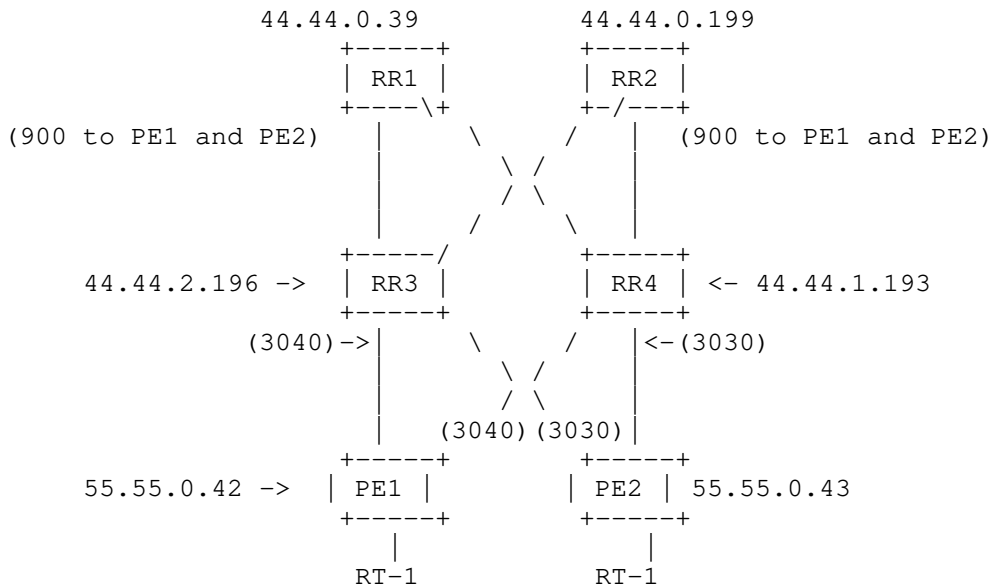


Figure 1. RT-Constraint with Hierarchical Route-reflector

In Figure 1, Hierarchical RRs are deployed. RR3 and RR4 are first level Router Reflectors and RR1 and RR2 are the second level Route Reflectors. Each RR is using its own router-id as its cluster-id. PE1 and PE2 are Route Reflector clients of RR3 and RR4 while RR3 and RR4 are Route Reflector clients of RR1 and RR2. Both PE1 and PE2 are advertising the route-target information RT-1 to the first level Router Reflectors RR3 and RR4. RR3 and RR4 are also advertising route-target information to the second level Router Reflectors RR1 and RR2. The numbers in the parentheses are next-hop metrics.

At step #1, RR3 has two paths for RT-1: one from PE1 and the other from PE2. The path from PE1 has next hop metric 3040 and the path from PE2 has next hop metric 3030. RR3 and RR4 select the path from PE2 as the best path (lower metric). RR3 and RR4 advertise their best paths for RT-1 to the second level router reflectors RR1 and RR2.

At step #2, RR1 has two paths for RT-1: one path from RR3 and the other path from RR4. The next hop metric to reach PE1 and PE2 are both 900. On RR1 and RR2, if both paths have the same ORIGINATOR_ID, then the lower peer address will be used to select the best path. RR1 selects the path from RR4 as best path because it has a lower peer address than RR3. RR1 advertises RT-1 back to RR3.

When announcing RT-1 to its client (RR3), RR1 will set the ORIGINATOR_ID to itself according to [RFC4684] section 3.2.i.

At step #3, RR3 has four paths: the first from PE1, the second from PE2, the third from RR1 and a fourth from RR2. For the purposes of this discussion, the path from RR2 is equivalent to that from RR1. The result is the same if either is chosen. RR3 selects the non-client path from RR1 to PE2 as best path because the ORIGINATOR_ID is lower than that of the paths from PE1 and PE2. Since there are client paths available to reach RT-1, RR3 advertises the path attribute of a client path to RR1 according to [RFC4684] section 3.2.ii. RR3 could choose either the path from PE1 or PE2. RR3 chooses the path attribute of RT-1 from PE1 at random.

At step #4, RR1 receives the updates and recalculates the best path. RR1 has a path from RR4 with ORIGINATOR_ID set to PE2's router-id and a path from RR3 with ORIGINATOR_ID set to PE1's router-id. RR1 selects the path from RR3 as best path because of lower ORIGINATOR_ID. RR1 sets the ORIGINATOR_ID to its own router-id and sends it back to RR3.

At step #5, RR3 receives the updates from RR1 and drops the updates since its own cluster-id is in the cluster list. Now RR3's routing state goes back to that at step #1 with 2 paths from its clients and the whole cycle starts again.

The same thing happens on RR4 as on RR3 and the same thing happens on RR2 as on RR1.

These iterations results in a persistent route oscillation for RT-1 prefix of RT-Constrain address-family on RR1, RR2, RR3 and RR4.

4. Solution

The solution is for the Route Reflector always to prefer the client paths when selecting a best path. This preference MUST be expressed before step f) of the BGP Decision Tie Breaking rules in Section 9.1.2.2 of [RFC4271]. It MAY be expressed at a higher step. So at Step #3 on RR3, the best path is still the path from PE2. The oscillation terminates with PE2's path.

Note that the scenario can not happen if RR1 and RR2 are in the same cluster. So at step #3, RR3 only has two client paths. The update from the top level Route Reflector will be dropped because of the cluster id check. The oscillation never happens with such a topology.

5. IANA Considerations

This draft makes no request of IANA.

6. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4271] and [RFC4684].

7. Acknowledgements

The authors would like to thank Shyam Sethuram, Nitin Kumar, Sameer Gulrajani, Mohammed Mirza and Mike Dubrovskiy.

8. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <<http://www.rfc-editor.org/info/rfc4684>>.

Authors' Addresses

Dongling Duan
Cisco
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: duan@cisco.com

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

Email: jheitz@cisco.com

Keyur Patel
Arrcus, Inc

Email: keyur@arrcus.com

Jeffrey Hass
Juniper Networks
1194 N. Mathida Ave
Sunnyvale, CA 94089
USA

Email: jhaas@juniper.net

IDR
Internet-Draft
Intended status: Standards Track
Expires: June 15, 2017

S. Previdi, Ed.
C. Filsfils
A. Lindem
K. Patel
A. Sreekantiah
Cisco Systems
S. Ray
Unaffiliated
H. Gredler
RtBrick Inc.
December 12, 2016

Segment Routing Prefix SID extensions for BGP
draft-ietf-idr-bgp-prefix-sid-04

Abstract

Segment Routing (SR) architecture allows a node to steer a packet flow through any topological path and service chain by leveraging source routing. The ingress node prepends a SR header to a packet containing a set of "segments". Each segment represents a topological or a service-based instruction. Per-flow state is maintained only at the ingress node of the SR domain.

This document describes the BGP extension for announcing BGP Prefix Segment Identifier (BGP Prefix SID) information.

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] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without any normative meaning.

Status of This Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on June 15, 2017.

Copyright Notice

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

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

Table of Contents

1. Segment Routing Documents	3
2. Introduction	3
3. BGP-Prefix-SID	4
3.1. MPLS Prefix Segment	4
3.2. IPv6 Prefix Segment	5
4. BGP-Prefix-SID Attribute	5
4.1. Label-Index TLV	6
4.2. IPv6 SID	7
4.3. Originator SRGB TLV	8
5. Receiving BGP-Prefix-SID Attribute	9
5.1. MPLS Dataplane: Labeled Unicast	9
5.2. IPv6 Dataplane	10
6. Announcing BGP-Prefix-SID Attribute	10
6.1. MPLS Dataplane: Labeled Unicast	11
6.2. IPv6 Dataplane	11
7. Error Handling of BGP-Prefix-SID Attribute	12
8. IANA Considerations	12
9. Security Considerations	13
10. Acknowledgements	13
11. Change Log	13
12. References	13
12.1. Normative References	13
12.2. Informative References	13
Authors' Addresses	15

1. Segment Routing Documents

The main references for this document are the SR architecture defined in [I-D.ietf-spring-segment-routing] and the related use case illustrated in [I-D.ietf-spring-segment-routing-msdc].

The Segment Routing Egress Peer Engineering architecture is described in [I-D.ietf-spring-segment-routing-central-epe].

The Segment Routing Egress Peer Engineering BGPLS extensions are described in [I-D.ietf-idr-bgpls-segment-routing-epe].

2. Introduction

Segment Routing (SR) architecture leverages the source routing paradigm. A group of inter-connected nodes that use SR forms a SR domain. The ingress node of the SR domain prepends a SR header containing "segments" to an incoming packet. Each segment represents a topological instruction such as "go to prefix P following shortest path" or a service instruction (e.g.: "pass through deep packet inspection"). By inserting the desired sequence of instructions, the ingress node is able to steer a packet via any topological path and/or service chain; per-flow state is maintained only at the ingress node of the SR domain.

Each segment is identified by a Segment Identifier (SID). As described in [I-D.ietf-spring-segment-routing], when SR is applied to the MPLS dataplane the SID consists of a label while when SR is applied to the IPv6 dataplane the SID consists of an IPv6 prefix (see [I-D.ietf-6man-segment-routing-header]).

A BGP-Prefix Segment (aka BGP-Prefix-SID), is a BGP segment attached to a BGP prefix. A BGP-Prefix-SID is always global within the SR/BGP domain and identifies an instruction to forward the packet over the ECMP-aware best-path computed by BGP to the related prefix. The BGP-Prefix-SID is the identifier of the BGP prefix segment.

This document describes the BGP extension to signal the BGP-Prefix-SID. Specifically, this document defines a new BGP attribute known as the BGP Prefix SID attribute and specifies the rules to originate, receive and handle error conditions of the new attribute.

As described in [I-D.ietf-spring-segment-routing-msdc], the newly proposed BGP Prefix-SID attribute can be attached to prefixes from AFI/SAFI:

Multiprotocol BGP labeled IPv4/IPv6 Unicast ([RFC3107]).

Multiprotocol BGP ([RFC4760]) unlabeled IPv6 Unicast.

[I-D.ietf-spring-segment-routing-msdc] describes use cases where the Prefix-SID is used for the above AFI/SAFI.

3. BGP-Prefix-SID

The BGP-Prefix-SID attached to a BGP prefix P represents the instruction "go to Prefix P" along its BGP bestpath (potentially ECMP-enabled).

3.1. MPLS Prefix Segment

The BGP Prefix Segment is realized on the MPLS dataplane in the following way:

As described in [I-D.ietf-spring-segment-routing-msdc] the operator assigns a globally unique "index", L_I, to a locally sourced prefix of a BGP speaker N which is advertised to all other BGP speakers in the SR domain.

According to [I-D.ietf-spring-segment-routing], each BGP speaker is configured with a label block called the Segment Routing Global Block (SRGB). While it is recommended to use the same SRGB across all the nodes within the SR domain, the SRGB of a node is a local property and could be different on different speakers. The drawbacks of the use case where BGP speakers have different SRGBs are documented in [I-D.ietf-spring-segment-routing] and [I-D.ietf-spring-segment-routing-msdc].

If traffic-engineering within the SR domain is required, each node may also be required to advertise topological information and Peering SID's for each of its links and peers. This information is required in order to perform the explicit path computation and to express any explicit path into a list of segments. The advertisement of topological information and Peer segments is assumed to be done through [I-D.ietf-idr-bgpls-segment-routing-epe].

If the BGP speakers are not all configured with the same SRGB, and if traffic-engineering within the SR domain is required, each node may be required to advertise its local SRGB in addition to the topological information.

This document assumes that BGP-LS is the preferred method for collecting both topological, peer segments and SRGB information through [RFC7752], [I-D.ietf-idr-bgpls-segment-routing-epe] and [I-D.ietf-idr-bgp-ls-segment-routing-ext]. However, as an

optional alternative for the advertisement of the local SRGB without the topology nor the peer SID's, hence without applicability for TE, the Originator SRGB TLV of the prefix-SID attribute, is specified in Section 4.3 of this document.

The index `L_I` is a 32 bit offset in the SRGB. Each BGP speaker derives its local MPLS label, `L`, by adding `L_I` to the start value of its own SRGB, and programs `L` in its MPLS dataplane as its incoming/local label for the prefix. See Section 5.1 for more details.

The outgoing label for the prefix is found in the NLRI of the Multiprotocol BGP labeled IPv4/IPv6 Unicast prefix advertisement. The index `L_I` is only used as a hint to derive the local/incoming label.

Section 4.1 of this document specifies the Label-Index TLV of the BGP Prefix-SID attribute; this TLV can be used to advertise the label index of a given prefix.

In order to advertise the label index of a given prefix `P` and, optionally, the SRGB, a new extension to BGP is needed: the BGP Prefix SID attribute. This extension is described in subsequent sections.

3.2. IPv6 Prefix Segment

As defined in [I-D.ietf-6man-segment-routing-header], and as illustrated in [I-D.ietf-spring-segment-routing-msdc], when SR is used over an IPv6 dataplane, the BGP Prefix Segment is instantiated by an IPv6 prefix originated by the BGP speaker.

Each node advertises a globally unique IPv6 address representing itself in the domain. This prefix (e.g.: its loopback interface address) is advertised to all other BGP speakers in the SR domain.

Also, each node MUST advertise its support of Segment Routing for IPv6 dataplane. This is realized using the flags contained in the Prefix SID Attribute defined below.

4. BGP-Prefix-SID Attribute

The BGP Prefix SID attribute is an optional, transitive BGP path attribute. The attribute type code is to be assigned by IANA (suggested value: 40). The value field of the BGP-Prefix-SID attribute has the following format:

The value field of the BGP Prefix SID attribute is defined here to be a set of elements encoded as "Type/Length/Value" (i.e., a set of TLVs). Following TLVs are defined:

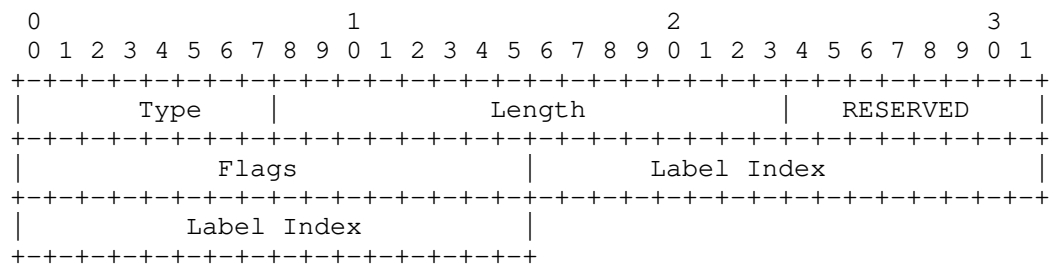
- o Label-Index TLV
- o IPv6 SID TLV
- o Originator SRGB TLV

Label-Index and Originator SRGB TLVs are used only when SR is applied to the MPLS dataplane.

IPv6 SID TLV is used only when SR is applied to the IPv6 dataplane.

4.1. Label-Index TLV

The Label-Index TLV MUST be present in the Prefix-SID attribute attached to Labeled IPv4/IPv6 unicast prefixes ([RFC3107]) and has the following format:

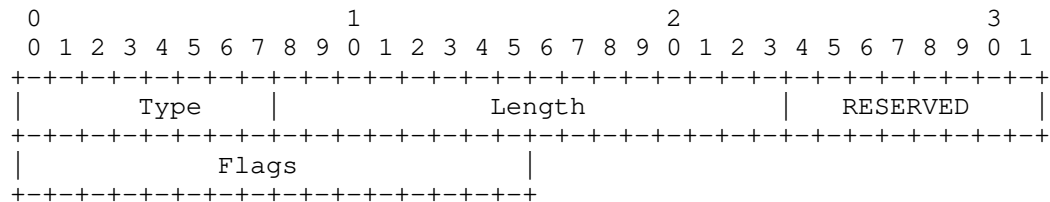


where:

- o Type is 1.
- o Length: is 7, the total length of the value portion of the TLV.
- o RESERVED: 8 bit field. SHOULD be 0 on transmission and MUST be ignored on reception.
- o Flags: 16 bits of flags. None are defined at this stage of the document. The flag field SHOULD be clear on transmission and MUST be ignored at reception.
- o Label Index: 32 bit value representing the index value in the SRGB space.

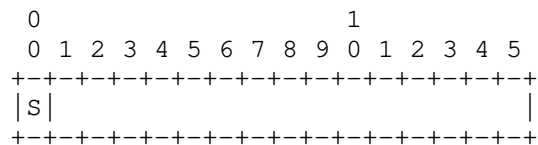
4.2. IPv6 SID

The IPv6-SID TLV MUST be present in the Prefix-SID attribute attached to MP-BGP unlabeled IPv6 unicast prefixes ([RFC4760]) and has the following format:



where:

- o Type is 2.
- o Length: is 3, the total length of the value portion of the TLV.
- o RESERVED: 8 bit field. SHOULD be 0 on transmission and MUST be ignored on reception.
- o Flags: 16 bits of flags defined as follow:



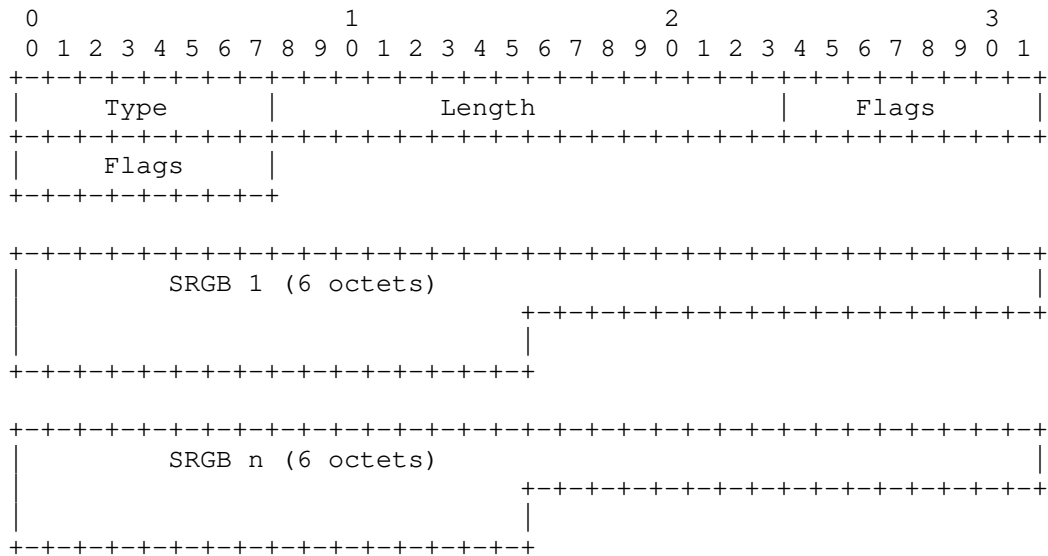
where:

- * S flag: if set then it means that the BGP speaker attaching the Prefix-SID Attribute to a prefix is capable of processing the IPv6 Segment Routing Header (SRH, [I-D.ietf-6man-segment-routing-header]) for the segment corresponding to the originated IPv6 prefix. The use case leveraging the S flag is described in [I-D.ietf-spring-segment-routing-msdc].

The other bits of the flag field SHOULD be clear on transmission and MUST be ignored at reception.

4.3. Originator SRGB TLV

The Originator SRGB TLV is an optional TLV and has the following format:



where:

- o Type is 3.
- o Length is the total length of the value portion of the TLV: 2 + multiple of 6.
- o Flags: 16 bits of flags. None are defined in this document. Flags SHOULD be clear on transmission and MUST be ignored at reception.
- o SRGB: 3 octets of base followed by 3 octets of range. Note that the SRGB field MAY appear multiple times. If the SRGB field appears multiple times, the SRGB consists of multiple ranges. The meaning of an SRGB with multiple ranges is explained in Section 3.2 ("SID/Label Range TLV") of [I-D.ietf-ospf-segment-routing-extensions].

The Originator SRGB TLV contains the SRGB of the router originating the prefix to which the BGP Prefix SID is attached and MUST be kept in the Prefix-SID Attribute unchanged during the propagation of the BGP update.

The originator SRGB describes the SRGB of the node where the BGP Prefix Segment end. It is used to build SRTE policies when different SRGB's are used in the fabric ([I-D.ietf-spring-segment-routing-msdc]).

The originator SRGB may only appear on Prefix-SID attribute attached to prefixes of SAFI 4 (labeled unicast, [RFC3107]).

5. Receiving BGP-Prefix-SID Attribute

A BGP speaker receiving a BGP Prefix-SID attribute from an EBGp neighbor residing outside the boundaries of the SR domain, SHOULD discard the attribute unless it is configured to accept the attribute from the EBGp neighbor. A BGP speaker MAY log an error for further analysis when discarding an attribute.

5.1. MPLS Dataplane: Labeled Unicast

A Multiprotocol BGP labeled IPv4/IPv6 Unicast ([RFC3107]) session type is required.

A BGP speaker may be locally configured with an SRGB=[GB_S, GB_E]. The preferred method for deriving the SRGB is a matter of local router configuration.

Given a label index L_I , we call $L = L_I + GB_S$ as the derived label. A BGP Prefix-SID attribute is called "unacceptable" for a speaker M if the derived label value L lies outside the SRGB configured on M. Otherwise the Label Index attribute is called "acceptable" to speaker M.

The mechanisms through which a given label_index value is assigned to a given prefix are outside the scope of this document. The label-index value associated with a prefix is locally configured at the BGP router originating the prefix.

The Prefix-SID attribute MUST contain the Label-Index TLV and MAY contain the Originator SRGB TLV. A BGP Prefix-SID attribute received without a Label-Index TLV MUST be considered as "unacceptable" by the receiving speaker.

When a BGP speaker receives a path from a neighbor with an acceptable BGP Prefix-SID attribute, it MUST program the derived label as the local label for the prefix in its MPLS dataplane. In case of any error, a BGP speaker MUST resort to the error handling rules specified in Section 7. A BGP speaker MAY log an error for further analysis.

When a BGP speaker receives a path from a neighbor with an unacceptable BGP Prefix-SID attribute or when a BGP speaker receives a path from a neighbor with a BGP-Prefix-SID attribute but is unable to process it (it does not have the capability or local policy disables the capability), it MUST treat the path as if it came without a Prefix-SID attribute. For the purposes of local label allocation, a BGP speaker MUST assign a local (also called dynamic) label (non-SRGB) for such a prefix as per classic Multiprotocol BGP labeled IPv4/IPv6 Unicast ([RFC3107]) operation. A BGP speaker MAY log an error for further analysis.

The outgoing label is always programmed as per classic Multiprotocol BGP labeled IPv4/IPv6 Unicast (RFC3107 [RFC3107]) operation.

Specifically, a BGP speaker receiving a prefix with a Prefix-SID attribute and a label NLRI field of implicit-null from a neighbor MUST adhere to standard behavior and program its MPLS dataplane to pop the top label when forwarding traffic to the prefix. The label NLRI defines the outbound label that MUST be used by the receiving node. The Label Index gives a hint to the receiving node on which local/incoming label the BGP speaker SHOULD use.

5.2. IPv6 Dataplane

When a SR IPv6 BGP speaker receives a IPv6 Unicast BGP Update with a prefix having the BGP Prefix SID attribute attached, it checks whether the IPv6 SID TLV is present and if the S-flag is set. If the IPv6 SID TLV is present and if the S-flag is not set, then the Prefix-SID attribute MUST be considered as "unacceptable" by the receiving speaker.

The Originator SRGB MUST be ignored on reception.

A BGP speaker receiving a BGP Prefix-SID attribute from an EBGp neighbor residing outside the boundaries of the SR domain, SHOULD discard the attribute unless it is configured to accept the attribute from the EBGp neighbor. A BGP speaker MAY log an error for further analysis when discarding an attribute.

6. Announcing BGP-Prefix-SID Attribute

The BGP Prefix-SID attribute MAY be attached to labeled BGP prefixes (IPv4/IPv6) [RFC3107] or to IPv6 prefixes [RFC4760]. In order to prevent distribution of the BGP Prefix-SID attribute beyond its intended scope of applicability, attribute filtering MAY be deployed.

6.1. MPLS Dataplane: Labeled Unicast

A BGP speaker that originates a prefix attaches the Prefix-SID attribute when it advertises the prefix to its neighbors via Multiprotocol BGP labeled IPv4/IPv6 Unicast ([RFC3107]). The value of the Label-Index in the Label-Index TLV is determined by configuration.

A BGP speaker that originates a Prefix-SID attribute MAY optionally announce Originator SRGB TLV along with the mandatory Label-Index TLV. The content of the Originator SRGB TLV is determined by the configuration.

Since the Label-index value must be unique within an SR domain, by default an implementation SHOULD NOT advertise the BGP Prefix-SID attribute outside an Autonomous System unless it is explicitly configured to do so.

A BGP speaker that advertises a path received from one of its neighbors SHOULD advertise the Prefix-SID received with the path without modification regardless of whether the Prefix-SID was acceptable. If the path did not come with a Prefix-SID attribute, the speaker MAY attach a Prefix-SID to the path if configured to do so. The content of the TLVs present in the Prefix-SID is determined by the configuration.

In all cases, the label field of the advertised NLRI ([RFC3107], [RFC4364]) MUST be set to the local/incoming label programmed in the MPLS dataplane for the given advertised prefix. If the prefix is associated with one of the BGP speakers interfaces, this label is the usual MPLS label (such as the implicit or explicit NULL label).

6.2. IPv6 Dataplane

A BGP speaker that originates a prefix attaches the Prefix-SID attribute when it advertises the prefix to its neighbors. The IPv6 SID TLV MUST be present and the S-flag MUST be set.

A BGP speaker that advertises a path received from one of its neighbors SHOULD advertise the Prefix-SID received with the path without modification regardless of whether the Prefix-SID was acceptable. If the path did not come with a Prefix-SID attribute, the speaker MAY attach a Prefix-SID to the path if configured to do so. The IPv6-SID TLV MUST be present in the Prefix-SID and with the S-flag set.

7. Error Handling of BGP-Prefix-SID Attribute

When a BGP Speaker receives a BGP Update message containing a malformed BGP Prefix-SID attribute, it MUST ignore the received BGP Prefix-SID attributes and not pass it to other BGP peers. This is equivalent to the -attribute discard- action specified in [RFC7606]. When discarding an attribute, a BGP speaker MAY log an error for further analysis.

If the BGP Prefix-SID attribute appears more than once in an BGP Update message, then, according to [RFC7606], all the occurrences of the attribute other than the first one SHALL be discarded and the BGP Update message shall continue to be processed.

When a BGP speaker receives an unacceptable Prefix-SID attribute, it MAY log an error for further analysis.

8. IANA Considerations

This document defines a new BGP path attribute known as the BGP Prefix-SID attribute. This document requests IANA to assign a new attribute code type (suggested value: 40) for BGP the Prefix-SID attribute from the BGP Path Attributes registry.

Currently, IANA temporarily assigned the following:

40 BGP Prefix-SID (TEMPORARY - registered 2015-09-30, expires 2016-09-30) [draft-ietf-idr-bgp-prefix-sid]

This document defines 3 new TLVs for BGP Prefix-SID attribute. These TLVs need to be registered with IANA. We request IANA to create a new registry for BGP Prefix-SID Attribute TLVs as follows:

Under "Border Gateway Protocol (BGP) Parameters" registry, "BGP Prefix SID attribute Types" Reference: draft-ietf-idr-bgp-prefix-sid Registration Procedure(s): Values 1-254 First Come, First Served, Value 0 and 255 reserved

Value	Type	Reference
0	Reserved	this document
1	Label-Index	this document
2	IPv6 SID	this document
3	Originator SRGB	this document
4-254	Unassigned	
255	Reserved	this document

9. Security Considerations

This document introduces no new security considerations above and beyond those already specified in [RFC4271] and [RFC3107].

10. Acknowledgements

The authors would like to thanks Satya Mohanty for his contribution to this document.

11. Change Log

Initial Version: Sep 21 2014

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <<http://www.rfc-editor.org/info/rfc3107>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<http://www.rfc-editor.org/info/rfc4364>>.
- [RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <<http://www.rfc-editor.org/info/rfc7606>>.

12.2. Informative References

- [I-D.ietf-6man-segment-routing-header]
Previdi, S., Filsfils, C., Field, B., Leung, I., Linkova, J., Aries, E., Kosugi, T., Vyncke, E., and D. Lebrun, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-02 (work in progress), September 2016.
- [I-D.ietf-idr-bgp-ls-segment-routing-ext]
Previdi, S., Psenak, P., Filsfils, C., Gredler, H., Chen, M., and j. jefftant@gmail.com, "BGP Link-State extensions for Segment Routing", draft-ietf-idr-bgp-ls-segment-routing-ext-00 (work in progress), November 2016.
- [I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Filsfils, C., Ray, S., Patel, K., Dong, J., and M. Chen, "Segment Routing BGP Egress Peer Engineering BGP-LS Extensions", draft-ietf-idr-bgpls-segment-routing-epe-06 (work in progress), November 2016.
- [I-D.ietf-ospf-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", draft-ietf-ospf-segment-routing-extensions-10 (work in progress), October 2016.
- [I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-10 (work in progress), November 2016.
- [I-D.ietf-spring-segment-routing-central-epe]
Filsfils, C., Previdi, S., Aries, E., and D. Afanasiev, "Segment Routing Centralized BGP Peer Engineering", draft-ietf-spring-segment-routing-central-epe-03 (work in progress), November 2016.
- [I-D.ietf-spring-segment-routing-msdc]
Filsfils, C., Previdi, S., Mitchell, J., Aries, E., and P. Lapukhov, "BGP-Prefix Segment in large-scale data centers", draft-ietf-spring-segment-routing-msdc-02 (work in progress), October 2016.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<http://www.rfc-editor.org/info/rfc4760>>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<http://www.rfc-editor.org/info/rfc7752>>.

Authors' Addresses

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

Email: sprevidi@cisco.com

Clarence Filsfils
Cisco Systems
Brussels
Belgium

Email: cfilsfils@cisco.com

Acee Lindem
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124 95134
USA

Email: acee@cisco.com

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124 95134
USA

Email: keyupate@cisco.com

Arjun Sreekantiah
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95124 95134
USA

Email: asreekan@cisco.com

Saikat Ray
Unaffiliated

Email: raysaikat@gmail.com

Hannes Gredler
RtBrick Inc.

Email: hannes@rtbrick.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 14, 2017

S. Previdi, Ed.
C. Filsfils
Cisco Systems, Inc.
K. Patel
Arrcus, Inc.
S. Ray
Individual Contributor
J. Dong
M. Chen
Huawei Technologies
March 13, 2017

Segment Routing BGP Egress Peer Engineering BGP-LS Extensions
draft-ietf-idr-bgpls-segment-routing-epe-11

Abstract

Segment Routing (SR) leverages source routing. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header. A segment can represent any instruction, topological or service-based. SR allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node of the SR domain.

The Segment Routing architecture can be directly applied to the MPLS dataplane with no change on the forwarding plane. It requires minor extension to the existing link-state routing protocols.

This document outline a BGP-LS extension 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.

Requirements Language

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

Status of This Memo

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

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

working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on September 14, 2017.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
2. Segment Routing Documents	3
3. BGP Peering Segments	4
4. Link NLRI for BGP-EPE Connectivity Description	5
4.1. BGP Router ID and Member ASN	5
4.2. BGP-EPE Node Descriptors	6
4.3. Link Attributes	7
5. Peer Node and Peer Adjacency Segments	9
5.1. Peer Node Segment (Peer-Node-SID)	9
5.2. Peer Adjacency Segment (Peer-Adj-SID)	10
5.3. Peer Set Segment	11
6. Illustration	12
6.1. Reference Diagram	12
6.2. Peer Node Segment for Node D	14
6.3. Peer Node Segment for Node F	14
6.4. Peer Node Segment for Node E	14
6.5. Peer Adjacency Segment for Node E, Link 1	15
6.6. Peer Adjacency Segment for Node E, Link 2	15
7. Implementation Status	16
8. IANA Considerations	17
8.1. New BGP-LS Protocol-ID	17

8.2. Node Descriptors and Link Attribute TLVs	17
9. Manageability Considerations	18
10. Security Considerations	18
11. Contributors	18
12. Acknowledgements	18
13. References	19
13.1. Normative References	19
13.2. Informative References	19
Authors' Addresses	20

1. Introduction

Segment Routing (SR) leverages source routing. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header. A segment can represent any instruction, topological or service-based. SR allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node of the SR domain.

The Segment Routing architecture can be directly applied to the MPLS dataplane with no change on the forwarding plane. It requires minor extension to the existing link-state routing protocols.

This document outline a BGP-LS extension 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 Egress Peer Engineering (BGP-EPE) policies and strategies.

This document defines new types of segments: a Peer Node segment describing the BGP session between two nodes; a Peer Adjacency Segment describing the link (one or more) that is used by the BGP session; the Peer Set Segment describing an arbitrary set of sessions or links between the local BGP node and its peers.

While an egress point topology usually refers to eBGP sessions between external peers, there's nothing in the extensions defined in this document that would prevent the use of these extensions in the context of iBGP sessions.

2. Segment Routing Documents

The main reference for this document is the SR architecture defined in [I-D.ietf-spring-segment-routing].

The Segment Routing BGP Egress Peer Engineering (BGP-EPE) architecture is described in [I-D.ietf-spring-segment-routing-central-epe].

3. BGP Peering Segments

As defined in [I-D.ietf-spring-segment-routing-central-epe], an BGP-EPE enabled Egress PE node MAY advertise segments corresponding to its attached peers. These segments are called BGP peering segments or BGP Peering SIDs. In case of eBGP, they enable the expression of source-routed inter-domain paths.

An ingress border router of an AS may compose a list of segments to steer a flow along a selected path within the AS, towards a selected egress border router C of the AS and through a specific peer. At minimum, a BGP-EPE policy applied at an ingress PE involves two segments: the Node SID of the chosen egress PE and then the BGP Peering Segment for the chosen egress PE peer or peering interface.

This document defines the BGP-EPE Peering Segments:

- o Peer Node Segment (Peer-Node-SID)
- o Peer Adjacency Segment (Peer-Adj-SID)
- o Peer Set Segment (Peer-Set-SID)

Each BGP session MUST be described by a Peer Node Segment. The description of the BGP session MAY be augmented by additional Adjacency Segments. Finally, each Peer Node Segment and Peer Adjacency Segment MAY be part of the same group/set so to be able to group EPE resources under a common Peer-Set Segment Identifier (SID).

Therefore, when the extensions defined in this document are applied to the use case defined in [I-D.ietf-spring-segment-routing-central-epe]:

- o One Peer Node Segment MUST be present.
- o One or more Peer Adjacency Segments MAY be present.
- o Each of the Peer Node and Peer Adjacency Segment MAY use the same Peer-Set.

While an egress point topology usually refers to eBGP sessions between external peers, there's nothing in the extensions defined in this document that would prevent the use of these extensions in the context of iBGP sessions.

4. Link NLRI for BGP-EPE Connectivity Description

This section describes the NLRI used for describing the connectivity of the BGP Egress router. The connectivity is based on links and remote peers/ASs and therefore the existing Link-Type NLRI (defined in [RFC7752]) is used. A new Protocol-ID is used: BGP (codepoint 7 assigned by IANA (Section 8) from the registry "BGP-LS Protocol-IDs").

The use of a new Protocol-ID allows separation and differentiation between the NLRIs carrying BGP-EPE descriptors from the NLRIs carrying IGP link-state information as defined in [RFC7752]. The Link NLRI Type uses descriptors and attributes already defined in [RFC7752] in addition to new TLVs defined in the following sections of this document.

The extensions defined in this document apply to both internal and external BGP-LS EPE advertisements.

[RFC7752] defines Link NLRI Type is as follows:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+
| Protocol-ID |
+-----+-----+-----+-----+
|                               Identifier                               |
|                               (64 bits)                               |
+-----+-----+-----+-----+
//      Local Node Descriptors      //
+-----+-----+-----+-----+
//      Remote Node Descriptors     //
+-----+-----+-----+-----+
//      Link Descriptors             //
+-----+-----+-----+-----+

```

Node Descriptors and Link Descriptors are defined in [RFC7752].

4.1. BGP Router ID and Member ASN

Two new Node Descriptors Sub-TLVs are defined in this document:

- o BGP Router Identifier (BGP Router-ID):

Type: 516 (assigned by IANA (Section 8) from the registry "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs").

Length: 4 octets

Value: 4 octet unsigned integer representing the BGP Identifier as defined in [RFC4271] and [RFC6286].

- o Confederation Member ASN (Member-ASN)

Type: 517 (assigned by IANA (Section 8) from the registry "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs").

Length: 4 octets

Value: 4 octet unsigned integer representing the Member ASN inside the Confederation. [RFC5065].

4.2. BGP-EPE Node Descriptors

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

- o BGP Router-ID, which contains the BGP Identifier of the local BGP-EPE capable node.
- o Autonomous System Number, which contains the local ASN or local confederation identifier (ASN) if confederations are used.
- o BGP-LS Identifier.

It has to be noted that [RFC6286] (section 2.1) requires the BGP identifier (router-id) to be unique within an Autonomous System. Therefore, the <ASN, BGP identifier> tuple is globally unique.

The following Node Descriptors Sub-TLVs MAY appear in the Link NLRI as Local Node Descriptors:

- o Member-ASN, which contains the ASN of the confederation member (when BGP confederations are used).
- o Node Descriptors as defined in [RFC7752].

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

- o BGP Router-ID, which contains the BGP Identifier of the peer node.
- o Autonomous System Number, which contains the peer ASN or the peer confederation identifier (ASN), if confederations are used.

The following Node Descriptors Sub-TLVs MAY appear in the Link NLRI as Remote Node Descriptors:

- o Member-ASN, which contains the ASN of the confederation member (when BGP confederations are used).
- o Node Descriptors as defined in defined in [RFC7752].

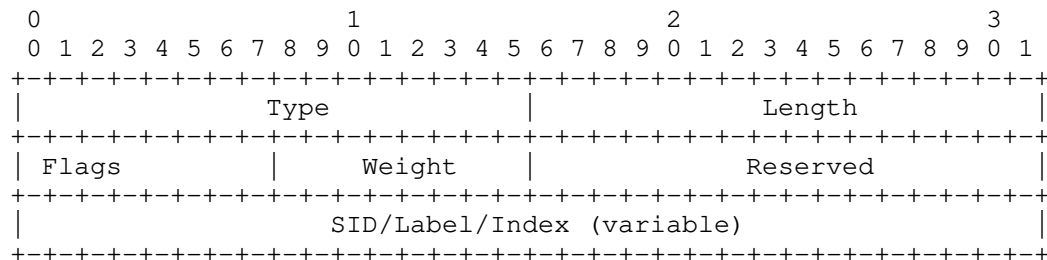
4.3. Link Attributes

The following BGP-LS Link attributes TLVs are used with the Link NLRI:

TLV Code Point	Description	Length
1101	Peer Node Segment Identifier (Peer-Node-SID)	variable
1102	Peer Adjacency Segment Identifier (Peer-Adj-SID)	variable
1103	Peer Set Segment Identifier (Peer-Set-SID)	variable

Figure 1: BGP-LS TLV code points for BGP-EPE

Peer-Node-SID, Peer-Adj-SID and Peer-Set-SID have all the same format defined here below:



where:

Figure 2

- o Type: 1101 or 1102 or 1103 (assigned by IANA (Section 8) from the registry "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs").

- o Length: variable.
- o Flags: following flags have been defined:

```

 0 1 2 3 4 5 6 7
+---+---+---+---+
|V|L|           |
+---+---+---+---+

```

where:

- * V-Flag: Value flag. If set, then the Adj-SID carries a value. By default the flag is SET.
 - * L-Flag: Local Flag. If set, then the value/index carried by the Adj-SID has local significance. By default the flag is SET.
 - * Other bits: MUST be zero when originated and ignored when received.
- o Weight: 1 octet. The value represents the weight of the SID for the purpose of load balancing. An example use of the weight is described in [I-D.ietf-spring-segment-routing].
 - o SID/Index/Label. According to the TLV length and to the V and L flags settings, it contains either:
 - * A 3 octet local label where the 20 rightmost bits are used for encoding the label value. In this case the V and L flags MUST be set.
 - * A 4 octet index defining the offset in the SRGB (Segment Routing Global Block as defined in [I-D.ietf-spring-segment-routing] advertised by this router. In this case the SRGB MUST be advertised using the extensions defined in [I-D.ietf-idr-bgp-ls-segment-routing-ext].
 - * A 16 octet IPv6 address. In this case the V flag MUST be set. The L flag MUST be unset if the IPv6 address is globally unique.

The values of the Peer-Node-SID, Peer-Adj-SID and Peer-Set-SID Sub-TLVs SHOULD be persistent across router restart.

The Peer-Node-SID MUST be present when BGP-LS is used for the use case described in [I-D.ietf-spring-segment-routing-central-epe] and MAY be omitted for other use cases.

The Peer-Adj-SID and Peer-Set-SID SubTLVs MAY be present when BGP-LS is used for the use case described in [I-D.ietf-spring-segment-routing-central-epe] and MAY be omitted for other use cases.

In addition, BGP-LS Nodes and Link Attributes, as defined in [RFC7752] MAY be inserted in order to advertise the characteristics of the link.

5. Peer Node and Peer Adjacency Segments

In this section the following Peer Segments are defined:

Peer Node Segment (Peer-Node-SID)

Peer Adjacency Segment (Peer-Adj-SID)

Peer Set Segment (Peer-Set-SID)

The Peer Node, Peer Adjacency and Peer Set segments can be either a local or a global segment (depending on the setting of the V and L flags defined in Figure 2. For example, when BGP-EPE is used in the context of a SR network over the IPv6 dataplane, it is likely the case that the IPv6 addresses used as SIDs will be global.

5.1. Peer Node Segment (Peer-Node-SID)

The Peer Node Segment describes the BGP session peer (neighbor). It MUST be present when describing a BGP-EPE topology as defined in [I-D.ietf-spring-segment-routing-central-epe]. The Peer Node Segment is encoded within the BGP-LS Link NLRI specified in Section 4.

The Peer Node Segment, at the BGP node advertising it, has the following semantic:

- o SR header operation: NEXT (as defined in [I-D.ietf-spring-segment-routing]).
- o Next-Hop: the connected peering node to which the segment is related.

The Peer Node Segment is advertised with a Link NLRI, where:

- o Local Node Descriptors contains

Local BGP Router-ID of the BGP-EPE enabled egress PE.
Local ASN.
BGP-LS Identifier.

- o Remote Node Descriptors contains
 - Peer BGP Router-ID (i.e.: the peer BGP ID used in the BGP session).
 - Peer ASN.
- o Link Descriptors Sub-TLVs, as defined in [RFC7752], contain the addresses used by the BGP session:
 - * IPv4 Interface Address (Sub-TLV 259) contains the BGP session IPv4 local address.
 - * IPv4 Neighbor Address (Sub-TLV 260) contains the BGP session IPv4 peer address.
 - * IPv6 Interface Address (Sub-TLV 261) contains the BGP session IPv6 local address.
 - * IPv6 Neighbor Address (Sub-TLV 262) contains the BGP session IPv6 peer address.
- o Link Attribute contains the Peer-Node-SID TLV as defined in Section 4.3.
- o In addition, BGP-LS Link Attributes, as defined in [RFC7752], MAY be inserted in order to advertise the characteristics of the link.

5.2. Peer Adjacency Segment (Peer-Adj-SID)

The Peer Adjacency Segment, at the BGP node advertising it, has the following semantic:

- o SR header operation: NEXT (as defined in [I-D.ietf-spring-segment-routing]).
- o Next-Hop: the interface peer address.

The Peer Adjacency Segment is advertised with a Link NLRI, where:

- o Local Node Descriptors contains
 - Local BGP Router-ID of the BGP-EPE enabled egress PE.
 - Local ASN.
 - BGP-LS Identifier.
- o Remote Node Descriptors contains
 - Peer BGP Router-ID (i.e.: the peer BGP ID used in the BGP session).
 - Peer ASN.

- o Link Descriptors Sub-TLVs, as defined in [RFC7752], MUST contain the following TLVs:
 - * Link Local/Remote Identifiers (Sub-TLV 258) contains the 4-octet Link Local Identifier followed by the 4-octet value 0 indicating the Link Remote Identifier in unknown [RFC5307].
- o In addition, Link Descriptors Sub-TLVs, as defined in [RFC7752], MAY contain the following TLVs:
 - * IPv4 Interface Address (Sub-TLV 259) contains the address of the local interface through which the BGP session is established.
 - * IPv6 Interface Address (Sub-TLV 261) contains the address of the local interface through which the BGP session is established.
 - * IPv4 Neighbor Address (Sub-TLV 260) contains the IPv4 address of the peer interface used by the BGP session.
 - * IPv6 Neighbor Address (Sub-TLV 262) contains the IPv6 address of the peer interface used by the BGP session.
- o Link attribute used with the Peer-Adj-SID contains the TLV as defined in Section 4.3.

In addition, BGP-LS Link Attributes, as defined in [RFC7752], MAY be inserted in order to advertise the characteristics of the link.

5.3. Peer Set Segment

The Peer Adjacency Segment, at the BGP node advertising it, has the following semantic:

- o SR header operation: NEXT (as defined in [I-D.ietf-spring-segment-routing]).
- o Next-Hop: load balance across any connected interface to any peer in the related set.

The Peer Set Segment is advertised within a Link NLRI (describing a Peer Node Segment or a Peer Adjacency segment) as a BGP-LS attribute.

The Peer Set Attribute contains the Peer-Set-SID TLV, defined in Section 4.3 identifying the set of which the Peer Node Segment or Peer Adjacency Segment is a member.

6. Illustration

6.1. Reference Diagram

The following reference diagram is used throughout this document. The solution is illustrated for IPv6 with MPLS-based segments and the BGP-EPE topology is based on eBGP sessions between external peers.

As stated in Section 3, the solution illustrated hereafter is equally applicable to an iBGP session topology. In other words, the solution also applies to the case where C, D, F, and E are in the same AS and run iBGP sessions between each other.

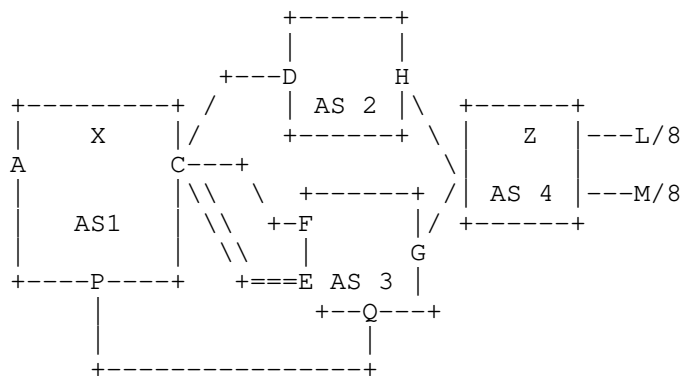


Figure 3: Reference Diagram

IP addressing:

- o C's IP address of interface to D: 2001:db8:cd::c/64, D's interface: 2001:db8:cd::d/64
- o C's IP address of interface to F: 2001:db8:cf::c/64, F's interface: 2001:db8:cf::f/64
- o C's IP address of upper interface to E: 2001:db8:ce1::c/64, E's interface: 2001:db8:ce1::e
- o C's local identifier of upper interface to E: 0.0.0.1.0.0.0.0
- o C's IP address of lower interface to E: 2001:db8:ce2::c, E's interface: 2001:db8:ce2::e
- o C's local identifier of lower interface to E: 0.0.0.2.0.0.0.0

- o Loopback of E used for eBGP multi-hop peering to C:
2001:db8:e::e/128

- o C's loopback is 2001:db8:c::c/128 with SID 64

BGP Router-IDs are C, D, F and E.

- o C's BGP Router-ID: 192.0.2.3
- o D's BGP Router-ID: 192.0.2.4
- o E's BGP Router-ID: 192.0.2.5
- o F's BGP Router-ID: 192.0.2.6

C's BGP peering:

- o Single-hop eBGP peering with neighbor 2001:db8:cd::d (D)
- o Single-hop eBGP peering with neighbor 2001:db8:cf::f (F)
- o Multi-hop eBGP peering with E on ip address 2001:db8:e::e (E)

C's resolution of the multi-hop eBGP session to E:

- o Static route 2001:db8:e::e/128 via 2001:db8:ce1::e
- o Static route 2001:db8:e::e/128 via 2001:db8:ce2::e

Node C configuration is such that:

- o A Peer Node segment (Peer-Node-SID) is allocated to each peer (D, F and E).
- o An Peer Adjacency segment (Peer-Adj-SID) is defined for each recursing interface to a multi-hop peer (CE upper and lower interfaces).
- o A Peer Set segment (Peer-Set-SID) is defined to include all peers in AS3 (peers F and E).

Local BGP-LS Identifier in router C is set to 10000.

The Link NLRI Type is used in order to encode C's connectivity. The Link NLRI uses the Protocol-ID value (to be assigned by IANA)

Once the BGP-LS update is originated by C, it may be advertised to internal (iBGP) as well as external (eBGP) neighbors supporting the BGP-LS EPE extensions defined in this document.

6.2. Peer Node Segment for Node D

Descriptors:

- o Local Node Descriptors (BGP Router-ID, local ASN, BGP-LS Identifier): 192.0.2.3, AS1, 10000
- o Remote Node Descriptors (BGP Router-ID, peer ASN): 192.0.2.4, AS2
- o Link Descriptors (BGP session IPv6 local address, BGP session IPv6 neighbor address): 2001:db8:cd::c, 2001:db8:cd::d

Attributes:

- o Peer-Node-SID: 1012
- o Link Attributes: see section 3.3.2 of [RFC7752]

6.3. Peer Node Segment for Node F

Descriptors:

- o Local Node Descriptors (BGP Router-ID, ASN, BGPLS Identifier): 192.0.2.3, AS1, 10000
- o Remote Node Descriptors (BGP Router-ID ASN): 192.0.2.6, AS3
- o Link Descriptors (BGP session IPv6 local address, BGP session IPv6 peer address): 2001:db8:cf::c, 2001:db8:cf::f

Attributes:

- o Peer-Node-SID: 1022
- o Peer-Set-SID: 1060
- o Link Attributes: see section 3.3.2 of [RFC7752]

6.4. Peer Node Segment for Node E

Descriptors:

- o Local Node Descriptors (BGP Router-ID, ASN, BGP-LS Identifier): 192.0.2.3, AS1, 10000

- o Remote Node Descriptors (BGP Router-ID, ASN): 192.0.2.5, AS3
- o Link Descriptors (BGP session IPv6 local address, BGP session IPv6 peer address): 2001:db8:c::c, 2001:db8:e::e

Attributes:

- o Peer-Node-SID: 1052
- o Peer-Set-SID: 1060

6.5. Peer Adjacency Segment for Node E, Link 1

Descriptors:

- o Local Node Descriptors (BGP Router-ID, ASN, BGP-LS Identifier): 192.0.2.3, AS1, 10000
- o Remote Node Descriptors (BGP Router-ID, ASN): 192.0.2.5, AS3
- o Link Descriptors (local interface identifier, IPv6 peer interface address): 0.0.0.1.0.0.0.0 , 2001:db8:ce1::e

Attributes:

- o Peer-Adj-SID: 1032
- o LinkAttributes: see section 3.3.2 of [RFC7752]

6.6. Peer Adjacency Segment for Node E, Link 2

Descriptors:

- o Local Node Descriptors (BGP Router-ID, ASN, BGP-LS Identifier): 192.0.2.3, AS1, 10000
- o Remote Node Descriptors (BGP Router-ID, ASN): 192.0.2.5, AS3
- o Link Descriptors (local interface identifier, IPv6 peer interface address): 0.0.0.2.0.0.0.0 , 2001:db8:ce2::e

Attributes:

- o Peer-Adj-SID: 1042
- o LinkAttributes: see section 3.3.2 of [RFC7752]

7. Implementation Status

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

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

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

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

Codepoint	Description
7	Protocol-ID BGP
516	BGP Router-ID
517	BGP Confederation Member
1101	Peer-Node-SID
1102	Peer-Adj-SID
1103	Peer-Set-SID

IANA has now confirmed the assignment of the above coidepoints. SeeSection 8.

8. IANA Considerations

This document defines:

A new Protocol-ID: BGP. The codepoint is from the "BGP-LS Protocol-IDs" registry.

Two new TLVs: BGP-Router-ID and BGP Confederation Member. The codepoints are in the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry.

Three new BGP-LS Attribute TLVs: Peer-Node-SID, Peer-Adj-SID and Peer-Set-SID. The codepoints are in the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry.

8.1. New BGP-LS Protocol-ID

This document defines a new value in the registry "BGP-LS Protocol-IDs":

Codepoint	Description	Status
7	BGP	Assigned by IANA

8.2. Node Descriptors and Link Attribute TLVs

This document defines 5 new TLVs in the registry "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs":

- o Two new node descriptor TLVs
- o Three new link attribute TLVs

All the new 5 codepoints are in the same registry: "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs". However, the registry is organized in ranges (node descriptors, link descriptors, node attributes, link attributes).

The following new Node Descriptors TLVs are defined:

Codepoint	Description	Status
516	BGP Router-ID	Assigned by IANA
517	BGP Confederation Member	Assigned by IANA

The following new Link Attribute TLVs are defined:

Codepoint	Description	Status
1101	Peer-Node-SID	Assigned by IANA
1102	Peer-Adj-SID	Assigned by IANA
1103	Peer-Set-SID	Assigned by IANA

9. Manageability Considerations

This BGP-LS ([RFC7752]) extensions that are described in this document consists of additional BGP-LS descriptors and TLVs that will follow the same manageability functions of BGP-LS, described in [RFC7752].

The operator MUST be capable of configuring, enabling, disabling the advertisement of each of the Peer-Node-SID, Peer-Adj-SID and Peer-Set-SID as well as to control which information is advertised to which internal or external peer. This is not different from what is required by a BGP speaker in terms of information origination and advertisement. In addition, the advertisement of EPE information MUST conform to standard BGP advertisement and propagation rules (iBGP, eBGP, Route-Reflectors, Confederations).

10. Security Considerations

[RFC7752] defines BGP-LS NLRIs to which the extensions defined in this document apply.

The Security Section of [RFC7752] also applies to:

- o New Node Descriptors Sub-TLVs: BGP-Router-ID and BGP-Confederation-Member;
- o New BGP-LS Attributes TLVs: Peer-Node-SID, Peer-Adj-SID and Peer-Set-SID.

11. Contributors

Acee Lindem gave a substantial contribution to this document.

12. Acknowledgements

The authors would like to thank Jakob Heitz, Howard Yang, Hannes Gredler, Peter Psenak, Ketan Jivan Talaulikar, Arjun Sreekantiah and Bruno Decraene for their feedback and comments.

13. References

13.1. Normative References

- [I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S.,
and R. Shakir, "Segment Routing Architecture", draft-ietf-
spring-segment-routing-11 (work in progress), February
2017.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119,
DOI 10.17487/RFC2119, March 1997,
<<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A
Border Gateway Protocol 4 (BGP-4)", RFC 4271,
DOI 10.17487/RFC4271, January 2006,
<<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous
System Confederations for BGP", RFC 5065,
DOI 10.17487/RFC5065, August 2007,
<<http://www.rfc-editor.org/info/rfc5065>>.
- [RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions
in Support of Generalized Multi-Protocol Label Switching
(GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008,
<<http://www.rfc-editor.org/info/rfc5307>>.
- [RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP
Identifier for BGP-4", RFC 6286, DOI 10.17487/RFC6286,
June 2011, <<http://www.rfc-editor.org/info/rfc6286>>.

13.2. Informative References

- [I-D.ietf-idr-bgp-ls-segment-routing-ext]
Previdi, S., Psenak, P., Filsfils, C., Gredler, H., Chen,
M., and j. jefftant@gmail.com, "BGP Link-State extensions
for Segment Routing", draft-ietf-idr-bgp-ls-segment-
routing-ext-01 (work in progress), February 2017.
- [I-D.ietf-spring-segment-routing-central-epe]
Filsfils, C., Previdi, S., Aries, E., and D. Afanasiev,
"Segment Routing Centralized BGP Egress Peer Engineering",
draft-ietf-spring-segment-routing-central-epe-05 (work in
progress), March 2017.

- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<http://www.rfc-editor.org/info/rfc7752>>.
- [RFC7942] Sheffer, Y. and A. Farrel, "Improving Awareness of Running Code: The Implementation Status Section", BCP 205, RFC 7942, DOI 10.17487/RFC7942, July 2016, <<http://www.rfc-editor.org/info/rfc7942>>.

Authors' Addresses

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

Email: sprevidi@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Brussels
BE

Email: cfilsfil@cisco.com

Keyur Patel
Arrcus, Inc.

Email: Keyur@arrcus.com

Saikat Ray
Individual Contributor

Email: raysaikat@gmail.com

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

Email: jie.dong@huawei.com

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

Email: mach.chen@huawei.com

IDR and SIDR
Internet-Draft
Intended status: Standards Track
Expires: September 7, 2017

K. Sriram
D. Montgomery
US NIST
B. Dickson

K. Patel
Arrcus
A. Robachevsky
Internet Society
March 6, 2017

Methods for Detection and Mitigation of BGP Route Leaks
draft-ietf-idr-route-leak-detection-mitigation-06

Abstract

RFC 7908 provides a definition of the route leak problem, and also enumerates several types of route leaks. This document first examines which of those route-leak types are detected and mitigated by the existing origin validation (OV) [RFC 6811]. It is recognized that OV offers a limited detection and mitigation capability against route leaks. This document specifies enhancements that significantly extend the route-leak prevention, detection, and mitigation capabilities of BGP. One solution component involves intra-AS messaging from ingress router to egress router using a BGP Community or Attribute. This intra-AS messaging prevents the AS from causing route leaks. Another solution component involves carrying a per-hop route-leak protection (RLP) field in BGP updates. The RLP fields are proposed to be carried in a new optional transitive attribute, called BGP RLP attribute. The RLP attribute helps with detection and mitigation of route leaks at ASes downstream from the leaking AS.

Status of This Memo

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

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

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

This Internet-Draft will expire on September 7, 2017.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
2. Related Prior Work	4
3. Do Origin Validation and BGPsec Assist in Route-Leak Detection?	4
4. Mechanisms for Prevention, Detection and Mitigation of Route Leaks	6
4.1. Ascertaining Peering Relationship	6
4.2. Prevention of Route Leaks at Local AS: Intra-AS Messaging	7
4.2.1. Non-Transitive BGP Community for Intra-AS Messaging	7
4.2.2. Non-Transitive BGP pRLP Attribute for Intra-AS Messaging	8
4.3. Route-Leak Protection (RLP) Field Encoding by Sending Router	8
4.3.1. BGP RLP Attribute	10
4.3.2. Carrying RLP Field Values in the BGPsec Flags	11
4.4. Recommended Actions at a Receiving Router for Detection of Route Leaks	11
4.5. Possible Actions at a Receiving Router for Mitigation	12
5. Stopgap Solution when Only Origin Validation is Deployed	12
6. Design Rationale and Discussion	13
6.1. Is route-leak solution without cryptographic protection a serious attack vector?	13
6.2. Combining results of route-leak detection, OV and BGPsec validation for path selection decision	15
6.3. Are there cases when valley-free violations can be considered legitimate?	15
6.4. Comparison with other methods (routing security BCPs)	16
6.5. Per-Hop RLP Field or Single RLP Flag per Update?	16
7. Security Considerations	18

8. IANA Considerations	19
9. Acknowledgements	19
10. References	19
10.1. Normative References	19
10.2. Informative References	19
Authors' Addresses	24

1. Introduction

[RFC7908] provides a definition of the route leak problem, and also enumerates several types of route leaks. This document first examines which of those route-leak types are detected and mitigated by the existing Origin Validation (OV) [RFC6811] method. OV and BGPsec path validation [I-D.ietf-sidr-bgpsec-protocol] together offer mechanisms to protect against re-originations and hijacks of IP prefixes as well as man-in-the-middle (MITM) AS path modifications. Route leaks (see [RFC7908] and references cited at the back) are another type of vulnerability in the global BGP routing system against which OV offers only partial protection. BGPsec (i.e. path validation) provides cryptographic protection for some aspects of BGP update messages, but in its current form BGPsec doesn't offer any protection against route leaks.

For the types of route leaks enumerated in [RFC7908], where the OV method does not offer a solution, this document specifies enhancements that significantly extend the route-leak prevention, detection, and mitigation capabilities of BGP. One solution component involves intra-AS messaging from ingress router to egress router using a BGP Community or Attribute. This intra-AS messaging prevents the AS from causing route leaks. Another solution component involves carrying a per-hop route-leak protection (RLP) field in BGP updates. The RLP fields are proposed to be carried in a new optional transitive attribute, called BGP RLP attribute. The RLP attribute helps with detection and mitigation of route leaks at ASes downstream from the leaking AS.

The solution is meant to be initially implemented as an enhancement of BGP without requiring BGPsec. However, when BGPsec is deployed in the future, the solution can be incorporated in BGPsec, enabling cryptographic protection for the RLP field. That would be one way of implementing the proposed solution in a secure way. It is not claimed that the solution detects all possible types of route leaks but it detects several types, especially considering some significant route-leak occurrences that have been observed in recent years. The document also includes a stopgap method (in Section 5) for detection and mitigation of route leaks for an intermediate phase when OV is deployed but BGP protocol on the wire is unchanged.

2. Related Prior Work

A mechanism embodied in the proposed solution is based on setting an attribute in BGP route announcement to manage the transmission/receipt of the announcement based on the type of neighbor (e.g. customer, transit provider, etc.). Documented prior work related to said basic idea and mechanism dates back to at least the 1980's. Some examples of prior work are: (1) Information flow rules described in [proceedings-sixth-ietf] (see pp. 195-196); (2) Link Type described in [RFC1105-obsolete] (see pp. 4-5); (3) Hierarchical Recording described in [draft-kunzinger-idrp-ISO10747-01] (see Section 6.3.1.12). The problem of route leaks and possible solution mechanisms based on encoding peering-link type information, e.g. P2C (i.e. Transit-Provider to Customer), C2P (i.e. Customer to Transit-Provider), p2p (i.e. peer to peer) etc., in BGPsec updates and protecting the same under BGPsec path signatures have been discussed in IETF SIDR WG at least since 2011.

[draft-dickson-sidr-route-leak-solns] attempted to describe these mechanisms in a BGPsec context. The draft expired in 2012.

[draft-dickson-sidr-route-leak-solns] defined neighbor relationships on a per link basis, but in the current document the relationship is encoded per prefix, as routes for prefixes with different peering relationships may be sent over the same link. Also

[draft-dickson-sidr-route-leak-solns] proposed a second signature block for the link type encoding, separate from the path signature block in BGPsec. By contrast, in the current document when BGPsec-based solution is considered, cryptographic protection is provided for Route-Leak Protection (RLP) encoding using the same signature block as that for path signatures (see Section 4.3.2).

3. Do Origin Validation and BGPsec Assist in Route-Leak Detection?

Referring to the enumeration of route leaks discussed in [RFC7908], Table 1 summarizes the route-leak detection capability offered by OV and BGPsec for different types of route leaks. (Note: Prefix filtering is not considered here in this table. Please see Section 5.)

A detailed explanation of the contents of Table 1 is as follows. It is readily observed that route leaks of Types 1, 2, 3, and 4 are not detected by OV or BGPsec in its current form. Clearly, Type 5 route leak involves re-origination or hijacking, and hence can be detected by OV. In the case of Type 5 route leak, there would be no existing ROAs to validate a re-originated prefix or more specific, but instead a covering ROA would normally exist with the legitimate AS, and hence the update will be considered Invalid by OV.

Type of Route Leak	Current State of Detection Coverage
Type 1: Hairpin Turn with Full Prefix	Neither OV nor BGPsec (in its current form) detects Type 1.
Type 2: Lateral ISP-ISP-ISP Leak	Neither OV nor BGPsec (in its current form) detects Type 2.
Type 3: Leak of Transit-Provider Prefixes to Peer	Neither OV nor BGPsec (in its current form) detects Type 3.
Type 4: Leak of Peer Prefixes to Transit Provider	Neither OV nor BGPsec (in its current form) detects Type 4.
Type 5: Prefix Re-Origination with Data Path to Legitimate Origin	OV detects Type 5.
Type 6: Accidental Leak of Internal Prefixes and More Specifics	For internal prefixes never meant to be routed on the Internet, OV helps detect their leak; they might either have no covering ROA or have an AS0-ROA to always filter them. In the case of accidental leak of more specifics, OV may offer some detection due to ROA maxLength.

Table 1: Examination of Route-Leak Detection Capability of Origin Validation and Current BGPsec Path Validation

In the case of Type 6 leaks involving internal prefixes that are not meant to be routed in the Internet, they are likely to be detected by OV. That is because such prefixes might either have no covering ROA or have an AS0-ROA to always filter them. In the case of Type 6 leaks that are due to accidental leak of more specifics, they may be detected due to violation of ROA maxLength. BGPsec (i.e. path validation) in its current form does not detect Type 6. However, route leaks of Type 6 are least problematic due to the following reasons. In the case of leak of more specifics, the offending AS is itself the legitimate destination of the leaked more-specific prefixes. Hence, in most cases of this type, the data traffic is neither misrouted nor denied service. Also, leaked announcements of Type 6 are short-lived and typically withdrawn quickly following the announcements. Further, the MaxPrefix limit may kick-in in some

receiving routers and that helps limit the propagation of sometimes large number of leaked routes of Type 6.

Realistically, BGPsec may take a much longer time being deployed than OV. Hence solution proposals for route leaks should consider both scenarios: (A) OV only (without BGPsec) and (B) OV plus BGPsec. Assuming an initial scenario A, and based on the above discussion and Table 1, it is evident that the solution method should focus primarily on route leaks of Types 1, 2, 3, and 4.

4. Mechanisms for Prevention, Detection and Mitigation of Route Leaks

There are two considerations for route leaks: (1) Prevention of route leaks from a local AS, and (2) Detection and mitigation of route leaks in ASes that are downstream from the leaking AS.

In Section 4.1, the method of ascertaining peering relationship per prefix is described. Section 4.2 describes intra-AS messaging methods for prevention of route leaks from local AS. Section 4.3 and Section 4.4 describe a simple addition to BGP that facilitates detection and mitigation of route leaks of Types 1, 2, 3, and 4 (see Section 3) at a downstream AS from the leaking AS.

4.1. 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 [I-D.ymbk-idr-bgp-open-policy] [Luckie]. On a per-prefix basis, the peering role types simplify to provider, customer, or lateral peer.

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.ymbk-idr-bgp-open-policy] introduces a method of confirming the BGP Role during BGP OPEN messaging. It defines a new BGP Role capability, which helps in re-confirming the relationship. 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.ymbk-idr-bgp-open-policy].

When the BGP relationship information has been correctly exchanged (i.e. free of contradictions) including the sets of prefixes with different roles (if complex), then this information SHOULD be used to 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.

4.2. Prevention of Route Leaks at Local AS: Intra-AS Messaging

Note: The intra-AS messaging for route leak prevention can be done using non-transitive BGP Community or Attribute. Both options are described below; one of them will be chosen after IDR working group consensus is established.

4.2.1. Non-Transitive BGP Community for Intra-AS Messaging

The following procedure (or similar) for intra-AS messaging (i.e. between ingress and egress routers) for prevention of route leaks is a fairly common practice used by large ISPs. (Note: This information was gathered from discussions on the NANOG mailing list [Nanog-thread-June2016] as well as through private discussions with operators of large ISP networks.)

Routes are tagged on ingress to an AS with communities for origin, including the type of eBGP peer it was learned from (customer, provider or lateral peer), geographic location, etc. The community attributes are carried across the AS with the routes. Routes that the AS originates directly are tagged with similar origin communities when they are redistributed into BGP from static, IGP, etc. These communities are used along with additional logic in route policies to determine which routes are to be announced to which eBGP peers and which are to be dropped. Route policy is applied to eBGP sessions based on what set of routes they should receive (transit, full routes, internal-only, default-only, etc.). In this process, the ISP's AS also ensures that routes learned from a transit-provider or a lateral peer (i.e. non-transit) at an ingress router are not leaked at an egress router to another transit-provider or lateral peer.

Additionally, in many cases, ISP network operators' outbound policies require explicit matches for expected communities before passing routes. This helps ensure that that if an update has made it into the routing table (i.e. RIB) but has missed its ingress community tagging (due to a missing/misapplied ingress policy), it will not be inadvertently leaked.

The above procedure (or a simplified version of it) is also applicable when an AS consists of a single eBGP router. It is

recommended that all AS operators SHOULD implement the procedure described above (or similar that is appropriate for their network) to prevent route leaks that they have direct control over.

4.2.2. Non-Transitive BGP pRLP Attribute for Intra-AS Messaging

It is possible to use an optional non-transitive BGP Attribute instead of the Community described above for intra-AS messaging for route leak prevention. The following description would be used in case the IDR working group decides on using a BGP Attribute.

A new optional non-transitive BGP Attribute called Preventive Route Leak Protection (pRLP) is used. The attribute type code for the pRLP attribute is to be assigned by IANA. The length of this attribute is 0 as it is used only as a flag.

Ingress (receiving) router action: The decision to set or not set the pRLP flag is made by a receiving router upon a route ingress. The flag is set when the route is received from a provider or a lateral peer. The flag is not set when the route is received from a customer. When the relationship is complex, the flag is set based on the per-prefix peering role information discussed in Section 4.1.

Egress (sending) router action: A sending router is allowed to send a route without the pRLP flag to any neighbor (transit-provider, customer, lateral peer). However, if the pRLP flag is present, then the route MUST NOT be sent to a transit-provider or a lateral peer.

An AS that follows the above set of receiver (ingress) and sender (egress) actions, prevents itself from causing route leaks.

4.3. Route-Leak Protection (RLP) Field Encoding by Sending Router

This section, Section 4.4 and Section 4.5 describe methods of detection and mitigation of route leaks in an AS downstream from the leaking AS.

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 update message that its customer AS (e.g. AS3 in Figure 1) SHOULD NOT have forwarded the update (towards the transit-provider AS). This means that at least one of the ASes in the AS path of the update has indicated that it sent the update to its customer or lateral (i.e. non-transit) peer, but forbade any subsequent 'Up' forwarding (i.e. from a customer AS to its transit-provider AS). For this purpose, a Route-Leak Protection (RLP) field to be set by a sending router is proposed to be used for each AS hop.

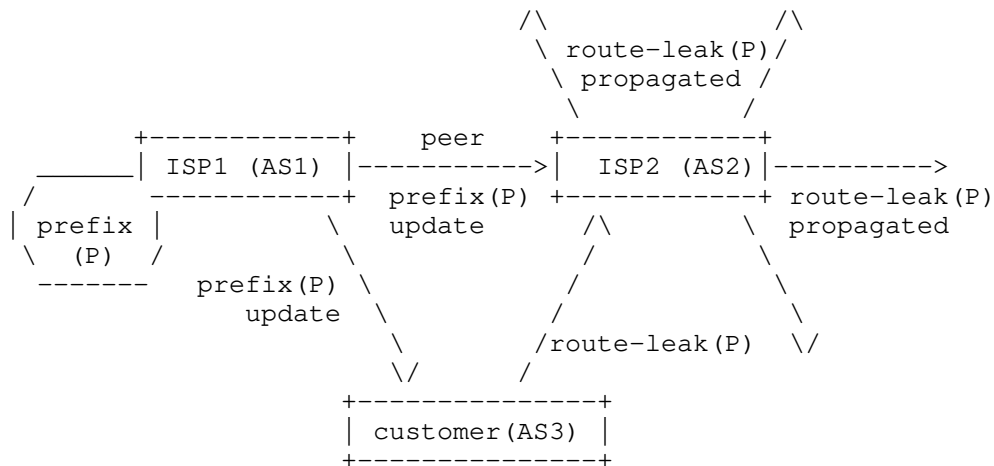


Figure 1: Illustration of the basic notion of a route leak.

For the purpose of route-leak detection and mitigation proposed in this document, the RLP field value SHOULD be set to one of two values as follows:

- o 0: This is the default value (i.e. "nothing specified"),
- o 1: This is the 'Do not Propagate Up or Lateral' indication; sender indicating that the route SHOULD NOT be forwarded 'Up' towards a transit-provider AS or to a lateral (i.e. non-transit) peer AS.

The RLP indications SHOULD be set on a per prefix basis. This is because some peering relations between neighbors can be complex (see Section 4.1). Further, the RLP indications are set on a per-hop (i.e. per AS) basis.

There are two different scenarios when a sending AS SHOULD set value 1 in the RLP field: (a) when sending the update to a customer AS, and (b) when sending the update to a lateral peer (i.e. non-transit) AS. In essence, in both scenarios, the intent of RLP = 1 is that the neighbor AS and any receiving AS along the subsequent AS path SHOULD NOT forward the update 'Up' towards its (receiving AS's) transit-provider AS or laterally towards its peer (i.e. non-transit) AS.

When sending an update 'Up' to a transit-provider AS, the RLP encoding SHOULD be set to the default value of 0. When a sending AS sets the RLP encoding to 0, it is indicating to the receiving AS that the update can be propagated in any direction (i.e. towards transit-provider, customer, or lateral peer).

The two-state specification in the RLP field (as described above) works for detection and mitigation of route leaks of Types 1, 2, 3, and 4 which are the focus here (see Section 4.4 and Section 4.5).

An AS MUST NOT rewrite/reset the values set by any preceding ASes in their respective RLP fields.

The proposed RLP encoding SHOULD be carried in BGP-4 [RFC4271] updates in a new BGP optional transitive attribute (see Section 4.3.1). In BGPsec, it SHOULD be carried in the Flags field (see Section 4.3.2).

4.3.1. BGP RLP Attribute

The BGP RLP attribute is a new BGP optional transitive attribute. The attribute type code for the RLP attribute is to be assigned by IANA. The length field of this attribute is 2 octets. The value field of the RLP attribute is defined as a set of one or more pairs of ASN (4 octets) and RLP (one octet) fields as described below (Figure 2).

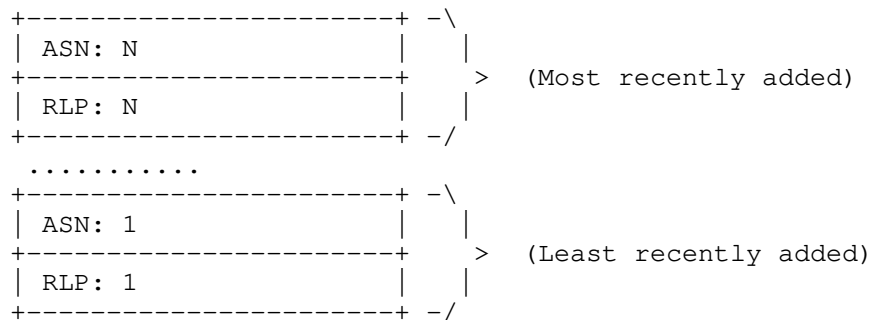


Figure 2: BGP RLP Attribute format.

The RLP Attribute value is a sequence of these two components (see Figure 2):

ASN: Four octets encoding the public registered AS number of a BGP speaker.

RLP Field: One octet encoding the RLP Field bits. The value of the RLP Field octet can be 0 (decimal) or 1 (decimal) as described above in Section 4.3.1. Its usage will be further discussed in subsequent sections.

If all ASes in the AS_PATH of a route are upgraded to participate in RLP, then the ASNs in the RLP TLV in Figure 2 will correspond one-to-one with sequence of ASes in the AS_PATH (excluding prepends). If some ASes do not participate, then one or more {ASN, RLP} tuples may be missing in the RLP attribute relative to the AS_PATH.

4.3.2. Carrying RLP Field Values in the BGPsec Flags

In BGPsec enabled routers, the RLP encoding SHOULD be accommodated in the existing Flags field in BGPsec updates. The Flags field is part of the Secure_Path Segment in BGPsec updates [I-D.ietf-sidr-bgpsec-protocol]. It is one octet long, and one Flags field is available for each AS hop, and currently only the first bit is used in BGPsec. So there are 7 bits that are currently unused in the Flags field. One of these bits can be designated for the RLP field value (see Section 4.3.1). This bit can be set to 0 when the RLP Field value is 0 and set to 1 when the RLP Field value is 1. Since the BGPsec protocol specification requires a sending AS to include the Flags field in the data that are signed over, the RLP field for each hop (assuming it would be part of the Flags field as described) will be protected under the sending AS's signature.

4.4. Recommended Actions at a Receiving Router for Detection of Route Leaks

The following receiver algorithm is RECOMMENDED for detecting route leaks:

A receiving router MUST mark an update as a 'Route Leak' if ALL of the following conditions hold true:

1. The update is received from a customer or lateral peer AS.
2. The update has the RLP Field set to 1 (i.e. 'Do not Propagate Up or Lateral') indication for one or more hops (excluding the most recent) in the AS path.

The reason for stating "excluding the most recent" in the above algorithm is as follows. An ISP should look at RLP values set by ASes preceding the immediate sending AS in order to ascertain a leak. The receiving router already knows that the most recent hop in the update is from its customer or lateral-peer AS to itself, and it does not need to rely on the RLP field value set by that AS (i.e the immediate neighbor AS in the AS path) for detection of route leaks.

If the RLP encoding is secured by BGPsec (see Section 4.3) and hence protected against tampering by intermediate ASes, then there would be

added certainty in the route-leak detection algorithm described above (see discussions in Section 6.1 and Section 6.2).

4.5. Possible Actions at a Receiving Router for Mitigation

After applying the above detection algorithm, a receiving router may use any policy-based algorithm of its own choosing to mitigate any detected route leaks. An example receiver algorithm for mitigating a route leak is as follows:

- o If an update from a customer or lateral peer AS is marked as a 'Route Leak' (see Section 4.4), then the receiving router SHOULD prefer an alternate unmarked route.
- o If no alternate unmarked route is available, then a route marked as a 'Route Leak' MAY be accepted.

A basic principle here is that if an AS receives and marks a customer route as 'Route Leak', then the AS should override the "prefer customer route" policy, and instead prefer an alternate 'clean' route learned from another customer, a lateral peer, or a transit provider. This can be implemented by adjusting the local preference for the routes in consideration.

5. Stopgap Solution when Only Origin Validation is Deployed

A stopgap method is described here for detection and mitigation of route leaks for the intermediate phase when OV is deployed but BGP protocol on the wire is unchanged. The stopgap solution can be in the form of construction of a prefix filter list from ROAs. A suggested procedure for constructing such a list comprises of the following steps:

- o ISP makes a list of all the ASes (Cust_AS_List) that are in its customer cone (ISP's own AS is also included in the list). (Some of the ASes in Cust_AS_List may be multi-homed to another ISP and that is OK.)
- o ISP downloads from the RPKI repositories a complete list (Cust_ROA_List) of valid ROAs that contain any of the ASes in Cust_AS_List.
- o ISP creates a list of all the prefixes (Cust_Prfx_List) that are contained in any of the ROAs in Cust_ROA_List.
- o Cust_Prfx_List is the allowed list of prefixes that is permitted by the ISP's AS, and will be forwarded by the ISP to upstream ISPs, customers, and peers.

- o A route for a prefix that is not in `Cust_Prfx_List` but announced by one of ISP's customers is 'marked' as a potential route leak. Further, the ISP's router SHOULD prefer an alternate route that is Valid (i.e. valid according to origin validation) and 'clean' (i.e. not marked) over the 'marked' route. The alternate route may be from a peer, transit provider, or different customer.

Special considerations with regard to the above procedure may be needed for DDoS mitigation service providers. They typically originate or announce a DDoS victim's prefix to their own ISP on a short notice during a DDoS emergency. Some provisions would need to be made for such cases, and they can be determined with the help of inputs from DDoS mitigation service providers.

For developing a list of all the ASes (`Cust_AS_List`) that are in the customer cone of an ISP, the AS path based Outbound Route Filter (ORF) technique [draft-ietf-idr-aspath-orf] can be helpful (see discussion in Section 6.4).

Another technique based on AS_PATH filters is described in [Snijders]. This method is applicable to very large ISPs (i.e. big networks) that have lateral peering. For a pair of such very large ISPs, say A and B, the method depends on ISP A communicating out-of-band (e.g. by email) with ISP B about whether or not it (ISP A) has any transit providers. This out-of-band knowledge enables ISP B to apply suitable AS_PATH filtering criteria for routes involving the presence of ISP A in the path and prevent certain kinds of route leaks (see [Snijders] for details).

6. Design Rationale and Discussion

This section provides design justifications for the methodology specified in Section 4, and also answers some questions that are anticipated or have been raised in the IETF IDR and SIDR working group meetings.

6.1. Is route-leak solution without cryptographic protection a serious attack vector?

It has been asked if a route-leak solution without BGPsec, i.e. when RLP Fields are not protected, can turn into a serious new attack vector. The answer seems to be: not really! Even the NLRI and AS_PATH in BGP updates are attack vectors, and RPKI/OV/BGPsec seek to fix that. Consider the following. Say, if 99% of route leaks are accidental and 1% are malicious, and if route-leak solution without BGPsec eliminates the 99%, then perhaps it is worth it (step in the right direction). When BGPsec comes into deployment, the route-leak protection (RLP) bits can be mapped into BGPsec (using the Flags

field) and then necessary security will be in place as well (within each BGPsec island as and when they emerge).

Further, let us consider the worst-case damage that can be caused by maliciously manipulating the RLP Field values in an implementation without cryptographic protection (i.e. sans BGPsec). Manipulation of the RLP bits can result in one of two types of attacks: (a) Upgrade attack and (b) Downgrade attack. Descriptions and discussions about these attacks follow. In what follows, P2C stands for transit provider to customer (Down); C2P stands for customer to transit provider (Up), and p2p stands for peer to peer (lateral or non-transit relationship).

(a) Upgrade attack: An AS that wants to intentionally leak a route would alter the RLP encodings for the preceding hops from 1 (i.e. 'Do not Propagate Up or Lateral') to 0 (default) wherever applicable. This poses no problem for a route that keeps propagating in the 'Down' (P2C) direction. However, for a route that propagates 'Up' (C2P) or 'Lateral' (p2p), the worst that can happen is that a route leak goes undetected. That is, a receiving router would not be able to detect the leak for the route in question by the RLP mechanism described here. However, the receiving router may still detect and mitigate it in some cases by applying other means such as prefix filters [RFC7454]. If some malicious leaks go undetected (when RLP is deployed without BGPsec) that is possibly a small price to pay for the ability to detect the bulk of route leaks that are accidental.

(b) Downgrade attack: RLP encoding is set to 1 (i.e. 'Do not Propagate Up or Lateral') when it should be set to 0 (default). This would result in a route being mis-detected and marked as a route leak. By default RLP encoding is set to 0, and that helps reduce errors of this kind (i.e. accidental downgrade incidents). Every AS or ISP wants reachability for prefixes it originates and for its customer prefixes. So an AS or ISP is not likely to change an RLP value 0 to 1 intentionally. If a route leak is detected (due to intentional or accidental downgrade) by a receiving router, it would prefer an alternate 'clean' route from a transit provider or peer over a 'marked' route from a customer. It may end up with a suboptimal path. In order to have reachability, the receiving router would accept a 'marked' route if there is no alternative that is 'clean'. So RLP downgrade attacks (intentional or accidental) would be quite rare, and the consequences do not appear to be grave.

6.2. Combining results of route-leak detection, OV and BGPsec validation for path selection decision

Combining the results of route-leak detection, OV, and BGPsec validation for path selection decision is up to local policy in a receiving router. As an example, a router may always give precedence to outcomes of OV and BGPsec validation over that of route-leak detection. That is, if an update fails OV or BGPsec validation, then the update is not considered a candidate for path selection. Instead, an alternate update is chosen that passed OV and BGPsec validation and additionally was not marked as route leak.

If only OV is deployed (and not BGPsec), then there are six possible combinations between OV and route-leak detection outcomes. Because there are three possible outcomes for OV (NotFound, Valid, and Invalid) and two possible outcomes for route-leak detection (marked as leak and not marked). If OV and BGPsec are both deployed, then there are twelve possible combinations between OV, BGPsec validation, and route-leak detection outcomes. As stated earlier, since BGPsec protects the RLP encoding, there would be added certainty in route-leak detection outcome if an update is BGPsec valid (see Section 6.1).

6.3. Are there cases when valley-free violations can be considered legitimate?

There are studies in the literature [Anwar] [Giotsas] [Wijchers] observing and analyzing the behavior of routes announced in BGP updates using data gathered from the Internet. In particular, the studies have focused on how often there appear to be valley-free (e.g. Gao-Rexford [Gao] model) violations, and if they can be explained [Anwar]. One important consideration for explanation of violations is per-prefix routing policies, i.e. routes for prefixes with different peering relationships may be sent over the same link. One encouraging result reported in [Anwar] is that when per-prefix routing policies are taken into consideration in the data analysis, more than 80% of the observed routing decisions fit the valley-free model (see Section 4.3 and SPA-1 data in Figure 2). [Anwar] also observes, "it is well known that this model [the basic Gao-Rexford model and some variations of it] fails to capture many aspects of the interdomain routing system. These aspects include AS relationships that vary based on the geographic region or destination prefix, and traffic engineering via hot-potato routing or load balancing." So there may be potential for explaining the remaining (20% or less) violations of valley-free as well.

One major design factor in the methodology described in this document is that the Route-Leak Protection (RLP) encoding is per prefix. So

the proposed solution is consistent with ISPs' per-prefix routing policies. Large global and other major ISPs will be the likely early adopters, and they are expected to have expertise in setting policies (including per prefix policies, if applicable), and make proper use of the RLP indications on a per prefix basis. When the large ISPs participate in this solution deployment, it is envisioned that they would form a ring of protection against route leaks, and co-operatively avoid many of the common types of route leaks that are observed. Route leaks may still happen occasionally within the customer cones (if some customer ASes are not participating or not diligently implementing RLP), but said leaks would be much less likely to propagate from one large participating ISP to another.

6.4. Comparison with other methods (routing security BCPs)

It is reasonable to ask if techniques considered in BCPs such as [RFC7454] (BGP Operations and Security) and [NIST-800-54] may be adequate to address route leaks. The prefix filtering recommendations in the BCPs may be complementary but not adequate. The difficulty is in ISPs' ability to construct prefix filters that represent their customer cones (CC) accurately, especially when there are many levels in the hierarchy within the CC. In the RLP-encoding based solution described here, AS operators signal for each route propagated, if it SHOULD NOT be subsequently propagated to a transit provider or peer.

AS path based Outbound Route Filter (ORF) described in [draft-ietf-idr-aspath-orf] is also an interesting complementary technique. It can be used as an automated collaborative messaging system (implemented in BGP) for ISPs to try to develop a complete view of the ASes and AS paths in their CCs. Once an ISP has that view, then AS path filters can be possibly used to detect route leaks. One limitation of this technique is that it cannot duly take into account the fact that routes for prefixes with different peering relationships may be sent over the same link between ASes. Also, the success of AS path based ORF depends on whether ASes at all levels of the hierarchy in a CC participate and provide accurate information (in the ORF messages) about the AS paths they expect to have in their BGP updates.

6.5. Per-Hop RLP Field or Single RLP Flag per Update?

The route-leak detection and mitigation mechanism described in this document is based on setting RLP Fields on a per-hop basis. There is another possible mechanism based on a single RLP flag per update.

Method A - Per-Hop RLP Field: The sender (eBGP router) on each hop in the AS path sets its RLP Field = 1 if sending the update to a

customer or lateral peer (see Section 4.3) and Section 4.3.1). No AS (if operating correctly) would rewrite the RLP Field set by any preceding AS.

Method B - Single RLP Flag per Update: As it propagates, the update would have at most one RLP flag. Once an eBGP router (in the update path) determines that it is sending an update towards a customer or lateral peer AS, it sets the RLP flag. The flag value equals the AS number of the eBGP router that is setting it. Once the flag is set, subsequent ASes in the path must propagate the flag as is.

To compare Methods A and B, consider the example illustrated in Figure 3. Consider a partial deployment scenario in which AS1, AS2, AS3 and AS5 participate in RLP, and AS4 does not. AS1 (2 levels deep in AS3's customer cone) has imperfect RLP operation. Each complying AS's route leak mitigation policy is to prefer an update not marked as route leak (see Section 4.5). If there is no alternative, then a transit-provider may propagate a marked update from a customer. In this example, multi-homed AS4 leaks a route received for prefix Q from transit-provider AS3 to transit-provider AS5.

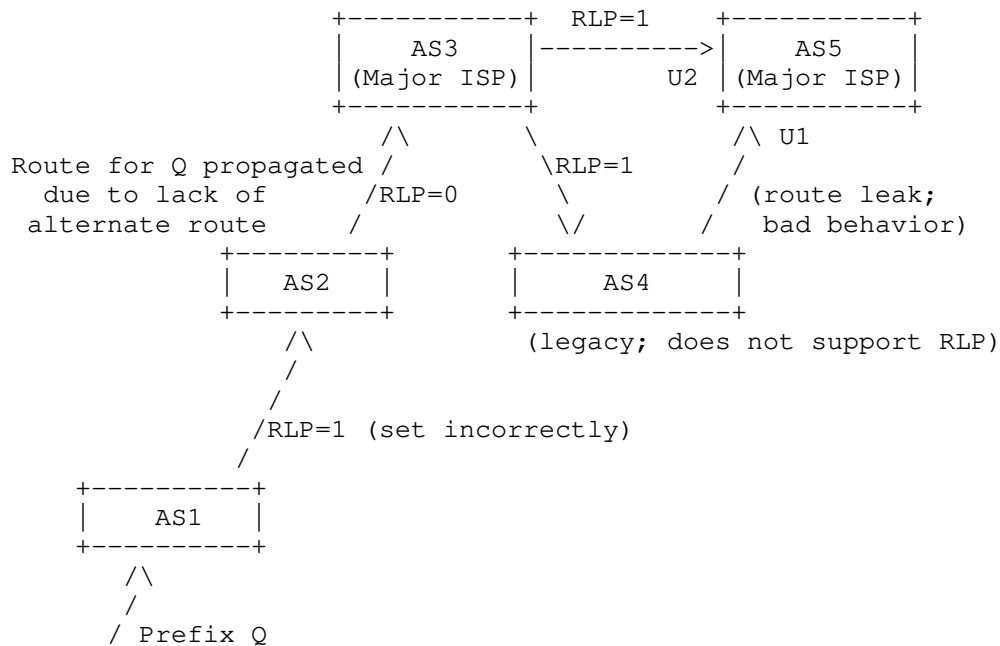


Figure 3: Example for comparison of Method A vs. Method B

If Method A is implemented in the network, the two BGP updates for prefix Q received at AS5 are (note that AS4 is not participating in RLP):

U1A: Q [AS4 AS3 AS2 AS1] {RLP3(AS3)=1, RLP2(AS2)=0, RLP1(AS1)=1}
..... from AS4

U2A: Q [AS3 AS2 AS1] {RLP3(AS3)=1, RLP2(AS2)=0, RLP1(AS1)=1}
from AS3

Alternatively, if Method B is implemented in the network, the two BGP updates for prefix Q received at AS5 are:

U1B: Q [AS4 AS3 AS2 AS1] {RLP(AS1)=1} from AS4

U2B: Q [AS3 AS2 AS1] {RLP(AS1)=1} from AS3

All received routes for prefix Q at AS5 are marked as route leak in either case (Method A or B). In the case of Method A, AS5 can use additional information gleaned from the RLP fields in the updates to possibly make a better best path selection. For example, AS5 can determine that U1A update received from its customer AS4 exhibits violation of two RLP fields (those set by AS1 and AS3) and one of them was set just two hops away. But U2A update exhibits that only one RLP field was violated and that was set three hops back. Based on this logic, AS5 may prefer U2A over U1A (even though U1A is a customer route). This would be a good decision. However, Method B does not facilitate this kind of more rational decision process. With Method B, both updates U1B and U2B exhibit that they violated only one RLP field (set by AS1 several hops away). AS5 may then prefer U1B over U2B since U1B is from a customer, and that would be bad decision. This illustrates that, due to more information in per-hop RLP Fields, Method A seems to be operationally more beneficial than Method B.

Further, for detection and notification of neighbor AS's non-compliance, Method A (per-hop RLP) is better than Method B (single RLP). With Method A, the bad behavior of AS4 would be explicitly evident to AS5 since it violated AS3's (only two hops away) RLP field as well. AS5 would alert AS4 and also AS2 would alert AS1 about lack of compliance (when Method A is used). With Method B, the alerting process may not be as expeditious.

7. Security Considerations

The proposed Route-Leak Protection (RLP) field requires cryptographic protection in order to prevent malicious route leaks. Since it is proposed that the RLP field be included in the Flags field in the

Secure_Path Segment in BGPsec updates, the cryptographic security mechanisms in BGPsec are expected to also apply to the RLP field. The reader is therefore directed to the security considerations provided in [I-D.ietf-sidr-bgpsec-protocol].

8. IANA Considerations

IANA is requested to register a new optional, non-transitive BGP Path Attribute, named "Preventive Route Leak Protection (pRLP)" in the BGP Path Attributes registry. The attribute type code is TBD. The reference for this new attribute is this document (i.e. the RFC that replaces this draft). The length of this new attribute is 0.

IANA is requested to register a new optional, transitive BGP Path Attribute, named "Route Leak Protection" in the BGP Path Attributes registry. The attribute type code is TBD. The reference for this new attribute is this document (i.e. the RFC that replaces this draft). The length field of this attribute is 2 octets, and the length of the value field of this attribute is variable (see Figure 2) in Section 4.3.1 of this document).

9. Acknowledgements

The authors wish to thank Jared Mauch, Jeff Haas, Job Snijders, Warren Kumari, Amogh Dhamdhere, Jakob Heitz, Geoff Huston, Randy Bush, Alexander Azimov, Ruediger Volk, Sue Hares, Wes George, Job Snijders, Chris Morrow, Sandy Murphy, Danny McPherson, and Eric Osterweil for comments, suggestions, and critique. The authors are also thankful to Padma Krishnaswamy, Oliver Borchert, and Okhee Kim for their review and comments.

10. References

10.1. 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, <<http://www.rfc-editor.org/info/rfc4271>>.

10.2. Informative References

- [Anwar] Anwar, R., Niaz, H., Choffnes, D., Cunha, I., Gill, P., and N. Katz-Bassett, "Investigating Interdomain Routing Policies in the Wild", ACM Internet Measurement Conference (IMC), October 2015, <<http://www.cs.usc.edu/assets/007/94928.pdf>>.

[Cowie2010]

Cowie, J., "China's 18 Minute Mystery", Dyn Research/Renesys Blog, November 2010, <<http://research.dyn.com/2010/11/chinas-18-minute-mystery/>>.

[Cowie2013]

Cowie, J., "The New Threat: Targeted Internet Traffic Misdirection", Dyn Research/Renesys Blog, November 2013, <<http://research.dyn.com/2013/11/mitm-internet-hijacking/>>.

[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>>.

[draft-ietf-idr-aspath-orf]

Patel, K. and S. Hares, "AS path Based Outbound Route Filter for BGP-4", IETF Internet Draft (expired), August 2007, <<https://tools.ietf.org/html/draft-ietf-idr-aspath-orf-09>>.

[draft-kunzinger-idrp-ISO10747-01]

Kunzinger, C., "Inter-Domain Routing Protocol (IDRP)", IETF Internet Draft (expired), November 1994, <<https://tools.ietf.org/pdf/draft-kunzinger-idrp-ISO10747-01.pdf>>.

[Gao]

Gao, L. and J. Rexford, "Stable Internet routing without global coordination", IEEE/ACM Transactions on Networking, December 2001, <<http://www.cs.princeton.edu/~jrex/papers/sigmetrics00.long.pdf>>.

[Gill]

Gill, P., Schapira, M., and S. Goldberg, "A Survey of Interdomain Routing Policies", ACM SIGCOMM Computer Communication Review, January 2014, <<https://www.cs.bu.edu/~goldbe/papers/survey.pdf>>.

[Giotsas]

Giotsas, V. and S. Zhou, "Valley-free violation in Internet routing - Analysis based on BGP Community data", IEEE ICC 2012, June 2012.

- [Hiran] Hiran, R., Carlsson, N., and P. Gill, "Characterizing Large-scale Routing Anomalies: A Case Study of the China Telecom Incident", PAM 2013, March 2013, <<http://www3.cs.stonybrook.edu/~phillipa/papers/CTelecom.html>>.
- [Huston2012] Huston, G., "Leaking Routes", March 2012, <<http://labs.apnic.net/blabs/?p=139/>>.
- [Huston2014] Huston, G., "What's so special about 512?", September 2014, <<http://labs.apnic.net/blabs/?p=520/>>.
- [I-D.ietf-sidr-bgpsec-protocol] Lepinski, M. and K. Sriram, "BGPsec Protocol Specification", draft-ietf-sidr-bgpsec-protocol-22 (work in progress), January 2017.
- [I-D.ymbk-idr-bgp-open-policy] Azimov, A., Bogomazov, E., Bush, R., Patel, K., and K. Sriram, "Route Leak Detection and Filtering using Roles in Update and Open messages", draft-ymbk-idr-bgp-open-policy-02 (work in progress), November 2016.
- [Kapela-Pilosov] Pilosov, A. and T. Kapela, "Stealing the Internet: An Internet-Scale Man in the Middle Attack", DEFCON-16 Las Vegas, NV, USA, August 2008, <<https://www.defcon.org/images/defcon-16/dc16-presentations/defcon-16-pilosov-kapela.pdf>>.
- [Kephart] Kephart, N., "Route Leak Causes Amazon and AWS Outage", ThousandEyes Blog, June 2015, <<https://blog.thousandeyes.com/route-leak-causes-amazon-and-aws-outage>>.
- [Khare] Khare, V., Ju, Q., and B. Zhang, "Concurrent Prefix Hijacks: Occurrence and Impacts", IMC 2012, Boston, MA, November 2012, <<http://www.cs.arizona.edu/~bzhang/paper/12-imc-hijack.pdf>>.
- [Labovitz] Labovitz, C., "Additional Discussion of the April China BGP Hijack Incident", Arbor Networks IT Security Blog, November 2010, <<http://www.arbornetworks.com/asert/2010/11/additional-discussion-of-the-april-china-bgp-hijack-incident/>>.

- [LRL] Khare, V., Ju, Q., and B. Zhang, "Large Route Leaks", Project web page, 2012, <<http://nrl.cs.arizona.edu/projects/lsrl-events-from-2003-to-2009/>>.
- [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>>.
- [Madory] Madory, D., "Why Far-Flung Parts of the Internet Broke Today", Dyn Research/Renesys Blog, September 2014, <<http://research.dyn.com/2014/09/why-the-internet-broke-today/>>.
- [Mauch] Mauch, J., "BGP Routing Leak Detection System", Project web page, 2014, <<http://puck.nether.net/bgp/leakinfo.cgi/>>.
- [Mauch-nanog] Mauch, J., "Detecting Routing Leaks by Counting", NANOG-41 Albuquerque, NM, USA, October 2007, <<https://www.nanog.org/meetings/nanog41/presentations/mauch-lightning.pdf>>.
- [Nanog-thread-June2016] "Intra-AS messaging for route leak prevention", NANOG Email List - Discussion Thread , June 2016, <<http://mailman.nanog.org/pipermail/nanog/2016-June/thread.html#86348>>.
- [NIST-800-54] Kuhn, D., Sriram, K., and D. Montgomery, "Border Gateway Protocol Security", NIST Special Publication 800-54, July 2007, <<http://csrc.nist.gov/publications/nistpubs/800-54/SP800-54.pdf>>.
- [Paseka] Paseka, T., "Why Google Went Offline Today and a Bit about How the Internet Works", CloudFare Blog, November 2012, <<http://blog.cloudflare.com/why-google-went-offline-today-and-a-bit-about/>>.
- [proceedings-sixth-ietf] Gross, P., "Proceedings of the April 22-24, 1987 Internet Engineering Task Force", April 1987, <<https://www.ietf.org/proceedings/06.pdf>>.

- [RFC1105-obsolete]
Lougheed, K. and Y. Rekhter, "A Border Gateway Protocol (BGP)", IETF RFC (obsolete), June 1989, <<https://tools.ietf.org/html/rfc1105>>.
- [RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, DOI 10.17487/RFC6811, January 2013, <<http://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, <<http://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, <<http://www.rfc-editor.org/info/rfc7908>>.
- [Snijders]
Snijders, J., "Practical everyday BGP filtering with AS_PATH filters: Peer Locking", NANOG-47 Chicago, IL, USA, June 2016, <https://www.nanog.org/sites/default/files/Snijders_Everyday_Practical_Bgp.pdf>.
- [Sriram] Sriram, K., Montgomery, D., Dickson, B., Patel, K., and A. Robachevsky, "Methods for Detection and Mitigation of BGP Route Leaks", IETF-95 IDR WG Meeting), April 2016, <<https://www.ietf.org/proceedings/95/slides/slides-95-idr-13.pdf>>.
- [Toonk] Toonk, A., "What Caused Today's Internet Hiccup", August 2014, <<http://www.bgppmon.net/what-caused-todays-internet-hiccup/>>.
- [Toonk2015-A]
Toonk, A., "What caused the Google service interruption", March 2015, <<http://www.bgppmon.net/what-caused-the-google-service-interruption/>>.
- [Toonk2015-B]
Toonk, A., "Massive route leak causes Internet slowdown", June 2015, <<http://www.bgppmon.net/massive-route-leak-cause-internet-slowdown/>>.

[Wijchers]

Wijchers, B. and B. Overeinder, "Quantitative Analysis of BGP Route Leaks", RIPE-69, November 2014, <<https://ripe69.ripe.net/presentations/157-RIPE-69-Routing-WG.pdf>>.

[Zmijewski]

Zmijewski, E., "Indonesia Hijacks the World", Dyn Research/Renesys Blog, April 2014, <<http://research.dyn.com/2014/04/indonesia-hijacks-world/>>.

Authors' Addresses

Kotikalapudi Sriram
US NIST

Email: ksriram@nist.gov

Doug Montgomery
US NIST

Email: doug@nist.gov

Brian Dickson

Email: brian.peter.dickson@gmail.com

Keyur Patel
Arrcus

Email: keyur@arrcus.com

Andrei Robachevsky
Internet Society

Email: robachevsky@isoc.org

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 25, 2021

R. Bush
Internet Initiative Japan
J. Haas
J. Scudder
Juniper Networks, Inc.
A. Nipper
C. Dietzel
DE-CIX
September 21, 2020

Making Route Servers Aware of Data Link Failures at IXPs
draft-ietf-idr-rs-bfd-09

Abstract

When BGP route servers are used, the data plane is not congruent with the control plane. Therefore, peers at an Internet exchange can lose data connectivity without the control plane being aware of it, and packets are lost. This document proposes the use of a newly defined BGP Subsequent Address Family Identifier (SAFI) both to allow the route server to request its clients use BFD to track data plane connectivity to their peers' addresses, and for the clients to signal that connectivity state back to the route server.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <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 March 25, 2021.

Copyright Notice

Copyright (c) 2020 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	2
2. Definitions	3
3. Overview	4
4. Next Hop Validation	5
4.1. ReachAsk	6
4.2. LocReach	6
4.3. ReachTell	7
4.4. NHIB	7
5. Advertising NH-Reach state in BGP	7
6. Client Procedures for NH-Reach Changes	9
7. Recommendations for Using BFD	9
8. Other Considerations	10
9. Acknowledgments	10
10. IANA Considerations	10
11. Security Considerations	10
12. References	11
12.1. Normative References	11
12.2. Informative References	12
Appendix A. Summary of Document Changes	12
Appendix B. Other Forms of Connectivity Checks	12
Authors' Addresses	13

1. Introduction

In configurations (typically Internet Exchange Points (IXPs)) where EBGp routing information is exchanged between client routers through the agency of a route server (RS) [RFC7947], but traffic is exchanged directly, operational issues can arise when partial data plane connectivity exists among the route server client routers. Since the

data plane is not congruent with the control plane, the client routers on the IXP can lose data connectivity without the control plane - the route server - being aware of it, resulting in significant data loss.

To remedy this, two basic problems need to be solved:

1. Client routers must have a means of verifying connectivity amongst themselves, and
2. Client routers must have a means of communicating the knowledge of the failure (and restoration) back to the route server.

The first can be solved by application of Bidirectional Forwarding Detection [RFC5880]. The second can be solved by exchanging BGP routes which use the NH-Reach Subsequent Address Family Identifier (SAFI) defined in this document.

Throughout this document, we generally assume that the route server being discussed is able to represent different RIBs towards different clients, as discussed in section 2.3.2.1 of [RFC7947]. If this is not the case, the procedures described here to allow BFD to be automatically provisioned between clients still have value; however, the procedures for signaling reachability back to the route server may not.

Throughout this document, we refer to the "route server", "RS" or just "server" and the "client" to describe the two BGP routers engaging in the exchange of information. We observe that there could be other applications for this extension. Our use of terminology is intended for clarity of description, and not to limit the future applicability of the proposal.

[I-D.ietf-idr-bgp-bestpath-selection-criteria] discusses enhancement of the route resolvability condition of section 9.1.2.1 of [RFC4271] to include next hop reachability and path availability checks. This specification represents in part an instance of such, implemented using BFD as the OAM mechanism.

2. Definitions

- o Indirect peer: If a route server is configured such that routes from a given client might be sent to some other client, or vice-versa, those two clients are considered to be indirect peers.
- o Indirect Peer's Address, IPA, next hop: We refer frequently to a next hop. It should generally be clear from context what is intended, almost always an address associated with an indirect peer (the exception, when an indirect peer sends a third party next hop, is discussed in Section 3). In Section 5 we discuss the

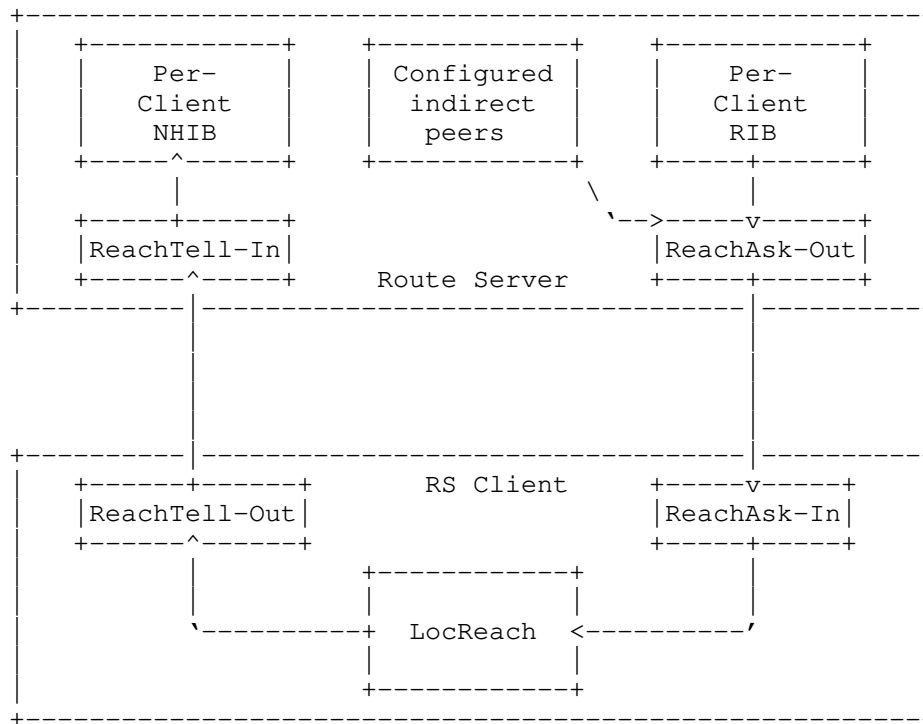
MP-BGP [RFC4760] Next Hop field; this is distinguished by its capitalization and should also be clear from context. Later in that section we define the Indirect Peer's Address field of the NLRI, also called "IPA". It will be clear to the reader that this refers to the "next hops" discussed elsewhere in the document, but we don't use the name "next hop" for this field to avoid confusion with the pre-existing next hop path attribute of [RFC4271] and attribute field of [RFC4760].

- o RS: Route Server. See [RFC7947].

3. Overview

As with the base BGP protocol, we model the function of this extension as the interaction between a conceptual set of databases:

- o ReachAsk: The reachability request database. A database of next hops (host addresses) for which data plane reachability is being queried.
- o ReachAsk-Out: A set of queries sent to the client.
- o ReachAsk-In: A set of queries received from the route server.
- o ReachTell: The reachability response database. A database of responses to ReachAsk queries, indicating what is known about data plane reachability.
- o ReachTell-Out: The responses being sent to the route server.
- o ReachTell-In: The response received from the client.
- o LocReach: The local reachability database.
- o NHIB: Next Hop Information Base. Stores what is known about the client's reachability to its next hops.



Route Server, RS Client, and Reachability Ask and Tell databases with In/Out Queues

In outline, the route server requests its client to track connectivity for all the potential next hops the RS might send to the client, by sending these next hops as ReachAsk "routes". The client tracks connectivity using BFD and reports its connectivity status to the RS using ReachTell "routes". Connectivity status may be that the next hop is reachable, unreachable, or unknown. Once the RS has been informed by the client of its connectivity, it uses this information to influence the route selection the RS performs on behalf of the client. Details are elaborated in the following sections.

4. Next Hop Validation

Below, we detail procedures where a route server tells its client router about other client next hops by sending it ReachAsk routes and the client router verifies connectivity to those other client routers and communicates its findings back to the RS using ReachTell routes. The RS uses the received ReachTell routes as input to the NHIB and hence the route selection process it performs on behalf of the client.

4.1. ReachAsk

The route server maintains a ReachAsk database for each client that supports this proposal, that is, for each client that has advertised support (Section 5) for the NH-Reach SAFI. This database is the union of:

- o The set of next hops found in the associated per-client Loc-RIB (see section 2.3.2.1 of [RFC7947]).
- o The set of addresses of this client's indirect peers (Section 2).
- o The RS MAY also add other entries, for example under configuration control.

We note that under most circumstances, the first (Loc-RIB next hops) set will be a subset of the second (indirect peers) set. For this not to be the case, a client would have to have sent a "third party" next hop [RFC4271] to the server. To cover such a case, an implementation MAY note any such next hops, and include them in its list of indirect peers. (This implies that if a third party next hop for client C is conveyed to client A, not only will C be placed in A's ReachAsk database, but A will be placed in C's ReachAsk database.)

The contents of the ReachAsk database are communicated to the client using the NLRI format and procedures described in Section 5.

4.2. LocReach

The client MUST attempt to track data plane connectivity to each host address depicted in the ReachAsk database. It MAY also track connectivity to other addresses. The use of BFD for this purpose is detailed in Section 6.

For each address being tracked, its state is maintained by the client in a LocReach entry. The state can be:

- o Unknown. Connectivity status is unknown. This may be due to a temporary or permanent lack of feasible OAM mechanism to determine the status.
- o Up. The address has been determined to be reachable.
- o Down. The address has been determined to be unreachable.

The LocReach database is used as input for the ReachTell database; it MAY also be used as input to the client's route resolvability condition (section 9.1.2.1 of [RFC4271]).

4.3. ReachTell

The ReachTell database contains an entry for every entry in the LocReach database.

The contents of the ReachTell database are communicated to the server using the NLRI format and procedures described in Section 5.

4.4. NHIB

The route server maintains a per-client Next Hop Information Base, or NHIB. This contains the information about next hop status received from ReachTell.

In computing its per-client Loc-RIB, the RS uses the content of the related per-client NHIB as input to the route resolvability condition (section 9.1.2.1 of [RFC4271]). The next hop being resolved is looked up in the NHIB and its state determined:

- o Up next hops are considered resolvable.
- o Unknown next hops MAY be considered resolvable. They MAY be less preferred for selection.
- o Down next hops MUST NOT be considered resolvable.
- o If a given next hop is not present in the NHIB, but is present in ReachAsk-Out, either the client has not responded yet (a transient condition) or an error exists. Similar to Unknown next hops, such routes MAY be considered resolvable; they MAY be less preferred.

5. Advertising NH-Reach state in BGP

A new BGP SAFI, the NH-Reach SAFI, is defined in this document. It has been assigned value TBD. A route server or a route server client using the procedures in this document MUST advertise support for this SAFI, for the IPv4 and/or IPv6 Address Family Identifier (AFI). The use of this SAFI with any other AFI is not defined by this document.

NH-Reach NLRI "routes" have a Length of Next Hop Network Address value of 0, therefore they have an empty Network Address of Next Hop field (section 3 of [RFC4760]).

Since as specified here, ReachTell "routes" from different clients populate distinct databases on the RS, there will generally be only a single path per "route"; this implies that route selection need not be performed (or equivalently, that it's trivial to perform).

In the other direction, a client might peer with multiple route servers and receive differing sets of ReachAsk routes from them. An implementation MAY handle this situation by implementing a distinct

ReachAsk and ReachTell per server, but it MAY also handle it by placing all servers' ReachAsk "routes" into a single ReachAsk, and sending the results to all servers from a single ReachTell. This would imply some route server(s) might get ReachTell results they had not asked for, but this is permissible in any case. Again, since the contents of ReachAsk are simply a set of host routes to be tested, route selection over a combined ReachAsk MAY be omitted.

ReachAsk and ReachTell entries are exchanged using the NH-Reach NLRI encoding:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|T|Reserved|Sta|Indirect Peer's Address (4 or 16 octets)|
+-----+-----+-----+-----+-----+-----+-----+-----+
.      ... Indirect Peer's Address (4 or 16 octets) ...      .
.
+-----+-----+-----+-----+-----+-----+-----+-----+

```

NH-Reach NLRI Format

- o T: Type is a one-bit field that can take the value 0, meaning the NLRI is a ReachAsk entry, or 1, meaning it is a ReachTell entry.
- o Reserved: These five bits are reserved. They MUST be sent as zero and MUST be disregarded on receipt.
- o Sta: State is a two-bit field used to signal the LocReach (Section 4.2) state:
 - * 0 or 3: Unknown.
 - * 1: Up.
 - * 2: Down.

Although either 0 or 3 is to be interpreted as "Unknown", the value 0 MUST be used on transmission. The value 3 MUST be accepted as an alias for 0 on receipt.

- o The Indirect Peer's Address ("IPA") field is an IPv4 or IPv6 host route, depending on whether the AFI is IPv4 or IPv6.

ReachAsk and ReachTell entries MUST NOT be propagated from one BGP peering session to another; the routes are not transitive.

The IPA field is the key for the NH-Reach NLRI type; the information encoded in the top octet is non-key information. It is possible in principle (although unlikely) for two NLRI to be validly present in an UPDATE message with identical IPA fields but different types. However, two NLRI with the same IPA field and different State fields MUST NOT be encoded in the same UPDATE message. If such is

encountered, the receiver MUST behave as though the state "Unknown" was received for the IPA in question.

6. Client Procedures for NH-Reach Changes

When an entry is added to a route server client's ReachAsk-In for a route server peering session, the client will then attempt to verify connectivity to the host depicted by that entry. The procedure described in this specification utilizes BFD.

If no existing BFD session exists to this next hop, a BFD session is provisioned to that IP address and the LocReach reachability state (Section 4.2) is set to Unknown.

If the client cannot establish a BFD session with an entry in its ReachAsk-In, the next hop remains in LocReach with its Reachable state Unknown.

Once the BFD session moves to the Up state, the LocReach reachability state is set to Up.

When the BFD session transitions out of the Up state to the Down state, the LocReach reachability state is set to Down.

If the BFD session transitions out of the Up state to the AdminDown state, the LocReach reachability state is set to Unknown.

When entries are removed from the route server client's ReachAsk-In for a route server peering session, the client MAY delay de-provisioning the BFD peering session. If the client delays de-provisioning the session, it should remove it if the BFD session transitions to the Down or AdminDown states.

7. Recommendations for Using BFD

The RECOMMENDED way a client router can confirm the data plane connectivity to its next hops is available, is the use of BFD in asynchronous mode. Echo mode MAY be used if both client routers running a BFD session support this. The use of authentication in BFD is OPTIONAL as there is a certain level of trust between the operators of the client routers at a particular IXP. If trust cannot be assumed, it is recommended to use pair-wise keys (how this can be achieved is outside the scope of this document). The ttl/hop limit values as described in section 5 [RFC5881] MUST be obeyed in order to shield BFD sessions against packets coming from outside the IXP.

The following values of the BFD configuration of client routers (see section 6.8.1 [RFC5880]) are RECOMMENDED:

- o DesiredMinTxInterval: 1,000,000 (microseconds)
- o RequiredMinRxInterval: 1,000,000 (microseconds)
- o DetectMult: 3

A client router administrator MAY select more appropriate values to meet the special needs of a particular deployment.

8. Other Considerations

For purposes of routing stability, implementations may wish to apply hysteresis ("holddown") to next hops that have transitioned from reachable to unreachable and back.

Implementations MAY restrict the range of addresses with which they will attempt to form BFD relationships. For example, an implementation might by default only allow BFD relationships with peers that share a subnet with the route server. An implementation MAY apply such restrictions by default.

In a route-server environment, use of this feature SHOULD be restricted to consider only routes that are advertised from within the IXP network. This might include checks on AS_PATH length.

9. Acknowledgments

The authors would like to thank Thomas King for his contributions toward this work.

10. IANA Considerations

IANA is requested to allocate a value from the Subsequent Address Family Identifiers (SAFI) Parameters registry for this proposal. Its Description in that registry shall be NH-Reach with a Reference of this RFC.

11. Security Considerations

The mechanism in this document permits a route server client to influence the contents of the route server's Adj-Ribs-Out through its reports of next hop reachability state using the NH-Reach SAFI. Since this state is per-client, if a route server client is able to inject NH-Reach routes for another route server's BGP session to a client, it can cause the route server to select different forwarding than otherwise expected. This issue may be mitigated using transport security on the BGP sessions between the route server and its clients. See [RFC4272].

The NH-Reach SAFI enables the server to trigger creation of a BFD session on its client. A malicious or misbehaving server could trigger an unreasonable number of sessions, a potential resource exhaustion attack. The sedate default timers proposed in Section 7 mitigate this; they also mitigate concerns about use of the client as a source of packets in a flooding attack. An implementation MAY also impose limits on the number of BFD sessions it will create at the request of the server.

The reachability tests between route server clients themselves may be a target for attack. Such attacks may include forcing a BFD session Down through injecting false BFD state. A less likely attack includes forcing a BFD session to stay Up when its real state is Down. These attacks may be mitigated using the BFD security mechanisms defined in [RFC5880].

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <<https://www.rfc-editor.org/info/rfc5880>>.
- [RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, DOI 10.17487/RFC5881, June 2010, <<https://www.rfc-editor.org/info/rfc5881>>.
- [RFC7947] Jasinska, E., Hilliard, N., Raszuk, R., and N. Bakker, "Internet Exchange BGP Route Server", RFC 7947, DOI 10.17487/RFC7947, September 2016, <<https://www.rfc-editor.org/info/rfc7947>>.

12.2. Informative References

- [I-D.chen-bfd-unsolicited]
Chen, E., Shen, N., and R. Raszuk, "Unsolicited BFD for Sessionless Applications", draft-chen-bfd-unsolicited-02 (work in progress), January 2018.
- [I-D.ietf-idr-bgp-bestpath-selection-criteria]
Asati, R., "BGP Bestpath Selection Criteria Enhancement", draft-ietf-idr-bgp-bestpath-selection-criteria-12 (work in progress), June 2019.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<https://www.rfc-editor.org/info/rfc4272>>.
- [RFC7880] Pignataro, C., Ward, D., Akiya, N., Bhatia, M., and S. Pallagatti, "Seamless Bidirectional Forwarding Detection (S-BFD)", RFC 7880, DOI 10.17487/RFC7880, July 2016, <<https://www.rfc-editor.org/info/rfc7880>>.

Appendix A. Summary of Document Changes

idr-06: Refresh -05.
idr-04 to idr-05: Added reference to "BGP Bestpath Selection Criteria Enhancement" draft. Rename "next hop" field of NLRI to "Indirect Peer's Address". Add suggestion about AS_PATH length checks.
idr-03 to idr-04: Note other forms of connectivity checks.
idr-02 to idr-03: Substantial rewrite. Introduce NLRI format that embeds state.
idr-01 to idr-02: Move from BGP-LS to NH-Reach SAFI. Lots of editorial changes.
idr-00 to idr-01: Add BGP Capability. Move from NH-Cost to BGP-LS.
ymbk-01 to idr-00: No technical changes; adopted by IDR.
ymbk-00 to ymbk-01: Clarifications to BFD procedures. Use BFD state as an input to BGP route selection.

Appendix B. Other Forms of Connectivity Checks

RFC 5880/5881 BFD is a well-deployed feature. For this reason, it was chosen as the connectivity check utilized for nexthop reachability by this document. As other forms of BFD become more widely deployed, they may also be utilized to provide the connectivity check functionality.

Examples of other such BFD mechanisms include:

- o Seamless BFD [RFC7880]
- o Unsolicited BFD for Sessionless Applications
[I-D.chen-bfd-unsolicited]

Implementations MUST support RFC 5880/5881 BFD to be compliant with this specification. Implementations MAY support other forms of connectivity check, including those mechanisms listed above, so long as they provide the ability to fall-back to RFC 5880/5881 BFD.

Authors' Addresses

Randy Bush
Internet Initiative Japan
5147 Crystal Springs
Bainbridge Island, Washington 98110
US

Email: randy@psg.com

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

Email: jhaas@juniper.net

John G. Scudder
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089
US

Email: jgs@juniper.net

Arnold Nipper
DE-CIX Management GmbH
Lichtstrasse 43i
Cologne 50825
Germany

Email: arnold.nipper@de-cix.net

Internet-Draft Making RSeS aware of IXP Data Link FailuresSeptember 2020

Christoph Dietzel
DE-CIX Management GmbH
Lichtstrasse 43i
Cologne 50825
Germany

Email: christoph.dietzel@de-cix.net

IDR
Internet-Draft
Updates: 5575 (if approved)
Intended status: Standards Track
Expires: September 4, 2018

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

Populate to FIB Action for FlowSpec
draft-li-idr-flowspec-populate-to-fib-02

Abstract

A bit, F bit, is defined in traffic action extended community, which is used by FlowSpec to indicate the associated specifications be populated in FIB (Forwarding Information Base) after appropriate process.

Status of This Memo

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

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

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

This Internet-Draft will expire on September 4, 2018.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	2
2. Requirements Language	3
3. Populate to FIB Action	3
4. Implementation Considerations	3
5. Security Considerations	4
6. IANA Considerations	4
7. Normative References	4
Authors' Addresses	5

1. Introduction

BGP FlowSpec [RFC5575] provides a flexible mechanism to distribute traffic flow specifications, where the matching rules are encoded in the Border Gateway Protocol Network Layer Reachability Information (BGP NLRI) with defined new format and the corresponding actions are encoded in BGP Extended communities.

In routers, traffic flow specifications distributed by BGP FlowSpec [RFC5575] are stored in distinct set of RIBs (Routing Information Base) according to their (AFI, SAFI) pairs. These RIBs are then populated to the dedicated hardware (most of them are TCAM based) usually shared with ACLs (Access Control Lists). The dedicated hardware is much more expensive and space limited when compared with the hardware used to store the FIB (Forwarding Information Base), which is usually sufficient to fit several millions of FIB entries. Although in some implementations, the hardware used to populate traffic flow specifications and FIB entries is the same, the size for each parts is fixed at design stage. As the number of ACL rules and FlowSpec specifications increases, especially when FlowSpec is used for dynamic traffic flow steering, which is one of the three BGP FlowSpec applications listed in [RFC5575] and [I-D.ietf-idr-rfc5575bis], hardware space requirement of FlowSpec specifications in the field network may exceed the size of the dedicated hardware. To save the limited and expensive space of the dedicated hardware, it is better to populate some FlowSpec specifications to FIB if possible. The destination prefix based FlowSpec specifications, for example, are suitable to be populated to FIB.

However, there is no method in the current version of BGP FlowSpec [RFC5575] and RFC5575bis [I-D.ietf-idr-rfc5575bis] to indicate the associated specifications are suitable to be populated to FIB. This

document defines a new bit, F bit (populate to FIB), in 0x8007 traffic action extended community to satisfy the requirement.

2. Requirements Language

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

3. Populate to FIB Action

F bit, populate to FIB bit, is defined in 0x8007 traffic action extended community [RFC5575] to indicate the associated BGP FlowSpec specifications are suitable to be populated to FIB. Thus the space of the dedicated hardware that is used to store the BGP FlowSpec specifications can be saved for other kinds of BGP FlowSpec specifications and ACL rules.

The encoding format of the traffic action extended community with F bit is shown below. The F bit is solicited to be assigned by IANA.

40	41	42	43	44	45	46	47
+---+---+---+---+---+---+---+---+							
reserved					F	S	T
+---+---+---+---+---+---+---+---+							

Traffic-action extended community consists of 2 bytes for type and subtype, the value of which MUST be 0x8007, and 6 bytes for value, of which only the 3 least significant bits of the 6th byte (from left to right) are currently defined. S and T are defined in BGP FlowSpec [RFC5575]. F is defined as:

- o F: Populate to FIB Action (bit 45, to be assigned by IANA): When this bit is set, the associated BGP FlowSpec specifications SHOULD be populated to FIB. If not set, the associated BGP FlowSpec specifications MUST NOT be populated to FIB. If this bit is set and the associated BGP FlowSpec specifications can not be populated to FIB, the associated BGP FlowSpec specifications MUST be ignored.

4. Implementation Considerations

FlowSpec rules are ordering sensitive. After ordering processing as per section 5.1 of [RFC5575], they are searched sequentially until a matching rule is found. FIB entries, on the contrary, have no ordering implication. Longest prefix matching is the rule to choose the matching FIB entry. Only the destination prefix based, F bit tagged FlowSpec rules that pass the validation (as per section 6 of

[RFC5575]) and ordering (as per section 5.1 of [RFC5575]) processing are suitable to be populated into FIB. When populating a FlowSpec rule into FIB, the following facts have to be taken into account.

- o FlowSpec rules have higher priority than corresponding IGP and BGP routing entries.
- o When populating the FIB, the FlowSpec rules with F bit tagged are preferred than the corresponding IGP and BGP routing entries.
- o When a FlowSpec rule is being populated into FIB, the FIB entries, including those come from IGP or BGP updates, covered by this FlowSpec rule MUST be removed or replaced by this FlowSpec rule.
- o The populated FlowSpec rules in the FIB MUST not be overridden by IGP or BGP updates.

5. Security Considerations

This document defines a new bit in the traffic action extended community to indicate the associated BGP FlowSpec specifications SHOULD be populated to FIB directly. This bit does not introduce any new security issues. The same security considerations as for the BGP FlowSpec [RFC5575] applies.

6. IANA Considerations

One bit, F bit, is solicited to be assigned from Traffic Action Fields registry. This bit is used by BGP FlowSpec to indicate the associated BGP FlowSpec specifications SHOULD be populated to FIB directly.

7. Normative References

[I-D.ietf-idr-rfc5575bis]

Hares, S., Loibl, C., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", draft-ietf-idr-rfc5575bis-06 (work in progress), October 2017.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

[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>>.

Authors' Addresses

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

Email: li_zhenqiang@hotmail.com

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

Email: jie.dong@huawei.com

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

Email: zhuangshunwan@huawei.com

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

S. Previdi, Ed.
C. Filsfils
A. Sreekantiah
S. Sivabalan
Cisco Systems, Inc.
P. Mattes
Microsoft
E. Rosen
Juniper Networks
S. Lin
Google
February 24, 2017

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

Abstract

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

Status of This Memo

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

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

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

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

Copyright Notice

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

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

Table of Contents

1. Introduction	3
1.1. Requirements Language	4
2. SR TE Policy Encoding	4
2.1. SR TE Policy SAFI and NLRI	4
2.2. SR TE Policy and Tunnel Encapsulation Attribute	6
2.3. Remote Endpoint and Color	7
2.4. SR TE Policy Sub-TLVs	7
2.4.1. Preference sub-TLV	7
2.4.2. SR TE Binding SID Sub-TLV	8
2.4.3. Segment List Sub-TLV	9
3. Extended Color Community	21
4. SR Policy Operations	21
4.1. Configuration and Advertisement of SR TE Policies	21
4.2. Reception of an SR Policy	22
4.2.1. Acceptance of a SR Policy Update	22
4.2.2. Passing an acceptable path to an SR Policy	24
4.2.3. Propagation of an SR Policy	24
4.3. Steering Traffic into a SR Policy	24
4.4. Flowspec and SR Policies	24
5. Acknowledgments	24
6. Implementation Status	25
7. IANA Considerations	25
7.1. Existing Registry: Subsequent Address Family Identifiers (SAFI) Parameters	26
7.2. Existing Registry: BGP Tunnel Encapsulation Attribute Tunnel Types	26
7.3. Existing Registry: BGP Tunnel Encapsulation Attribute sub-TLVs	26
7.4. New Registry: SR Policy List Sub-TLVs	26
8. Security Considerations	27
9. References	27
9.1. Normative References	27
9.2. Informational References	28
Authors' Addresses	29

1. Introduction

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

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

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

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

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

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

Typically, a controller defines the set of policies and advertise them to BGP routers (typically ingress routers). The policy advertisement uses BGP extensions defined in this document. The policy advertisement is, in most but not all of the cases, tailored for the receiver. In other words, a policy advertised to a given BGP speaker has significance only for that particular router and is not intended to be propagated anywhere else. Then, the receiver of the policy instantiate the policy in its routing and forwarding tables and steer traffic into it based on both the policy and destination prefix color and next-hop.

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

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

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

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

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

- o A new Subsequent Address Family Identifier (SAFI) identifying the content of the BGP message (i.e.: the SR Policy).
- o A new NLRI identifying the SR Policy.
- o A set of new TLVs to be inserted into the Tunnel Encapsulation Attribute (as defined in [I-D.ietf-idr-tunnel-encaps]) and describing the SR Policy.
- o An IPv4 address format route-target extended community ([RFC4360]) attached to the SR Policy advertisement and that indicates the intended receiver of such SR Policy advertisement.
- o The Extended Color Community (as defined in [I-D.ietf-idr-tunnel-encaps]) and used in order to steer traffic into an SR Policy. This document (Section 3) modifies the format of the Extended Color Community by using the two leftmost bits of the RESERVED field.

1.1. Requirements Language

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

2. SR TE Policy Encoding

2.1. SR TE Policy SAFI and NLRI

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

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

Distinguisher (4 octets)
Policy Color (4 octets)
Endpoint (4 or 16 octets)

where:

- o Distinguisher: 4-octet value uniquely identifying the policy in the context of <color, endpoint> tuple. The distinguisher has no semantic and it's solely used by the SR Policy originator in order to make unique (from a NLRI perspective) multiple occurrences of the same SR Policy.
- o Policy Color: 4-octet value identifying (with the endpoint) the policy. The color is used to match the color of the destination prefixes in order to steer traffic into the SR Policy [I-D.filsfils-spring-segment-routing-policy].
- o Endpoint: identifies the endpoint of a policy. The Endpoint may represent a single node or a set of nodes (e.g.: an anycast address or a summary address). The Endpoint is an IPv4 (4-octet) address or an IPv6 (16-octet) address according to the AFI of the NLRI.

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

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

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

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

2.2. SR TE Policy and Tunnel Encapsulation Attribute

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

The SR Policy Encoding structure is as follows:

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>

Attributes:

 Tunnel Encaps Attribute (23)

 Tunnel Type: SR Policy

 Binding SID

 Preference

 Segment List

 Weight

 Segment

 Segment

 ...

 ...

where:

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

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

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

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

2.3. Remote Endpoint and Color

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

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

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

2.4. SR TE Policy Sub-TLVs

This section defines the SR Policy sub-TLVs.

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

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

2.4.1. Preference sub-TLV

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

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

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
Type										Length										Flags										RESERVED									
Preference (4 octets)																																							

where:

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

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

[I-D.filsfils-spring-segment-routing-policy] defines the use of the Preference value.

2.4.2. SR TE Binding SID Sub-TLV

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

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

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
Type										Length										Flags										RESERVED									
Binding SID (variable, optional)																																							

where:

- o Type: TBD4 (to be assigned by IANA from the "BGP Tunnel Encapsulation Attribute sub-TLVs" registry).
- o Length: specifies the length of the value field not including Type and Length fields. Can be 2 or 6 or 18.
- o Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o Binding SID: if length is 2, then no Binding SID is present. If length is 6 then the Binding SID contains a 4-octet SID. If length is 18 then the Binding SID contains a 16-octet IPv6 SID.

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

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

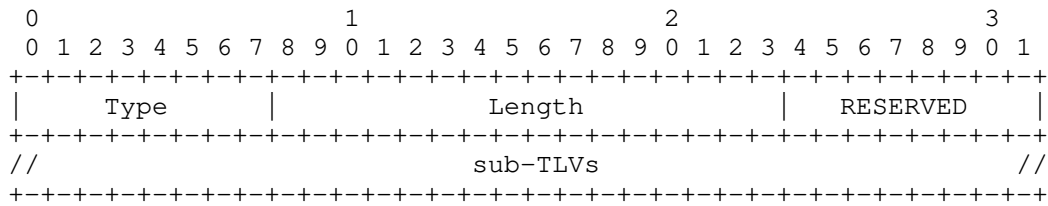
The Binding SID concept is defined in [I-D.ietf-spring-segment-routing] and its use in the context of SR Policies is defined in [I-D.filsfils-spring-segment-routing-policy].

2.4.3. Segment List Sub-TLV

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

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

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



where:

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

The Segment List sub-TLV is mandatory.

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

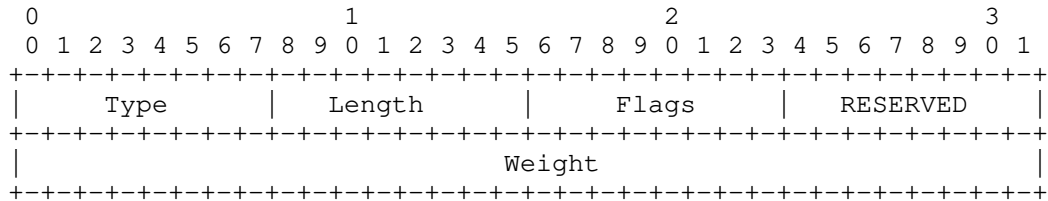
When multiple occurrences of the Segment List sub-TLV appear in the SR Policy, the traffic is load-balanced across them either through an ECMP scheme (if no Weight sub-TLV is present) or through a weighted ECMP scheme according to Section 2.4.3.1.

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

2.4.3.1. Weight Sub-TLV

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

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



where:

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

Length: 6.

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

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

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

2.4.3.2. Segment Sub-TLV

The Segment sub-TLV describes a single segment in a segment list (i.e.: a single element of the explicit path). Multiple Segment sub-TLVs constitute an explicit path of the SR Policy.

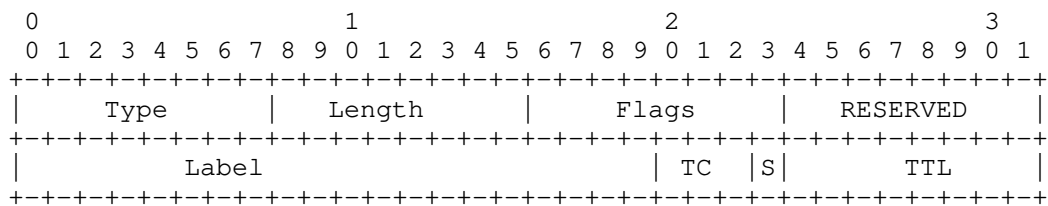
The Segment sub-TLV is mandatory and MAY appear multiple times in the Segment List sub-TLV.

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

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

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

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



where:

- o Type: 1 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- o Length is 6.
- o Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o Label: 20 bits of label value.
- o TC: 3 bits of traffic class.
- o S: 1 bit of bottom-of-stack.
- o TTL: 1 octet of TTL.

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

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

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

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

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

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   Type   |   Length   |   Flags   |   RESERVED   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
//                               IPv6 SID (16 octets)                               //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

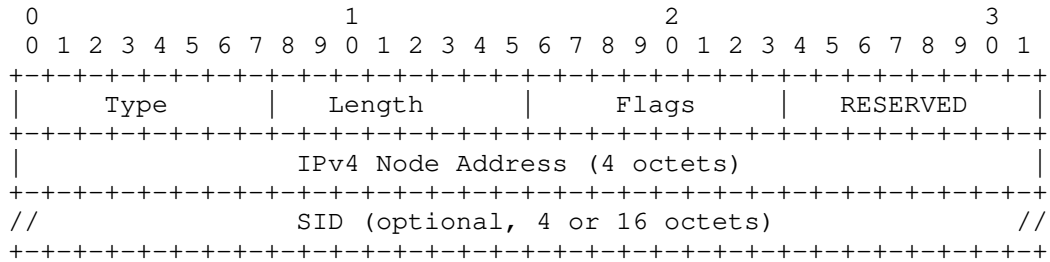
where:

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

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

2.4.3.2.3. Type 3: IPv4 Node Address with optional SID

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



where:

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

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

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

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

2.4.3.2.4. Type 4: IPv6 Node Address with optional SID

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

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Type      |      Length      |      Flags      | RESERVED |
+-----+-----+-----+-----+-----+-----+-----+-----+
//                IPv6 Node Address (16 octets)                //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                SID (optional, 4 or 16 octets)                //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

where:

- o Type: 4 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- o Length is 18 or 22 or 34.
- o Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o IPv6 Node Address: a 16 octet IPv6 address representing a node.
- o SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

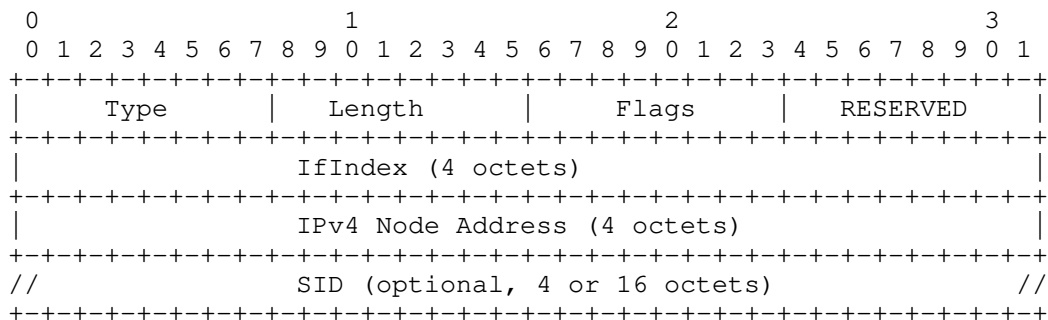
The following applies to the Type-4 Segment sub-TLV:

- o The IPv6 Node Address MUST be present.
- o The SID is optional and MAY be of one of the following formats:
 - * MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.
 - * IPV6 SID: a 16 octet IPv6 address.
- o If length is 18, then only the IPv6 Node Address is present.

- o If length is 22, then the IPv6 Node Address and the MPLS SID are present.
- o If length is 34, then the IPv6 Node Address and the IPv6 SID are present.

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

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



where:

- o Type: 5 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- o Length is 10 or 14 or 26.
- o Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o IfIndex: 4 octets of interface index.
- o IPv4 Node Address: a 4 octet IPv4 address representing a node.
- o SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

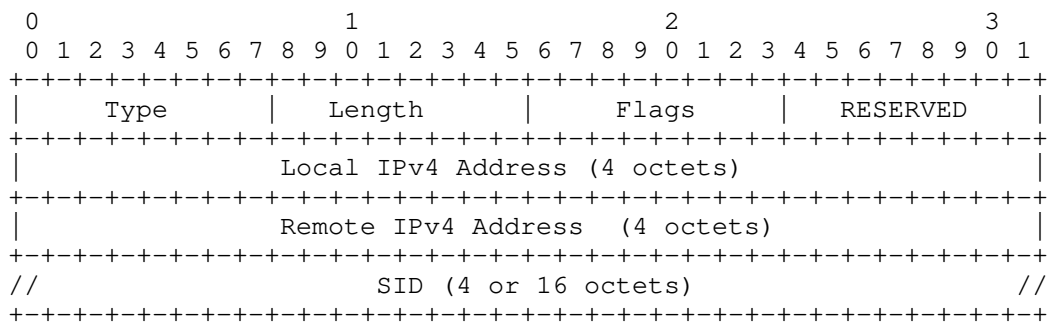
The following applies to the Type-5 Segment sub-TLV:

- o The IPv4 Node Address MUST be present.
- o The Interface Index (IfIndex) MUST be present.

- o The SID is optional and MAY be of one of the following formats:
 - * MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.
 - * IPV6 SID: a 16 octet IPv6 SID.
- o If length is 10, then the IPv4 Node Address and IfIndex are present.
- o If length is 14, then the IPv4 Node Address, the IfIndex and the MPLS SID are present.
- o If length is 26, then the IPv4 Node Address, the IfIndex and the IPv6 SID are present.

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

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



where:

- o Type: 6 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- o Length is 10 or 14 or 26.
- o Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.

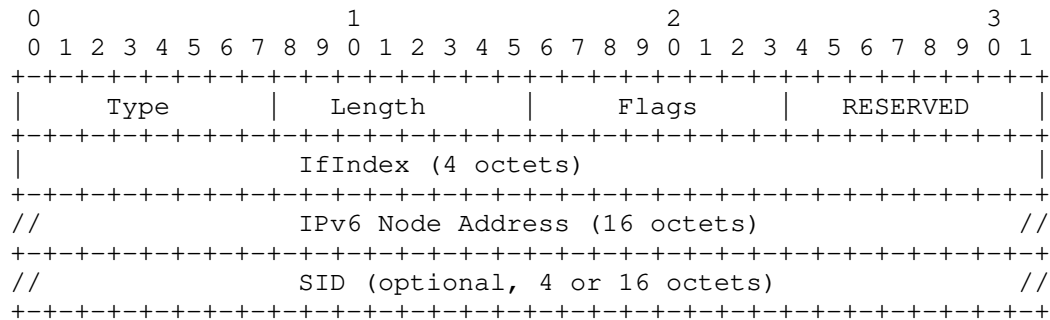
- o Local IPv4 Address: a 4 octet IPv4 address.
- o Remote IPv4 Address: a 4 octet IPv4 address.
- o SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

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

- o The Local IPv4 Address MUST be present and represents an adjacency local address.
- o The Remote IPv4 Address MUST be present and represents the remote end of the adjacency.
- o The SID is optional and MAY be of one of the following formats:
 - * MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.
 - * IPV6 SID: a 16 octet IPv6 address.
- o If length is 10, then only the IPv4 Local and Remote addresses are present.
- o If length is 14, then the IPv4 Local address, IPv4 Remote address and the MPLS SID are present.
- o If length is 26, then the IPv4 Local address, IPv4 Remote address and the IPv6 SID are present.

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

The Type-7 Segment Sub-TLV encodes an IPv6 node address, an interface index and an optional SID in the form of either an MPLS label or an IPv6 address. The format is as follows:



where:

- o Type: 7 (to be assigned by IANA from the registry "SR Policy List Sub-TLVs" defined in this document).
- o Length is 22 or 26 or 38.
- o Flags: 1 octet of flags. None is defined at this stage. Flags SHOULD be unset on transmission and MUST be ignored on receipt.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o IfIndex: 4 octets of interface index.
- o IPv6 Node Address: a 16 octet IPv6 address representing a node.
- o SID: either 4 octet MPLS SID or a 16 octet IPv6 SID.

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

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

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

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

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

0																1																2																3															
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9														
Type																Length																Flags																RESERVED															
Local IPv6 Address (16 octets)																																																															
Remote IPv6 Address (16 octets)																																																															
SID (4 or 16 octets)																																																															

where:

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

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

- o The Local IPv6 Address MUST be present and represents an adjacency local address.
- o The Remote IPv6 Address MUST be present and represents the remote end of the adjacency.
- o The SID is optional and MAY be of one of the following formats:
 - * MPLS SID: a 4 octet label containing label, TC, S and TTL as defined in Section 2.4.3.2.1.
 - * IPV6 SID: a 16 octet IPv6 address.
- o If length is 34, then only the IPv6 Local and Remote addresses are present.
- o If length is 38, then the IPv6 Local address, IPv4 Remote address and the MPLS SID are present.
- o If length is 50, then the IPv6 Local address, IPv4 Remote address and the IPv6 SID are present.

3. Extended Color Community

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

The RESERVED field is changed as follows:

```

      1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+---+---+---+---+---+---+---+---+---+
|C O|          RESERVED          |
+---+---+---+---+---+---+---+---+

```

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

4. SR Policy Operations

4.1. Configuration and Advertisement of SR TE Policies

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

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

When advertised in BGP, each SR path of the same SR Policy will have a different Distinguisher in order to prevent BGP selection among these SR paths along the distribution of BGP updates.

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

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

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

4.2. Reception of an SR Policy

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

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

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

4.2.1. Acceptance of a SR Policy Update

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

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

- o The SR Policy update MUST have either the NO_ADV community or at least one route-target extended community in IPv4-address format.
- o The Tunnel Encapsulation Attribute MUST be attached to the BGP Update and MUST have the Tunnel Type set to SR Policy (value to be assigned by IANA).
- o Within the SR Policy, at least one Segment List sub-TLV MUST be present.
- o Within the Segment List sub-TLV at least one Segment sub-TLV MUST be present.

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

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

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

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

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

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

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

4.2.2. Passing an acceptable path to an SR Policy

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

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

4.2.3. Propagation of an SR Policy

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

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

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

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

4.3. Steering Traffic into a SR Policy

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

4.4. Flowspec and SR Policies

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

5. Acknowledgments

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

6. Implementation Status

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

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

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

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

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

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

7. IANA Considerations

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

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

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

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

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

Codepoint	Description	Reference
73	SR Policy SAFI	This document

7.2. Existing Registry: BGP Tunnel Encapsulation Attribute Tunnel Types

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

Codepoint	Description	Reference
15	SR Policy Type	This document

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

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

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

7.4. New Registry: SR Policy List Sub-TLVs

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

Following Sub-TLV codepoints are defined:

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

8. Security Considerations

TBD.

9. References

9.1. Normative References

- [I-D.ietf-idr-tunnel-encaps]
 Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-03 (work in progress), November 2016.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <<http://www.rfc-editor.org/info/rfc4360>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<http://www.rfc-editor.org/info/rfc4760>>.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <<http://www.rfc-editor.org/info/rfc5226>>.

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

9.2. Informational References

- [I-D.filsfils-spring-segment-routing-policy]
Filsfils, C., Sivabalan, S., Yoyer, D., Nanduri, M., Lin, S., bogdanov@google.com, b., Horneffer, M., Clad, F., Steinberg, D., Decraene, B., and S. Litkowski, "Segment Routing Policy for Traffic Engineering", draft-filsfils-spring-segment-routing-policy-00 (work in progress), February 2017.
- [I-D.ietf-6man-segment-routing-header]
Previdi, S., Filsfils, C., Field, B., Leung, I., Linkova, J., Aries, E., Kosugi, T., Vyncke, E., and D. Lebrun, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header-05 (work in progress), February 2017.
- [I-D.ietf-idr-flowspec-redirect-ip]
Uttaro, J., Haas, J., Texier, M., Andy, A., Ray, S., Simpson, A., and W. Henderickx, "BGP Flow-Spec Redirect to IP Action", draft-ietf-idr-flowspec-redirect-ip-02 (work in progress), February 2015.
- [I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-11 (work in progress), February 2017.
- [I-D.ietf-spring-segment-routing-mpls]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Shakir, R., jeffrant@gmail.com, j., and E. Crabbe, "Segment Routing with MPLS data plane", draft-ietf-spring-segment-routing-mpls-07 (work in progress), February 2017.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <<http://www.rfc-editor.org/info/rfc4456>>.

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

Authors' Addresses

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

Email: sprevidi@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Brussels
BE

Email: cfilsfil@cisco.com

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

Email: asreekan@cisco.com

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

Email: msiva@cisco.com

Paul Mattes
Microsoft
One Microsoft Way
Redmond, WA 98052
USA

Email: pamattes@microsoft.com

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

Email: erosen@juniper.net

Steven Lin
Google

Email: stevenlin@google.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: February 19, 2020

A. Przygienda
Juniper
A. Lingala
AT&T
C. Mate
NIIF/Hungarnet
J. Tantsura
Nuage Networks
August 18, 2019

Compressed BGP Update Message
draft-przygienda-idr-compressed-updates-07

Abstract

This document provides specification of an optional compressed BGP update message format to allow family independent reduction in BGP control traffic volume.

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 February 19, 2020.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
2. Terminology	4
3. IANA Considerations	4
4. Procedures	5
4.1. Decompression Capability Negotiation	5
4.2. Compressed BGP Update Messages	5
4.3. Compressor Overflow	6
4.4. Compressor Restarts	7
4.5. Error Handling	7
5. Special Considerations	7
5.1. Impact on Network Sniffing Tools	7
6. Packet Formats	8
6.1. Decompressor Capability	8
6.2. Compressed Update Messages	8
7. Security Considerations	9
8. Acknowledgements	10
9. Normative References	10
Authors' Addresses	11

1. Introduction

BGP as a protocol evolved over the years to carry larger and larger volumes of information and this trend seems to continue unabated. And while lots of the growth can be contributed to the advent of new address families spurred by [RFC2283], steady increase in attributes and their size amplifies this tendency. Recently, even the same NLRI may be advertised multiple times by the means of ADD-PATH [RFC7911] extensions. All those developments drive up the volume of information BGP needs to exchange to synchronize RIBs of the peers.

Although BGP update format provides a simple "semantic" compression mechanism that avoids the repetition of attributes if multiple NLRIs share them already, in practical terms, the packing of updates has proven a difficult challenge. The packing attempts are further undermined by the plethora of "per NLRI-tagging" attributes such as extended communities [RFC4360].

One could of course dismiss the growing, raw volume of the data necessary to exchange BGP information between two peers as a mere trifle given the still rising link bandwidths, alas we are facing other sustained trends that would make the reduction of data volume exchanged by BGP highly desirable:

- o Link delays will remain constant until radically new transmission mechanisms become common place [QUANT]. Bare those developments, and given the prevailing constant ethernet MTU, increasing volume of BGP traffic will cause more and more IP packets to be sent with the BGP synchronization speed being limited by the expanding bandwidth-delay product.
- o The data volume, which for one peer may be reasonable, becomes less so when many of those need to be refreshed due to [RFC4724] and [RFC7313] interactions. Use of those techniques is expected to increase due to increasing demands on BGP reliability and novel variants of state synchronization between peers.
- o BGP message length is limited to 4K which in itself is a recognized problem. Extensions to the message length [ID.draft-ietf-idr-bgp-extended-messages-21] are being worked on but this puts its own requirements and memory pressure on the implementations and ultimately will not help with attributes exceeding 4K size limit in mixed environments.
- o Virtualization techniques introduce an increasing amount of context switches an IP packet has to cross between two BGP instances. Coupled with difficulties in estimating a reasonable TCP MSS in virtualized environments and the number of IP packets TCP generates, more and more context switching overhead per update is necessary before good-put BGP processing can happen.

Obviously, unless we change BGP encoding drastically by e.g. introducing more context to allow for semantic compression, we cannot expect a reduction in data volume without paying some kind of price. Ideas such as changing BGP format to allow for decoupling of attribute value updates from the NLRI updates could be a viable course of action. The challenges of such a scheme are significant and since such "compression" would extend the semantics and formats of the updates as we have them today, former and future drafts may interact with such an approach in ways not discernible today. Last but not least, attempting to introduce a smarter, context-rich encoding is likely to cause dependency problems and slow-down in BGP encoding procedures.

Fortunately, some observations can be made and emerging trends exploited to attempt a reduction in BGP data volumes without the mentioned disadvantages:

- o BGP updates are very repetitive. Smallest change in attribute values causes extensive repetition of all attributes and any difference prevents packing of NLRIs in same update. On top, each update message BGP still carries a marker that largely lost its practical value some time ago. One could generalize those facts by saying that BGP updates tend to exhibit very low entropy.
- o CPU cycles available to run control protocols are getting more and more abundant as does to a certain extent memory. They tend to not be available anymore in easily harvested "single core with higher frequency" form factors but as multiple cores that introduce the usual pitfalls of parallelization. In short, getting a lot of independent work done is getting cheaper and cheaper while speeding up a single strain of execution depending on previous results less so. This opens nevertheless the possibility to apply different filters on BGP streams, possibly even executing in parallel threads. One possible filter can compress the data in a manner completely transparent to the rest of existing implementation.

Hence, we suggest in this document the removal of redundancy in the BGP update stream via Huffman codes which can be applied as filter to a BGP update stream concurrently to the rest of the BGP processing and per peer. Subsequently, this document describes an optional scheme to compress BGP update traffic with a deflate variant of Huffman encoding [RFC1950], [RFC1951].

In broadest terms, such a scheme will be beneficial if a BGP implementation finds itself in an I/O constrained scenario while having spare CPU cycles disponible. Compression will ease the pressure on TCP processing and synchronization as well as reduce raw number of IP packets exchanged between peers.

2. Terminology

3. IANA Considerations

This document will request IANA to assign new BGP message type value and a new optional capability value in the BGP Capability Codes registry. The suggested value for the Compressed Updates message type in this process will be 7 and for the Capability Code the suggested value will be 76.

IANA will be requested as well to assign a new subcode in the "BGP Cease NOTIFICATION message subcodes" registry. The suggested name for the code point will be "Decompression Error". The suggested value will be 10.

4. Procedures

4.1. Decompression Capability Negotiation

The capability to *decompress* a new, optional message type carrying compressed updates is advertised via the usual BGP optional capability negotiation technique.

A peer MUST NOT send any compressed updates towards peers that did not advertise the capability to decompress. A peer MAY send compressed updates towards peers that advertised such capability.

4.2. Compressed BGP Update Messages

A new BGP message is introduced under the name of "Compressed BGP Update". It contains inside arbitrary number of following message types

- o normal BGP updates
- o Enhanced Route Refresh [RFC7313] subtype 1 and 2 (BoRR and EoRR)
- o Route Refresh with Options [ID.draft-idr-bgp-route-refresh-options-03] subtype 4 and 5 (BoRR and EoRR with options)

following each other and compressed while following the rules below:

1. Compressed and uncompressed BGP updates MAY follow each other in arbitrary order with exception of compressor overflow scenario per Section 4.3.
2. After decompression of the stream of interleaved compressed and uncompressed BGP update messages the resulting uncompressed sequence does not have to be identical to the sequence in a stream that would be generated without compression. However, the processing of the uncompressed sequence MUST ensure that the ultimate semantics of the message stream is the same to the peer as of a correct uncompressed case.
3. The sender is explicitly permitted to generate outgoing updates in a manner that reorders them as compared to uncompressed stream, but if it does so it MUST ensure that the resulting

stream of updates retains the original semantics as if compression was not in use.

4. The updates and refreshes contained within the compressed BGP update message MUST be stripped of the initial marker while preserving the BGP update or route refresh message header. The length field in the BGP header retains its original value.
5. Each compressed BGP Update MUST carry a sequence of non-fragmented original messages, i.e. it cannot e.g. contain a part of an original BGP update. Section 4.3 presents the only exception to this rule.
6. Each compressed BGP Update MUST be sent as a block, i.e. the decompression MUST be able to yield decompressed results of the update without waiting for further compressed updates. This is different from the normally used stream compression mode. Section 4.3 presents the only exception to this rule.
7. The compressed update message MAY exceed the maximum message size but in such case compressor overflow per Section 4.3 MUST be invoked.

4.3. Compressor Overflow

To achieve optimal compression rates it is desirable to provide to the compressor enough data so the resulting compressed update is as close to the maximum BGP update size as possible. Unfortunately, a Huffman with adapting dictionary compresses at always varying ratio which can lead to an overflow unless it is used very conservatively. A special provision, optionally to be used at the sender's discretion, allows for such overruns and simplifies the handling of overflow events.

In case the compressed block size exceeds the maximum BGP update size, the compressing peer MUST set the according bit in the compressed update generated and MUST proceed it with one and only one compressed update with the overflow and compressor restart bit cleared and the remainder of the block. No other BGP update messages are allowed in the TCP stream between the compressed update of a certain compressor and its overflow fragment. In case of any deviations, the error procedures of Section 4.5 MUST be followed.

The receiving peer MUST concatenate the first compressed update and the following overflow update as a single compressed block and apply decompression to it.

The first update MAY be smaller than the maximum BGP update size.

4.4. Compressor Restarts

In certain scenarios it is beneficial for the compressing peer to be able to restart any of the compressors at any point in the ongoing BGP session. To indicate such an occurrence, each compressed update CAN carry a flag signaling to the decompressing peer that it MUST restart the given de-compressor before attempting to handle the update.

4.5. Error Handling

If the decompression fails for any reason, the failure MUST cause immediate CEASE notification with a newly introduced subcode of "Decompression Error" (as documented in the IANA BGP Error Codes registry). The peer which experienced the failure MAY initiate the connection again but it SHOULD NOT advertise the decompressor capability until an administrative reset of the session or re-configuration of the peer. This will achieve self-stabilization of the feature in case of implementation problems.

The compressing peer MAY send such CEASE notification as well and close the peer. It is at the discretion of the decompressing peer given such a notification to omit the decompression capability on the next OPEN.

5. Special Considerations

5.1. Impact on Network Sniffing Tools

Network sniffing tool today have the capability to monitor an ongoing BGP session and try to reconstruct the state of the peers from the updates parsed. Obviously, with compression enabled, such a monitor cannot follow the compressed updates unless the session is monitored from the first compressed update on.

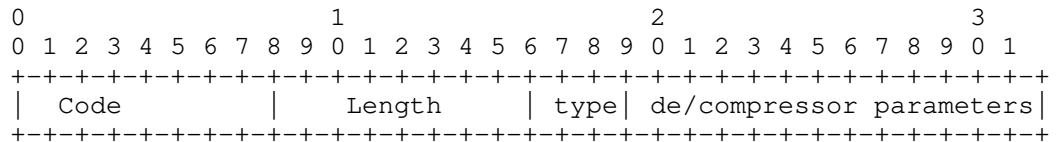
Several possibilities to deal with the problem exist, the simplest one being the restart of the compressors on a periodic basis to allow the monitoring tool to 'sync up'. It goes without saying that this will be detrimental to the compression ratio achieved.

Another possibility would have been to periodically send the Huffman dictionary over the wire but this complexity has been left out as to not overburden this specification. Moreover, at the current time, such a capability is not part of any standard Huffman implementation that could be easily referred to.

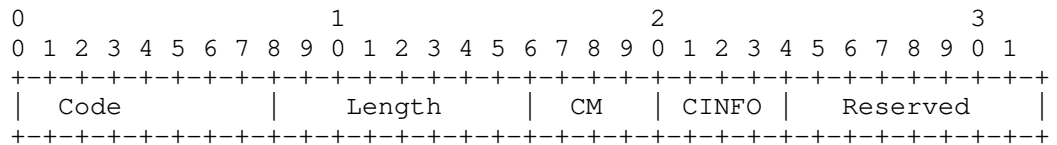
6. Packet Formats

6.1. Decompressor Capability

Decompressor Capability is following the normal procedures of [RFC5492]. In its generic form the option can support different compressors in the future.



This document specifies only DEFLATE Huffman support per [RFC1950].



Code: To be obtained by early allocation, suggested value in this process will be 76.

Length: 1 octet.

CM: 4 bits of CM indicating DEFLATE compressed format value as specified in [RFC1950].

CINFO: 4 bits of CINFO as specified in [RFC1950]. Invalid values MUST lead to the capability being ignored. The compressing peer MUST use this value for the parametrization of its algorithm.

6.2. Compressed Update Messages

This carries the original updates in a single message with content adhering to Section 4.2.

```

0          1          2          3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     |         |R|O| ULI | ID# |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| compressed data      ...
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Type: To be obtained by early allocation, suggested value in this process will be 7.

Length: 2 octets.

ID#: 3 bits. Indicates the number of the compressor used. Up to 8 compressors MAY be used by the compressing peer to allow for multiple thread of execution to compress the BGP update stream. Accordingly the decompressing side MUST support up to 8 independent decompressors.

R: If the bit is set, the according de-compressor MUST be initialized before the following compressed data is decompressed per Section 4.4. The bit MAY be set on first compressed update sent for the compressor on the session or is otherwise implied sapienti sat. The bit MUST NOT be set on the overflow fragment in case of overflow.

0: If the bit is set, procedures in Section 4.3 MUST be applied. If both the R-bit and the O-bit are set, the de-compressor must be re-initialized before the update and its overflow is assembled and decompression attempted.

ULI: Original uncompressed length indication as to be interpreted as 2**(11+ULI). This MUST indicate a buffer large enough the decompressed data (including overflow) will fit in. The indication MAY be ignored by the receiver but should allow for efficient buffer allocation. The field MUST be ignored on overflow fragment.

7. Security Considerations

This document introduces no new security concerns to BGP or other specifications referenced in this document.

8. Acknowledgements

Thanks to John Scudder for some bar discussions that primed the creative process. Thanks to Eric Rosen, Jeff Haas and Acee Lindem for their careful reviews. Thanks to David Lamperter for discussions on reordering issues.

9. Normative References

- [ID.draft-idr-bgp-route-refresh-options-03]
Patel et al., K., "Extension to BGP's Route Refresh Message", internet-draft draft-idr-bgp-route-refresh-options-03.txt, May 2017.
- [ID.draft-ietf-idr-bgp-extended-messages-21]
Bush et al., R., "Extended Message support for BGP", internet-draft draft-ietf-idr-bgp-extended-messages-21.txt, May 2016.
- [QUANT] Zyga, L., "Worldwide Quantum Web May Be Possible with Help from Graphs", New Journal on Physics , June 2016.
- [RFC1950] Deutsch, P. and J-L. Gailly, "ZLIB Compressed Data Format Specification version 3.3", RFC 1950, DOI 10.17487/RFC1950, May 1996, <<https://www.rfc-editor.org/info/rfc1950>>.
- [RFC1951] Deutsch, P., "DEFLATE Compressed Data Format Specification version 1.3", RFC 1951, DOI 10.17487/RFC1951, May 1996, <<https://www.rfc-editor.org/info/rfc1951>>.
- [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>>.
- [RFC2283] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 2283, DOI 10.17487/RFC2283, February 1998, <<https://www.rfc-editor.org/info/rfc2283>>.
- [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>>.

- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <<https://www.rfc-editor.org/info/rfc4724>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<https://www.rfc-editor.org/info/rfc5492>>.
- [RFC7313] Patel, K., Chen, E., and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", RFC 7313, DOI 10.17487/RFC7313, July 2014, <<https://www.rfc-editor.org/info/rfc7313>>.
- [RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <<https://www.rfc-editor.org/info/rfc7911>>.

Authors' Addresses

Tony Przygienda
Juniper
1137 Innovation Way
Sunnyvale, CA
USA

Email: prz@juniper.net

Avinash Lingala
AT&T
200 S Laurel Ave
Middletown, NJ
USA

Email: ar977m@att.com

Csaba Mate
NIIF/Hungarnet
18-22 Victor Hugo
Budapest 1132
Hungary

Email: matecs@niif.hu

Jeff Tantsura
Nuage Networks
755 Ravendale Drive
Mountain View, CA 94043
USA

Email: jefftant.ietf@gmail.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 6, 2017

J. Tantsura
Individual
U. Chunduri
Huawei Technologies
G. Mirsky
ZTE Corp.
S. Sivabalan
Cisco
June 04, 2017

Signaling Maximum SID Depth using Border Gateway Protocol Link-State
draft-tantsura-idr-bgp-ls-segment-routing-msd-05

Abstract

This document proposes a way to signal Maximum SID Depth (MSD) supported by a node at node and/or link granularity by a BGP-LS speaker. In a Segment Routing (SR) enabled network a centralized controller that programs SR tunnels needs to know the MSD supported by the head-end at node and/or link granularity to push the SID stack of an appropriate depth. MSD is relevant to the head-end of a SR tunnel or Binding-SID anchor node where Binding-SID expansions might result in creation of a new SID stack.

Status of This Memo

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

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

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

This Internet-Draft will expire on December 6, 2017.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
1.1. Conventions used in this document	3
1.1.1. Terminology	3
1.1.2. Requirements Language	3
2. Problem Statement	3
3. MSD supported by a node	4
4. MSD supported on a link	4
5. IANA Considerations	5
6. Security Considerations	5
7. Acknowledgements	5
8. References	5
8.1. Normative References	5
8.2. Informative References	6
Authors' Addresses	6

1. Introduction

When Segment Routing tunnels are computed by a centralized controller, it is critical that the controller learns the MSD "Maximum SID Depth" of the node or link SR tunnel exits over, so the SID stack depth of a path computed doesn't exceed the number of SIDs the node is capable of imposing. This document describes how to use BGP-LS to signal the MSD of a node or link to a centralized controller.

PCEP SR extensions draft [I-D.ietf-pce-segment-routing] signals MSD in SR PCE Capability TLV and METRIC Object. However, if PCEP is not supported/configured on the head-end of a SR tunnel or a Binding-SID anchor node and controller does not participate in IGP routing, it has no way to learn the MSD of nodes and links which has been configured. BGP-LS [RFC7752] defines a way to expose topology and associated attributes and capabilities of the nodes in that topology to a centralized controller.

MSD of sub-type 1, called Base MSD as defined in Section 3 is used to signal the number of SID's a node is capable of imposing, to be used

by a path computation element/controller. In case, there are additional labels (e.g. service) that are to be pushed to the stack - this would be signaled with an another MSD type (TBD), no adjustment to the Base MSD should be made. In the future, new MSD types could be defined to signal additional capabilities: entropy labels, labels that can be pushed thru recirculation, or another dataplane e.g IPv6.

1.1. Conventions used in this document

1.1.1. Terminology

BGP-LS: Distribution of Link-State and TE Information using Border Gateway Protocol

MSD: Maximum SID Depth

PCC: Path Computation Client

PCE: Path Computation Element

PCEP: Path Computation Element Protocol

SID: Segment Identifier

SR: Segment routing

1.1.2. 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 [RFC2119].

2. Problem Statement

In existing technology only PCEP has extension to signal the MSD (SR PCE Capability TLV/ METRIC Object as defined in [I-D.ietf-pce-segment-routing], If PCEP is not supported by the node (head-end of the SR tunnel) controller has no way to learn the MSD of the node/link configured. OSPF and IS-IS extensions are defined in:

[I-D.ietf-ospf-segment-routing-msd]

[I-D.ietf-isis-segment-routing-msd]

3. MSD supported by a node

Node MSD is encoded in a new Node Attribute TLV, as defined in [RFC7752]

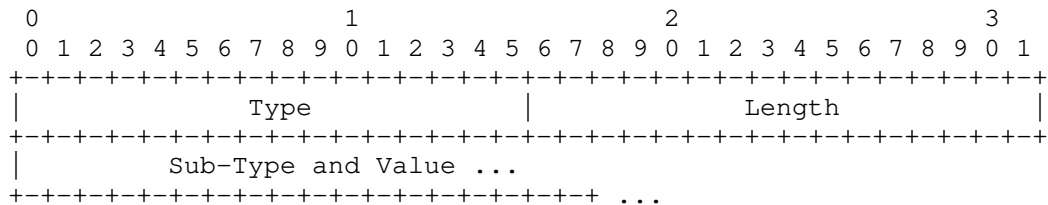


Figure 1: Node attribute format

Type : A 2-octet field specifying code-point of the new TLV type.
 Code-point: 1050 (Suggested value - to be assigned by IANA) from BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs registry

Length: A 2-octet field that indicates the length of the value portion

Sub-Type and value fields are as defined in corresponding OSPF [I-D.ietf-ospf-segment-routing-msd] and IS-IS [I-D.ietf-isis-segment-routing-msd] extensions.

4. MSD supported on a link

Link MSD is encoded in a New Link Attribute TLV, as defined in [RFC7752]

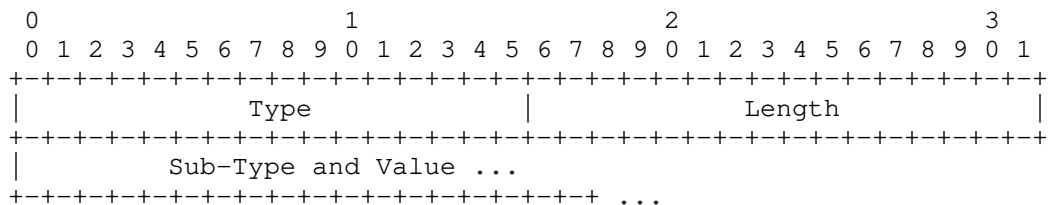


Figure 2: Link attribute format

Type : A 2-octet field specifying code-point of the new TLV type.
 Code-point: 1110 (Suggested value - to be assigned by IANA) from BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs registry

Length: A 2-octet field that indicates the length of the value portion

Sub-Type and value fields are as defined in corresponding OSPF [I-D.ietf-ospf-segment-routing-msd] and IS-IS [I-D.ietf-isis-segment-routing-msd] extensions.

5. IANA Considerations

This document requests IANA to assign 2 new code-points from the BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs registry as specified in sections 3 and 4.

6. Security Considerations

This document does not introduce security issues beyond those discussed in [RFC7752]

7. Acknowledgements

We like to thank Nikos Triantafyllis, Stephane Litkowski and Bruno Decraene for their reviews and valuable comments.

8. References

8.1. Normative References

- [I-D.ietf-isis-segment-routing-msd]
Tantsura, J., Chunduri, U., Aldrin, S., and L. Ginsberg,
"Signaling MSD (Maximum SID Depth) using IS-IS", draft-
ietf-isis-segment-routing-msd-03 (work in progress), March
2017.
- [I-D.ietf-ospf-segment-routing-msd]
Tantsura, J., Chunduri, U., Aldrin, S., and P. Psenak,
"Signaling MSD (Maximum SID Depth) using OSPF", draft-
ietf-ospf-segment-routing-msd-04 (work in progress), March
2017.
- [I-D.ietf-pce-segment-routing]
Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W.,
and J. Hardwick, "PCEP Extensions for Segment Routing",
draft-ietf-pce-segment-routing-09 (work in progress),
April 2017.

- [I-D.ietf-spring-segment-routing-mpls]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B.,
Litkowski, S., and R. Shakir, "Segment Routing with MPLS
data plane", draft-ietf-spring-segment-routing-mpls-08
(work in progress), March 2017.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119,
DOI 10.17487/RFC2119, March 1997,
<<http://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,
<<http://www.rfc-editor.org/info/rfc7752>>.

8.2. Informative References

- [I-D.ietf-isis-segment-routing-extensions]
Previdi, S., Filsfils, C., Bashandy, A., Gredler, H.,
Litkowski, S., Decraene, B., and j. jeffrant@gmail.com,
"IS-IS Extensions for Segment Routing", draft-ietf-isis-
segment-routing-extensions-12 (work in progress), April
2017.
- [I-D.ietf-ospf-segment-routing-extensions]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H.,
Shakir, R., Henderickx, W., and J. Tantsura, "OSPF
Extensions for Segment Routing", draft-ietf-ospf-segment-
routing-extensions-16 (work in progress), May 2017.

Authors' Addresses

Jeff Tantsura
Individual

Email: jeffrant.ietf@gmail.com

Uma Chunduri
Huawei Technologies

Email: uma.chunduri@huawei.com

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Siva Sivabalan
Cisco

Email: msiva@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 13, 2017

X. Xu
K. Bi
Huawei
J. Tantsura
Individual
March 12, 2017

BGP Neighbor Autodiscovery
draft-xu-idr-neighbor-autodiscovery-01

Abstract

BGP has been used as the routing protocol in many hyper-scale data centers. This document proposes a BGP neighbor autodiscovery mechanism which can be used to simplify the BGP deployment greatly. This mechanism is very useful for those hyper-scale data centers where BGP is used as the routing 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 <http://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on September 13, 2017.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	2
1.1. Requirements Language	3
2. Terminology	3
3. BGP Hello Message Format	3
4. Hello Message Procedure	5
5. HELLO Message Error Handling	6
6. Acknowledgements	6
7. IANA Considerations	6
7.1. BGP Hello Message	6
7.2. TLVs of BGP Hello Message	7
8. Security Considerations	7
9. References	7
9.1. Normative References	7
9.2. Informative References	8
Authors' Addresses	8

1. Introduction

BGP has been used as the routing protocol instead of IGP in many hyper-scale data centers [RFC7938]. Furthermore, there is an attempt to leverages BGP Link-State distribution and the Shortest Path First algorithm similar to Internal Gateway Protocols (IGPs) such as OSPF [I-D.keyupate-idr-bgp-spf]. In a word, there is a strong motivation to replace IGP by BGP in hyper-scale data centers.

However, BGP is not good as IGP from the perspective of deployment automation and simplicity. For instance, the IP address and Autonomous System Number (ASN) of each BGP neighbor have to be manually configured on BGP routers although these BGP peers are directly connected. In addition, for those directly connected BGP routers, it's usually not ideal to establish BGP sessions over their directly connected interface addresses due to the following reasons: 1) it's not convient to do trouble-shooting; 2) the BGP update volume is unnecessarily increased when there are multiple physical links between them and those links couldn't be configured as a Link Aggregation Group (LAG) due to whatever reason (e.g., diffferent link type or speed). As a result, it's more common that loopback interface addresses of those directly connected BGP peers are used for BGP session establishment. To make those loopback addresses of directly connected BGP peers reachable from one another, either static routes have to be configured or some kind of IGP has to be enabled. The former is not good from the automation perspective

while the latter is in conflict with the original intention of using BGP as IGP.

This draft specifies a BGP neighbor autodiscovery mechanism by borrowing some ideas from the Label Distribution Protocol (LDP) [RFC5036]. More specifically, directly connected BGP routers could automatically discover the loopback address and the ASN of one other through the exchange of the to-be-defined BGP HELLO messages. The BGP session establishment process as defined in [RFC4271] is triggered once directly connected BGP neighbors are discovered from one another. Note that the BGP session should be established over the discovered loopback address of the BGP neighbor. In addition, to eliminate the need of configuring static routes or enabling IGP for the loopback addresses, a certain type of routes towards the BGP neighbor's loopback addresses are dynamically created once the BGP neighbor has been discovered. The administrative distance of such type of routes MUST be smaller than their equivalents which are learnt via the normal BGP update messages. Otherwise, circular dependency problem would occur once these loopback addresses are advertised via the normal BGP update messages as well.

1.1. Requirements Language

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

2. Terminology

This memo makes use of the terms defined in [RFC4271].

3. BGP Hello Message Format

To automatically discover directly connected BGP neighbors, a BGP router periodically sends BGP HELLO messages out those interfaces on which BGP neighbor autodiscovery are enabled. The BGP HELLO message is a new BGP message which has the same fixed-size BGP header as the existing BGP messages. However, the HELLO message MUST be sent as UDP packets addressed to the to-be-assigned BGP discovery port (179 is the suggested port value) for the "all routers on this subnet" group multicast address (i.e., 224.0.0.2 in the IPv4 case and FF02::2 in the IPv6 case). The IP source address is set to the address of the interface over which the message is sent out.

In addition to the fixed-size BGP header, the HELLO message contains the following fields:

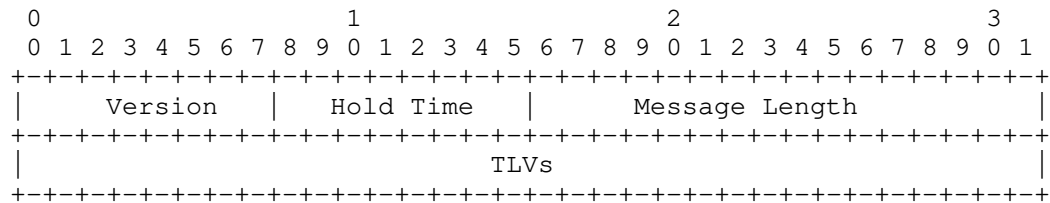


Figure 1: BGP Hello Message

Version: This 1-octet unsigned integer indicates the protocol version number of the message. The current BGP version number is 4.

Hold Time: Hello hold timer in seconds. Hello Hold Time specifies the time the sending BGP peer will maintain its record of Hellos from the receiving BGP peer without receipt of another Hello. A pair of BGP peers negotiates the hold times they use for Hellos from each other. Each proposes a hold time. The hold time used is the minimum of the hold times proposed in their Hellos. A value of 0 means use the default 15 seconds.

Message Length: This 2-octet unsigned integer specifies the length in octets of the ASN TLV, Connection Address TLV and other TLVs.

TLVs: This field contains ASN TLV, Connection Address TLV and other TLVs.

The ASN TLV format is show as follows:

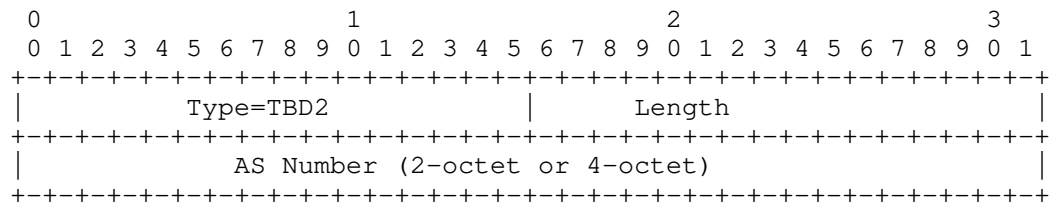


Figure 2: ASN TLV

Type: TBD2.

Length: Specifies the length of the Value field in octets.

AS Number: This variable-length field indicates the 2-octet or 4-octet ASN of the sender.

The Connection Address TLV format is shown as follows:

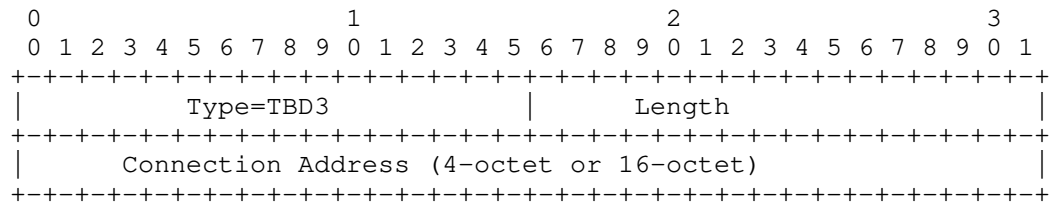


Figure 3: Connection Address TLV

Type: TBD3

Length: Specifies the length of the Value field in octets.

Connection Address: This variable-length field indicates the IPv4 or IPv6 loopback address which is used for establishing BGP sessions.

The Router ID TLV format is shown as follows:

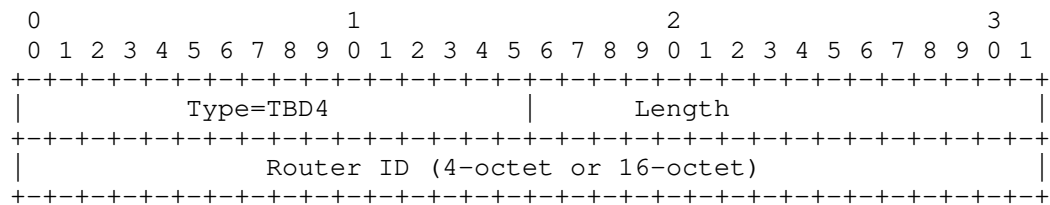


Figure 4: Router ID TLV

Type: TBD3

Length: Specifies the length of the Value field in octets and it's set to 4 for the IPv4-address-formatted BGP Router ID.

Router ID: This variable-length field indicates the BGP router ID which is used for performing the BGP-SPF algorithm as described in [I-D.keyupate-idr-bgp-spf].

4. Hello Message Procedure

A BGP peer receiving Hellos from another peer maintains a Hello adjacency corresponding to the Hellos. The peer maintains a hold timer with the Hello adjacency, which it restarts whenever it receives a Hello that matches the Hello adjacency. If the hold timer for a Hello adjacency expires the peer discards the Hello adjacency.

We recommend that the interval between Hello transmissions be at most one third of the Hello hold time.

A BGP session with a peer has one or more Hello adjacencies.

A BGP session has multiple Hello adjacencies when a pair of BGP peers is connected by multiple links that have the same connection address; for example, multiple PPP links between a pair of routers. In this situation, the Hellos a BGP peer sends on each such link carry the same Connection Address. In addition, to eliminate the need of configuring static routes or enabling IGP for the loopback addresses, a certain type of routes towards the BGP neighbor's loopback addresses (e.g., carried in the Connection Address TLV) are dynamically created once the BGP neighbor has been discovered. The administrative distance of such type of routes MUST be smaller than their equivalents which are learnt via the normal BGP update messages. Otherwise, circular dependency problem would occur once these loopback addresses are advertised via the normal BGP update messages as well.

BGP uses the regular receipt of BGP Discovery Hellos to indicate a peer's intent to keep BGP session identified by the Hello. A BGP peer maintains a hold timer with each Hello adjacency that it restarts when it receives a Hello that matches the adjacency. If the timer expires without receipt of a matching Hello from the peer, BGP concludes that the peer no longer wishes to keep BGP session for that link or that the peer has failed. The BGP peer then deletes the Hello adjacency. When the last Hello adjacency for an BGP session is deleted, the BGP peer terminates the BGP session by sending a Notification message and closing the transport connection.

5. HELLO Message Error Handling

TBD

6. Acknowledgements

The authors would like to thank

7. IANA Considerations

7.1. BGP Hello Message

This document requests IANA to allocate a new UDP port for BGP Hello message.

Value	TLV Name	Reference
Service Name:	BGP-HELLO	
Transport Protocol(s):	UDP	
Assignee:	IESG <iesg@ietf.org>	
Contact:	IETF Chair <chair@ietf.org>.	
Description:	BGP Hello Message.	
Reference:	This document -- draft-xu-idr-neighbor-autodiscovery.	
Port Number:	TBD1 (179 is the suggested value) -- To be assigned by IANA.	

7.2. TLVs of BGP Hello Message

This document requests IANA to create a new registry "TLVs of BGP Hello Message" with the following registration procedure:

Registry Name: TLVs of BGP Hello Message.

Value	TLV Name	Reference
0	Reserved	This document
1	ASN	This document
2	Connection Address	This document
3	Router ID	This document
4-65500	Unassigned	
65501-65534	Experimental	This document
65535	Reserved	This document

8. Security Considerations

For security purposes, BGP speakers usually only accept TCP connection attempts to port 179 from the specified BGP peers or those within the configured address range. With the BGP auto-discovery mechanism, it's configurable to enable or disable sending/receiving BGP hello messages on the per-interface basis and BGP hello messages are only exchanged between physically connected peers that are trustworthy. Therefore, the BGP auto-discovery mechanism doesn't introduce additional security risks associated with BGP.

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, <<http://www.rfc-editor.org/info/rfc2119>>.

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

9.2. Informative References

- [I-D.keyupate-idr-bgp-spf]
Patel, K., Lindem, A., Zandi, S., and G. Velde, "Shortest Path Routing Extensions for BGP Protocol", draft-keyupate-idr-bgp-spf-02 (work in progress), December 2016.
- [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <<http://www.rfc-editor.org/info/rfc5036>>.
- [RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <<http://www.rfc-editor.org/info/rfc7938>>.

Authors' Addresses

Xiaohu Xu
Huawei

Email: xuxiaohu@huawei.com

Kunyang Bi
Huawei

Email: bikunyang@huawei.com

Jeff Tantsura
Individual

Email: jefftant@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 14, 2017

A. Azimov
E. Bogomazov
Qrator Labs
R. Bush
Internet Initiative Japan
K. Patel
Arccus, Inc.
K. Sriram
US NIST
March 13, 2017

Route Leak Prevention using Roles in Update and Open messages
draft-ymbk-idr-bgp-open-policy-03

Abstract

Route Leaks are the propagation of BGP prefixes which violate assumptions of BGP topology relationships; e.g. passing a route learned from one peer to another peer or to a transit provider, passing a route learned from one transit provider to another transit provider or to a peer. Today, approaches to leak prevention rely on marking routes according to operator configuration options without any check that the configuration corresponds to that of the BGP neighbor, or enforcement that the two BGP speakers agree on the relationship. This document enhances BGP Open to establish agreement of the (peer, customer, provider, internal) relationship of two neighboring BGP speakers to enforce appropriate configuration on both sides. Propagated routes are then marked with an iOTC attribute according to agreed relationship allowing prevention of route leaks.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in RFC 2119 [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning.

Status of This Memo

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

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

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

This Internet-Draft will expire on September 14, 2017.

Copyright Notice

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

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

Table of Contents

1. Preamble	3
1.1. Peering Relationships	3
2. Introduction	3
3. Role Definitions	3
4. BGP Role	4
5. Role capability	5
6. Role correctness	5
6.1. Strict mode	6
7. Restrictions on the Complex role	6
8. BGP Internal Only To Customer attribute	6
9. Compatibility with BGPsec	7
10. Additional Considerations	7
11. IANA Considerations	7
12. Security Considerations	8
13. Acknowledgments	8
14. References	8
14.1. Normative References	8
14.2. Informative References	9
Authors' Addresses	9

1. Preamble

1.1. Peering Relationships

Despite uses of words such as "Customer," "Peer." etc. the intent is not business relationships, who pays whom, etc. These are common terms to represent restrictions on BGP propagation, some times known as Gao/Rexford. E.g. if A is a "peer" of B and C, A does not propagate B's prefixes to C. If D is a "customer" of E and F, D does not propagate prefixes learned from E to F.

As the whole point of route leak detection and prevention is to prevent vioation of these relationships, they are inescapable.

2. Introduction

This document specifies a new BGP Capability Code, [RFC5492] Sec 4, which two BGP speakers MAY use to ensure that they MUST agree on their relationship; i.e. customer and provider or peers. Either or both may optionally be configured to require that this option be exchanged for the BGP Open to succeed.

Also this document specifies a way to mark routes according to BGP Roles established in OPEN and a way to create double-boundary filters for prevention of route leaks via new BGP Path Attribute.

For the purpose of this document, BGP route leaks are when a BGP route was learned from transit provider or peer and is announced to another provider or peer. See [I-D.ietf-grow-route-leak-problem-definition]. These are usually the result of misconfigured or absent BGP route filtering or lack of coordination between two BGP speakers.

[I-D.ietf-idr-route-leak-detection-mitigation] The mechanism proposed in that draft provides the opportunity to detect route leaks made by third parties but provides no support to strongly prevent route leak creation.

Also, route tagging which relies on operator maintained policy configuration is too easily and too often misconfigured.

3. Role Definitions

As many of these terms are used differently in various contexts, it is worth being explicit.

A Provider: sends their own routes and (possibly) a subset of routes learned from their other customers, peers, and transit providers to their customer.

A Customer: accepts 'transit routes' from its provider(s) and announces their own routes and the routes they have learned from the transitive closure of their customers (AKA their 'customer cone') to their provider(s).

A Peer: announces their routes and the routes from their customer cone to other Peers.

An Internal: announces all routes, accepts all routes.

A Complex: BGP relationship is an attempt to allow those whose policy may vary by prefix. It is aptly named and the authors question its real utility.

Of course, any BGP speaker may apply policy to reduce what is announced, and a recipient may apply policy to reduce the set of routes they accept.

4. BGP Role

BGP Role is new mandatory configuration option. It reflects the real-world agreement between two BGP speakers about their peering relationship.

Allowed Role values are:

- o Provider - sender is a transit provider to neighbor;
- o Customer - sender is customer of neighbor;
- o Peer - sender and neighbor are peers;
- o Internal - sender and neighbor is part of same organization. This includes but is not limited to situation when sender and neighbor are in same AS.
- o Complex - sender has a non-standard relationship and wants to use manual per-prefix based role policies.

Since BGP Role reflects the relationship between two BGP speakers, it could also be used for more than route leak mitigation.

5. Role capability

The TLV (type, length, value) of the BGP Role capability are:

- o Type - <TBD1>;
- o Length - 1 (octet);
- o Value - integer corresponding to speaker' BGP Role.

Value	Role name
0	Undefined
1	Sender is Peer
2	Sender is Provider
3	Sender is Customer
4	Sender is Internal
5	Sender is Complex

Table 1: Predefined BGP Role Values

6. Role correctness

Section 4 described how BGP Role is a reflection of the relationship between two BGP speakers. But the mere presence of BGP Role doesn't automatically guarantee role agreement between two BGP peers.

To enforce correctness, the BGP Role check is used with a set of constrains on how speakers' BGP Roles MUST corresponded. Of course, each speaker MUST announce and accept the BGP Role capability in the BGP OPEN message exchange.

If a speaker receives a BGP Role capability, it SHOULD check value of the received capability with its own BGP Role. The allowed pairings are (first a sender's Role, second the receiver's Role):

Sender Role	Receiver Role
Peer	Peer
Provider	Customer
Customer	Provider
Internal	Internal
Complex	Complex

Table 2: Allowed Role Capabilities

In all other cases speaker MUST send a Role Mismatch Notification (code 2, sub-code <TBD2>).

6.1. Strict mode

A new BGP configuration option "strict mode" is defined with values of true or false. If set to true, then the speaker MUST refuse to establish a BGP session with peers which do not announce the BGP Role capability in their OPEN message. If a speaker rejects a connection, it MUST send a Connection Rejected Notification [RFC4486] (Notification with error code 6, subcode 5). By default strict mode SHOULD be set to false for backward compatibility with BGP speakers, that do not yet support this mechanism.

7. Restrictions on the Complex role

The Complex role should be set only if the relationship between BGP neighbors can not be described using simple Customer/Provider/Peer roles. For a example, if neighbor is literal peer, but for some prefixes it provides full transit; the complex role SHOULD be set on both sides. In this case roles Customer/Provider/Peer should be set on per-prefix basis, keeping the abstraction from filtering mechanisms (Section 8).

If role is not Complex all per-prefix role settings MUST be ignored.

8. BGP Internal Only To Customer attribute

The Internal Only To Customer (iOTC) attribute is a new optional, non-transitive BGP Path attribute with the Type Code <TBD3>. This attribute has zero length as it is used only as a flag.

There are four rules for setting the iOTC attribute:

1. The iOTC attribute MUST be added to all incoming routes if the receiver's Role is Customer or Peer;

2. The iOTC attribute MUST be added to all incoming routes if the receiver's Role is Complex and the prefix Role is Customer or Peer;
3. Routes with the iOTC attribute set MUST NOT be announced by a sender whose Role is Customer or Peer;
4. Routes with the iOTC attribute set MUST NOT be announced if by a sender whose Role is Complex and the prefix Role is Customer or Peer;

These four rules provide mechanism that strongly prevents route leak creation by an AS.

9. Compatibility with BGPsec

As the iOTC field is non-transitive, it is not seen by or signed by BGPsec [I-D.ietf-sidr-bgpsec-protocol].

10. Additional Considerations

As the BGP Role reflects the relationship between neighbors, it can also have other uses. As an example, BGP Role might affect route priority, or be used to distinguish borders of a network if a network consists of multiple AS.

Though such uses may be worthwhile, they are not the goal of this document. Note that such uses would require local policy control.

This document doesn't provide any security measures to check correctness of per-prefix roles, so the Complex role should be used with great caution. It is as dangerous as current BGP peering.

11. IANA Considerations

This document defines a new Capability Codes option [to be removed upon publication: <http://www.iana.org/assignments/capability-codes/capability-codes.xhtml>] [RFC5492], named "BGP Role", assigned value <TBD1> . The length of this capability is 1.

The BGP Role capability includes a Value field, for which IANA is requested to create and maintain a new sub-registry called "BGP Role Value". Assignments consist of Value and corresponding Role name. Initially this registry is to be populated with the data in Table 1. Future assignments may be made by a standard action procedure [RFC5226].

This document defines new subcode, "Role Mismatch", assigned value <TBD2> in the OPEN Message Error subcodes registry [to be removed upon publication: <http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-6>] [RFC4271].

This document defines a new optional, non-transitive BGP Path Attributes option, named "Internal Only To Customer", assigned value <TBD3> [To be removed upon publication: <http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-2>] [RFC4271]. The length of this attribute is 0.

12. Security Considerations

This document proposes a mechanism for prevention of route leaks that are the result of BGP policy misconfiguration.

Deliberate sending of a known conflicting BGP Role could be used to sabotage a BGP connection. This is easily detectable.

BGP Role is disclosed only to an immediate BGP neighbor, so it will not itself reveal any sensitive information to third parties.

13. Acknowledgments

The authors wish to thank Douglas Montgomery, Brian Dickson, and Andrei Robachevsky for their contributions to a variant of this work.

14. References

14.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<http://www.rfc-editor.org/info/rfc4271>>.
- [RFC4486] Chen, E. and V. Gillet, "Subcodes for BGP Cease Notification Message", RFC 4486, DOI 10.17487/RFC4486, April 2006, <<http://www.rfc-editor.org/info/rfc4486>>.

- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.

14.2. Informative References

- [I-D.ietf-grow-route-leak-problem-definition]
Sriram, K., Montgomery, D., McPherson, D., Osterweil, E., and B. Dickson, "Problem Definition and Classification of BGP Route Leaks", draft-ietf-grow-route-leak-problem-definition-06 (work in progress), May 2016.
- [I-D.ietf-idr-route-leak-detection-mitigation]
Sriram, K., Montgomery, D., Dickson, B., Patel, K., and A. Robachevsky, "Methods for Detection and Mitigation of BGP Route Leaks", draft-ietf-idr-route-leak-detection-mitigation-03 (work in progress), May 2016.
- [I-D.ietf-sidr-bgpsec-protocol]
Lepinski, M. and K. Sriram, "BGPsec Protocol Specification", draft-ietf-sidr-bgpsec-protocol-15 (work in progress), March 2016.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <<http://www.rfc-editor.org/info/rfc5226>>.

Authors' Addresses

Alexander Azimov
Qrator Labs

Email: aa@qrator.net

Eugene Bogomazov
Qrator Labs

Email: eb@qrator.net

Randy Bush
Internet Initiative Japan

Email: randy@psg.com

Keyur Patel
Arrcus, Inc.

Email: keyur@arrcus.com

Kotikalapudi Sriram
US NIST

Email: ksriram@nist.gov

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 19, 2018

A. Azimov
E. Bogomazov
Qrator Labs
R. Bush
Internet Initiative Japan
K. Patel
Arccus, Inc.
K. Sriram
US NIST
November 15, 2017

New definition of ISP internal eBGP border using BGP Roles
draft-ymbk-idr-isp-border-02

Abstract

This document proposes a new definition of ISP borders using BGP Roles. It may be used to improve the BGP best path selection algorithm for better support of hot-potato routing between different internal ASNs of an ISP. It may also be used to enable transmission of local attributes between different internal ASNs of an ISP.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in RFC 2119 [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning.

Status of This Memo

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

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

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

This Internet-Draft will expire on May 19, 2018.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
2. Changes in BGP decision process	3
3. Local Attributes Transmission	3
4. IANA Considerations	3
5. Normative References	3
Authors' Addresses	4

1. Introduction

The BGP best path selection algorithm (Section 9.1.2.2 of [RFC4271]) has a very clear definition of a network border: different ASNs - different networks. It differs from some real world situations when two networks become one business entity and want to operate as one network.

Today BGP does not provide any robust or automated support for such merging networks:

- o There is no support for carrying local attributes through this border,
- o Hot-potato routing, implemented by eBGP being preferred to iBGP, does not work, and
- o Route Leak prevention inside such a united network can not be easily automated.

In [I-D.ietf-idr-bgp-open-policy] BGP Roles were introduced - a configuration option that strongly enforces agreement on real-world peering relations between two BGP speakers. This configuration option can accept values of: Peering, Customer, Provider and

Internal. These values could be used in a new ISP border definition: Internal vs. External. With this definition of network borders, it becomes easy to allow robust propagation of local attributes between different ASNs of one ISP. It could be also used to improve the hot-potato routing mechanism: where routes learned from External BGP connections should be preferred over Internal, even those which cross the ISP's internal AS/AS boundary.

2. Changes in BGP decision process

To improve hot-potato routing for networks with multiple ASNs we propose to insert before d) in Section 9.1.2.2 of [RFC4271] next step:

If at least one of the candidate routes was received via a BGP session with External (Peer, Provider, Customer) role, remove from consideration all routes that were received via BGP sessions with an Internal role.

While this step will improve traffic control for ISPs with multiple ASNs it will have no affect on ISPs with single ASN.

3. Local Attributes Transmission

Propagation of local attributes through an ISP's internal AS/AS border could be enabled only if both sides set Internal roles in their BGP Open negotiation. Different attributes may still have different transmission policy:

- o iOTC attribute from [I-D.ietf-idr-bgp-open-policy] MUST be sent to enforce route leak prevention,
- o LOCAL_PREF attribute MAY be sent, and
- o MED attribute MAY be sent without changes.

4. IANA Considerations

This document has no IANA actions.

5. Normative References

- [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-01 (work in progress), July 2017.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.

Authors' Addresses

Alexander Azimov
Qrator Labs

Email: aa@qrator.net

Eugene Bogomazov
Qrator Labs

Email: eb@qrator.net

Randy Bush
Internet Initiative Japan

Email: randy@psg.com

Keyur Patel
Arrcus, Inc.

Email: keyur@arrcus.com

Kotikalapudi Sriram
US NIST

Email: ksriram@nist.gov