

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 22 June 2022

H. Chen
M. McBride
Futurewei
R. Chen
ZTE Corporation
G. Mishra
Verizon Inc.
A. Wang
China Telecom
Y. Liu
China Mobile
Y. Fan
Casa Systems
B. Khasanov
Yandex LLC
L. Liu
Fujitsu
X. Liu
Volta Networks
19 December 2021

BGP for BIER-TE Path
draft-chen-idr-bier-te-path-03

Abstract

This document describes extensions to Border Gateway Protocol (BGP) for distributing a Bit Index Explicit Replication Traffic/Tree Engineering (BIER-TE) path. A new Tunnel Type for BIER-TE path is defined to encode the information about a BIER-TE path.

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] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

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

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

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

This Internet-Draft will expire on 22 June 2022.

Copyright Notice

Copyright (c) 2021 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. Terminologies	3
2. Overview of BGP for BIER-TE Path	4
2.1. Example BIER-TE Topology with BGP	4
2.2. Distributing Path to Ingress	5
3. Extensions to BGP	6
3.1. New SAFI and NLRI	6
3.2. New Tunnel Type for BIER-TE	7
3.3. Path BitPositions Sub-TLV	8
3.4. Path Name Sub-TLV	9
3.5. Traffic Description Sub-TLVs	10
4. Security Considerations	11
5. Acknowledgements	11
6. IANA Considerations	11
6.1. Existing Registry: SAFI Parameters	11
6.2. Existing Registry: BGP TEA Tunnel Types	12
6.3. Existing Registry: BGP TEA sub-TLVs	12
7. References	12
7.1. Normative References	12
7.2. Informative References	13
Appendix A. Extensions to PMSI_TUNNEL Attribute	13
A.1. New Tunnel Type for BIER-TE	14
Authors' Addresses	14

1. Introduction

[I-D.ietf-bier-te-arch] introduces Bit Index Explicit Replication (BIER) Tree Engineering (BIER-TE). It is an architecture for per-packet stateless explicit point to multipoint (P2MP) multicast path/tree, which is called BIER-TE path, and based on the BIER architecture defined in [RFC8279].

A Bit-Forwarding Router (BFR) in a BIER-TE domain has a BIER-TE Bit Index Forwarding Table (BIFT). A BIER-TE BIFT on a BFR comprises a forwarding entry for a BitPosition (BP) assigned to each of the adjacencies of the BFR. If the BP represents a forward connected adjacency, the forwarding entry for the BP forwards the multicast packet with the BP to the directly connected BFR neighbor of the adjacency. If the BP represents a BFER (i.e., egress node) or say a local decap adjacency, the forwarding entry for the BP decapsulates the multicast packet with the BP and passes a copy of the payload of the packet to the packet's NextProto within the BFR.

A Bit-Forwarding Ingress Router (BFIR) in a BIER-TE domain receives the information or instructions about which multicast flows/packets are mapped to which BIER-TE paths that are represented by BitPositions or say BitStrings. After receiving the information or instructions, the ingress node/router encapsulates the multicast packets with the BitPositions for the corresponding BIER-TE paths, replicates and forwards the packets with the BitPositions along the BIER-TE paths.

This document proposes some procedures and extensions to BGP for distributing a BIER-TE path to the Bit-Forwarding Ingress Router (BFIR) of the path. It specifies a way of encoding the information about a BIER-TE path in a BGP UPDATE message, which can be distributed to the BFIR of the path.

1.1. Terminologies

The following terminologies are used in this document.

BIER: Bit Index Explicit Replication.

BIER-TE: BIER Tree Engineering.

BFR: Bit-Forwarding Router.

BFIR: Bit-Forwarding Ingress Router.

BFER: Bit-Forwarding Egress Router.

BFR-id: BFR Identifier. It is a number in the range [1,65535].

BFR-NBR: BFR Neighbor.

BFR-prefix: An IP address (either IPv4 or IPv6) of a BFR.

BIRT: Bit Index Routing Table. It is a table that maps from the BFR-id (in a particular sub-domain) of a BFER to the BFR-prefix of that BFER, and to the BFR-NBR on the path to that BFER.

BIFT: Bit Index Forwarding Table.

P-tunnel: A multicast tunnel through the network of one or more SPs.

PMSI: Provider Multicast Service Interface. PMSI is an abstraction that represents a multicast service for carrying packets. A PMSI is instantiated via one or more P-tunnels.

I-PMSI A-D Route: Inclusive PMSI Auto-Discovery route.

S-PMSI A-D route: Selective PMSI Auto-Discovery route.

x-PMSI A-D route: A route that is either an I-PMSI A-D route or an S-PMSI A-D route.

2. Overview of BGP for BIER-TE Path

This section briefs the BGP for BIER-TE path and illustrates some details through a simple example BIER-TE topology.

2.1. Example BIER-TE Topology with BGP

An example BIER-TE domain topology using SDN controller with a BGP to distribute BIER-TE path is shown in Figure 1. There are 8 nodes/BFRs A, B, C, D, E, F, G and H in the domain. Nodes/BFRs A, H, E, F and D are BFIRs (i.e., ingress nodes) or BFERs (i.e., egress nodes). The controller has a BGP session with each of the edge nodes in the domain, including BFIRs (i.e., ingress nodes A, H, E, F and D), and each of the non edge nodes in the domain (i.e., nodes B, C and G). Note that some of connections and the BGP on each edge node are not shown in the figure.

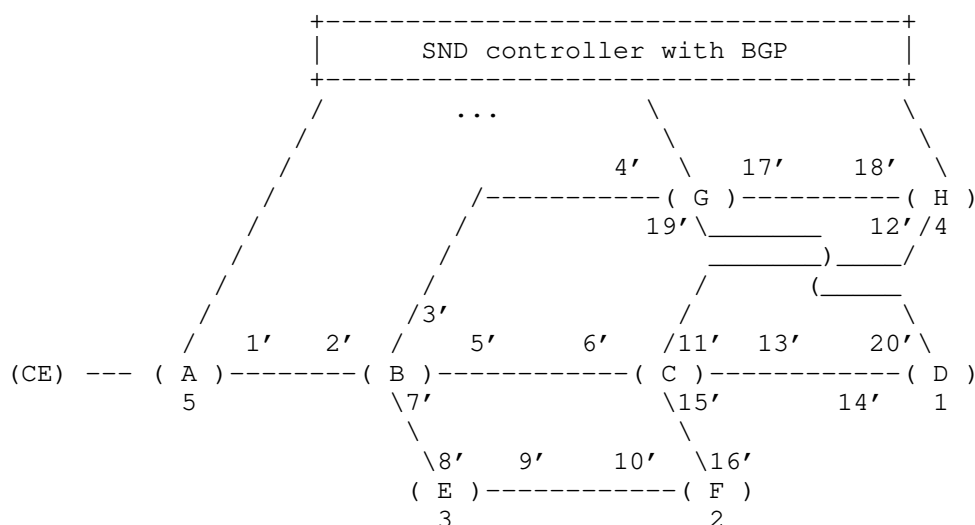


Figure 1: Example BIER-TE Topology with Controller

Nodes/BFRs D, F, E, H and A are BFERs (or BFIRs) and have local decap adjacency BitPositions 1, 2, 3, 4, and 5 respectively.

The BitPositions for the forward connected adjacencies are represented by i' , where i is from 1 to 20.

2.2. Distributing Path to Ingress

This section describes how the SDN controller distributes a BIER-TE path to its ingress node.

There are two scenarios for distributing the BIER-TE path information. In the first scenario, the ingress node is directly connected to the controller. The path information should not be propagated beyond the ingress node. In the second scenario, the ingress node is not directly connected to the controller. The path information should be propagated throughout the domain until it reaches the ingress node.

Suppose that node A in Figure 1 wants to have a BIER-TE path from ingress node A to egress nodes H and F. The path satisfies a set of constraints. The controller obtains the request from an application or user configuration. It finds a BIER-TE path satisfying the constraints and distributes the path to ingress node A.

If A is directly connected to the controller (e.g., as the example network in Figure 1), then the controller sends A the information about the path in a Update message in one of two ways. In one way, the controller sends each of its BGP peers, including the BGP peer running on node A, a Update message about the explicit path, with a route target matching the BGP identifier of ingress node A, and NO_ADVERTISE community. Ingress node A accepts this message from the controller and installs a forwarding entry for the BIER-TE path, but will not advertise it to any peer. All the other peers do not accept the message.

In another way, the controller sends A a Update message directly through the local session between them, but does not send the message to any other peers. The message contains the information about the path, a route target matching the BGP identifier of ingress node A and the NO_ADVERTISE community. Ingress node A accepts this message from the controller and installs a forwarding entry for the BIER-TE path, but will not advertise it.

If A is not directly connected to the controller, then the controller distributes the information about the explicit path to the ingress node A across the network. To achieve this, the controller advertises a BGP Update message to all its BGP peers, where the message contains the information about the path, a route target matching the BGP identifier of ingress node A, but does not have NO_ADVERTISE community. Each of the BGP peers advertises the received Update to its BGP neighbors according to normal BGP propagation rules. Eventually, ingress node A accepts this message and installs a forwarding entry for the BIER-TE path, which imports the packets to be transported by the path into the path.

3. Extensions to BGP

This section defines a new Tunnel Type (or say TLV) for BIER-TE path/tunnel under Tunnel Encapsulation Attribute and a new SAFI. This new SAFI and the existing AFI for IPv4/IPv6 pair uses a new NLRI for indicating a BIER-TE Path.

3.1. New SAFI and NLRI

A new SAFI, called BIER-TE path SAFI, is defined. Its codepoint (TBD1) is to be assigned by IANA. This new SAFI and the existing AFI for IPv4/IPv6 pair uses a new NLRI, which is defined as follows:

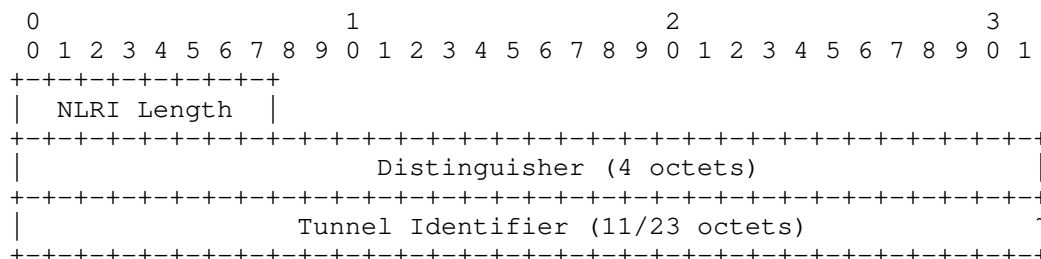


Figure 2: NLRI Format

Where:

NLRI Length: 1 octet represents the length of NLRI. If the Length is anything other than 15 or 27, the NLRI is corrupt and the enclosing UPDATE message MUST be ignored.

Distinguisher: 4 octet value uniquely identifies the content/BIER-TE path.

Tunnel Identifier: 11/23 octet value contains:

- * sub-domain-id (1 octet): It is id of the sub domain through which the BIER-TE tunnel crosses.
- * BFR-id (2 octets): It is the BFR-id of the BFIR of the BIER-TE tunnel.
- * Tunnel-ID (4 octets): It is a number uniquely identifying a BIER-TE tunnel within the BFIR and sub domain.
- * BFR-prefix (4/16 octets): It is a BFR-prefix of the BFIR of the BIER-TE tunnel. It occupies 4 octets for IPv4 and 16 octets for IPv6.

3.2. New Tunnel Type for BIER-TE

A new Tunnel Type (or say TLV), called BIER-TE Path or Tunnel, is defined under Tunnel Encapsulation Attribute in [RFC9012]. Its codepoint is to be assigned by IANA. This new TLV with a number of new sub-TLVs encodes the information about a BIER-TE path.

The structure encoding the information about a BIER-TE path is shown below.

Attributes:

```

    Tunnel Encaps Attribute (23)
      Tunnel Type (TBD2): BIER-TE Path
      Path BitPositions sub-TLV
      Path Name sub-TLV
      Traffic Description sub-TLV

```

Where:

- * Tunnel Type (TBD2) is to be assigned by IANA.
- * Path BitPositions sub-TLV encodes the bit positions of the BIER-TE path.
- * Path Name sub-TLV encodes the name of a BIER-TE path.
- * Traffic Description sub-TLV encodes the multicast traffic that is transported by the BIER-TE path.

3.3. Path BitPositions Sub-TLV

The bit positions of a BIER-TE path are encoded in a Path BitPositions sub-TLV. The format of the sub-TLV is illustrated below.

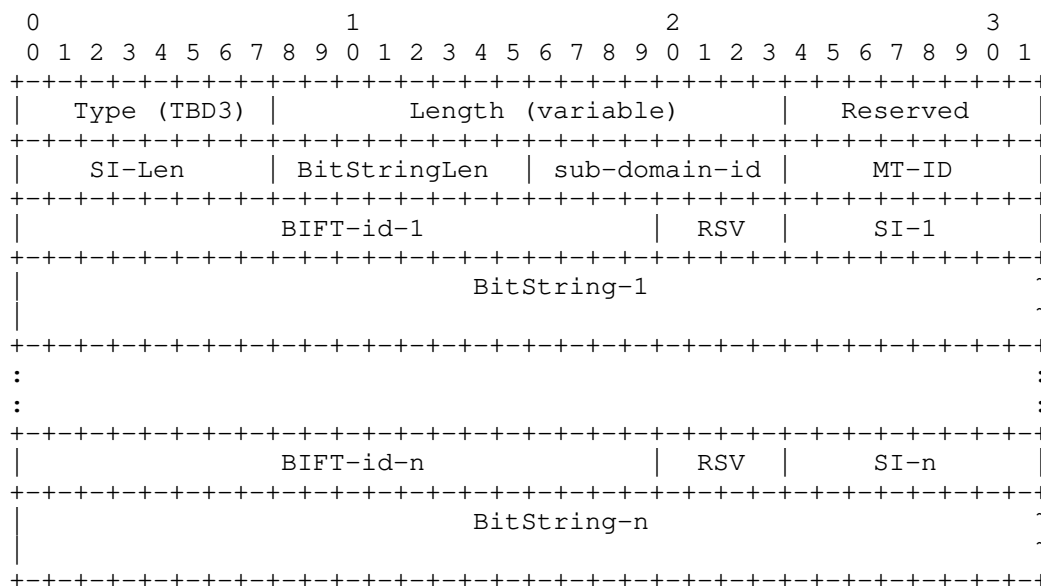


Figure 3: Path BitPositions Sub-TLV Format

Type: Its value (TBD3) is to be assigned by IANA.

Length: It is variable.

Reserved/RSV: MUST be set to zero by the sender and MUST be ignored by the receiver.

SI-Len (SI Length) - 8 bits: The length in bits of the SI field.

BitStringLength (Bit String Length) - 8 bits: The length in bits of the BitString field according to [RFC8296]. If k is the length of the BitString, the value of BitStringLength is $\log_2(k)-5$. For example, BitStringLength = 1 indicates $k = 64$, BitStringLength = 7 indicates $k = 4096$.

sub-domain-id: Unique value identifying the BIER sub-domain within the BIER domain.

MT-ID: Multi-Topology ID identifying the topology that is associated with the BIER sub-domain.

<BIFT-id, SI, BitString> tuple: Each tuple <BIFT-id- i , SI- i , BitString- i > ($i = 1, 2, \dots, n$) represents/encodes a set of bit positions on the BIER-TE path with a BIFT ID. All the tuples in the sub-TLV represent/encode the BIER-TE path (i.e., all the bit positions of the BIER-TE path).

3.4. Path Name Sub-TLV

The name of a BIER-TE path is encoded in a Path Name sub-TLV. The format of the sub-TLV is illustrated below.

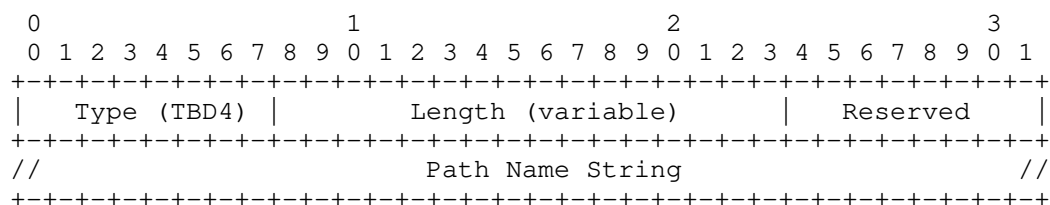


Figure 4: Path Name Sub-TLV Format

Type: Its value (TBD4) is to be assigned by IANA.

Length: It is variable.

Reserved: MUST be set to zero by the sender and MUST be ignored by

the receiver.

Path Name String: It represents/encodes the name of the BIER-TE path in a string of chars.

3.5. Traffic Description Sub-TLVs

A Traffic Description Sub-TLV describes the traffic to be imported into a BIER-TE path. Two Traffic Description Sub-TLVs are defined. They are multicast traffic sub-TLVs for IPv4 and IPv6.

The multicast traffic sub-TLVs for IPv4 and IPv6 are shown in Figure 5 and Figure 6 respectively.

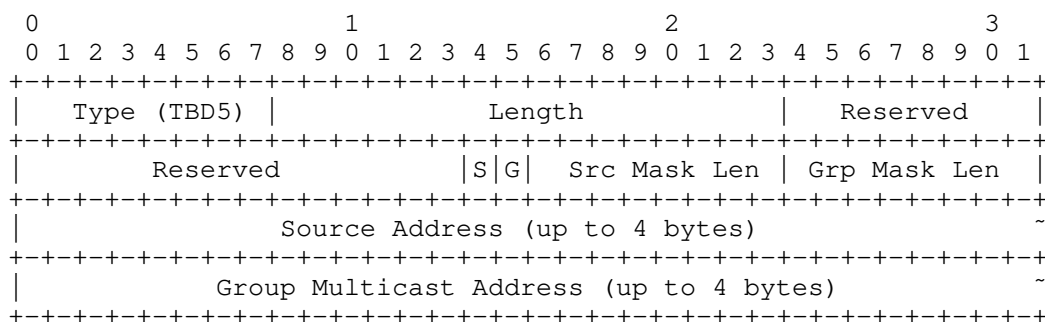


Figure 5: Multicast Traffic for IPv4 Sub-TLV

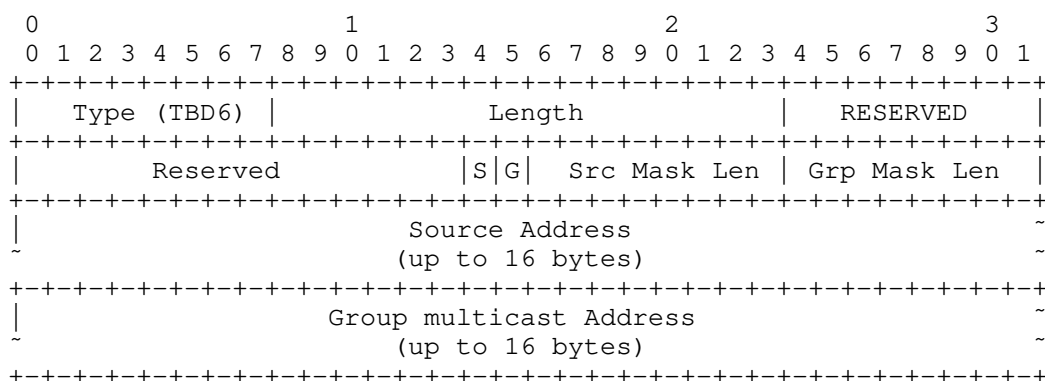


Figure 6: Multicast Traffic for IPv6 Sub-TLV

The address fields and address mask lengths of the two Multicast Traffic sub-TLVs contain source and group prefixes for matching against packets noting that the two address fields are up to 32 bits for an IPv4 Multicast Traffic and up to 128 bits for an IPv6 Multicast Traffic.

The Reserved field MUST be set to zero and ignored on receipt.

Two bit flags (S and G) are defined to describe the multicast wildcarding in use. If the S bit is set, then source wildcarding is in use and the values in the Source Mask Length and Source Address fields MUST be ignored. If the G bit is set, then group wildcarding is in use and the values in the Group Mask Length and Group multicast Address fields MUST be ignored. The G bit MUST NOT be set unless the S bit is also set: if a Multicast Traffic sub-TLV is received with S bit = 0 and G bit = 1 the receiver MUST respond with an error (Malformed Multicast Traffic).

The three multicast mappings may be achieved as follows:

(S, G): S bit = 0, G bit = 0, the Source Address and Group multicast Address prefixes are both used to define the multicast traffic.

(*, G): S bit = 1, G bit = 0, the Group multicast Address prefix is used to define the multicast traffic, but the Source Address prefix is ignored.

(*, *): S bit = 1, G bit = 1, the Source Address and Group multicast Address prefixes are both ignored.

4. Security Considerations

Protocol extensions defined in this document do not affect the BGP security other than those as discussed in the Security Considerations section of [RFC9012].

5. Acknowledgements

The authors of this document would like to thank Tony Przygienda, Susan Hares, and Jeffrey Zhang for their comments.

6. IANA Considerations

6.1. Existing Registry: SAFI Parameters

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

Code Point	Description	Reference
TBD1 (179 suggested)	BIER-TE Policy SAFI	This document

6.2. Existing Registry: BGP TEA Tunnel Types

This document requests assigning a new Tunnel-Type in the registry "BGP Tunnel Encapsulation Attribute Tunnel Types" as follows:

Code Point	Description	Reference
TBD2 (16 suggested)	BIER-TE Tunnel/Path	This document

6.3. Existing Registry: BGP TEA sub-TLVs

This document requests assigning a few of new sub-TLVs in the registry "BGP Tunnel Encapsulation Attribute sub-TLVs" as follows:

Code Point	Description	Reference
TBD3 (16 suggested)	Path BitPositions	This document
TBD4 (17 suggested)	Path Name	This document
TBD5 (18 suggested)	IPv4 Multicast Traffic	This document
TBD6 (19 suggested)	IPv6 Multicast Traffic	This document

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, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.

- [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>>.
- [RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <<https://www.rfc-editor.org/info/rfc8279>>.
- [RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <<https://www.rfc-editor.org/info/rfc8296>>.
- [RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <<https://www.rfc-editor.org/info/rfc9012>>.

7.2. Informative References

- [I-D.ietf-bier-te-arch]
Eckert, T., Cauchie, G., and M. Menth, "Tree Engineering for Bit Index Explicit Replication (BIER-TE)", Work in Progress, Internet-Draft, draft-ietf-bier-te-arch-11, 15 November 2021, <<https://www.ietf.org/archive/id/draft-ietf-bier-te-arch-11.txt>>.
- [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>>.
- [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>>.

Appendix A. Extensions to PMSI_TUNNEL Attribute

This section defines a new Tunnel Type (or TLV) for BIER-TE path/tunnel under the PMSI_TUNNEL Attribute (PTA) defined in [RFC6514]. It describes a couple of new sub-TLVs encoding the information about a BIER-TE path.

A.1. New Tunnel Type for BIER-TE

The PMSI Tunnel attribute carried by an x-PMSI A-D route identifies P-tunnel for PMSI. For the PTA with Tunnel Type BIER-TE, the PTA is constructed by the SDN controller and distributed to the ingress node of the BIER-TE tunnel.

The format of the PMSI_TUNNEL Attribute with the new Tunnel Type (TBD) for BIER-TE is shown in Figure 7.

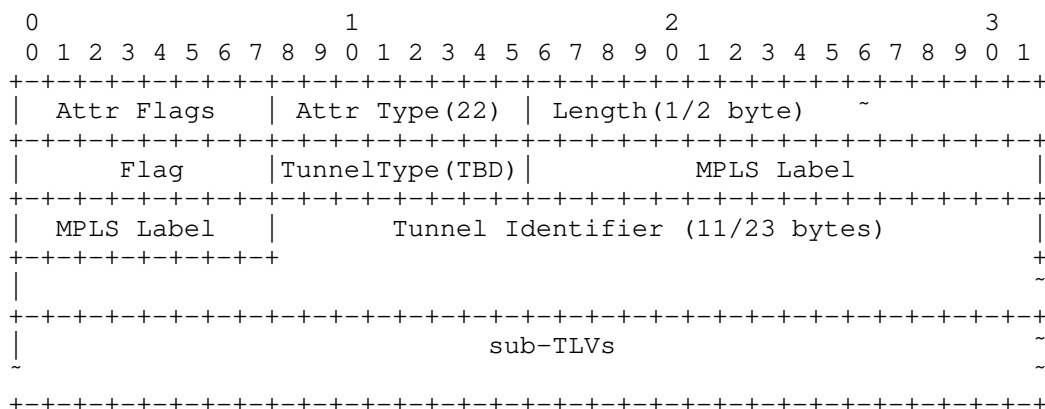


Figure 7: PTA with Tunnel Type for BIER-TE

For BIER-TE tunnel/path, the fields in the PTA are set as follows:

- o Tunnel Type: It is set to be TBD, indicating BIER-TE tunnel.
- o Tunnel Identifier: It contains: sub-domain-id of 1 byte, BIER-TE tunnel BFIR's BFR-id of 2 bytes, Tunnel-ID of 4 bytes, and BIER-TE tunnel BFIR's BFR-prefix of 4/16 bytes for IPv4/IPv6.
- o sub-TLVs: It contains a Path BitPositions sub-TLV encoding an explicit BIER-TE path. It may include a Path Name sub-TLV for the name of the BIER-TE path.
- o Others: The other fields are set according to [RFC6514].

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA,
United States of America

Email: huaimo.chen@futurewei.com

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

Ran Chen
ZTE Corporation

Email: chen.ran@zte.com.cn

Gyan S. Mishra
Verizon Inc.
13101 Columbia Pike
Silver Spring, MD 20904
United States of America

Phone: 301 502-1347
Email: gyan.s.mishra@verizon.com

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

Email: wangaj3@chinatelecom.cn

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Yanhe Fan
Casa Systems
United States of America

Email: yfan@casa-systems.com

Boris Khasanov
Yandex LLC
Moscow

Email: bhassanov@yahoo.com

Lei Liu
Fujitsu
United States of America

Email: liulei.kddi@gmail.com

Xufeng Liu
Volta Networks
McLean, VA
United States of America

Email: xufeng.liu.ietf@gmail.com

Network Working Group
Internet Draft
Intended status: Standard
Expires: August 23, 2022

L. Dunbar
Futurewei
K. Majumdar
CommScope
H. Wang
Huawei
G. Mishra
Verizon
February 23, 2022

BGP Update for 5G Edge Computing Service Metadata
draft-dunbar-idr-5g-edge-compute-app-meta-data-06

Abstract

This draft describes a new AppMetaData subTLV carried by Tunnel Encap[RFC9012] Path Attribute for egress router to advertise the running status and environment for the directly attached 5G Edge Computing (EC) servers. The AppMetaData can be used by the ingress routers in the 5G Local Data Network to make path selection not only based on the routing distance but also the running environment of the destinations. The goal is to improve latency and performance for 5G EC services.

The extension enables an EC server at one specific location to be more preferred than the others with the same IP address to receive data flows from a specific source (UE).

Status of this Memo

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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

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

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

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 7, 2021.

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 (<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. 5G Edge Computing Background.....	3
1.2. 5G Edge Computing Network Properties.....	4
1.3. Problem#1: ANYCAST in 5G EC Environment.....	5
1.4. Problem #2: Unbalanced Anycast Distribution due to UE Mobility.....	7
1.5. Problem 3: Application Server Relocation.....	8
2. Conventions used in this document.....	8
3. Usage of AppMetaData for 5G Edge Computing.....	9
3.1. Assumptions.....	9
3.2. IP Layer Metrics to Gauge Application Behavior.....	10
3.3. AppMetaData Constrained Optimal Path Selection.....	11

4. BGP Protocol Extension to advertise Load & Capacity.....	12
4.1. Ingress Node BGP Path Selection Behavior.....	12
4.1.1. AppMetaData Influenced BGP Path Selection.....	12
4.1.2. Ingress Router Forwarding Behavior.....	12
4.1.3. Forwarding Behavior when UEs moving to new 5G Sites.....	14
5. The Sub-TLVs for AppMetaData.....	14
5.1. Load Measurement sub-TLV format.....	15
5.2. Capacity Index sub-TLV format.....	16
5.3. The Site Preference Index sub-TLV format.....	16
6. AppMetaData Propagation Scope.....	17
7. Minimum Interval for Metrics Change Advertisement.....	17
8. Soft Anchoring of an ANYCAST Flow.....	17
9. Manageability Considerations.....	19
10. Security Considerations.....	19
11. IANA Considerations.....	19
12. References.....	20
12.1. Normative References.....	20
12.2. Informative References.....	20
13. Acknowledgments.....	21

1. Introduction

This document describes a new subTLV, AppMetaData, for egress routers to advertise the running status and environment for the directly attached Edge Computing (EC) servers. The AppMetaData can be used by the ingress routers in the 5G Local Data Network to make path selection not only based on the routing distance but also the running environment of the destinations. The goal is to improve latency and performance for 5G Edge Computing services.

1.1. 5G Edge Computing Background

In 5G Edge Computing (EC), one Application can be hosted on multiple Servers in different EC data centers that are close in proximity. The 5G Local Data Networks (LDN) that connect the EC data centers with the 5G Base stations consist of a small number of dedicated routers.

When a User Equipment (UE) initiates application packets using the destination address from a DNS reply or its cache, the packets from the UE are carried in a PDU session through 5G Core [5GC] to the 5G UPF-PSA (User Plan Function - PDU Session Anchor). The UPF-PSA decapsulates the 5G GTP outer header and

forwards the packets from the UEs to its directly connected Ingress router of the 5G LDN. The LDN for 5G EC is responsible for forwarding the packets to the intended destinations.

When the UE moves out of coverage of its current gNB (next-generation Node B) and anchors to a new gNB, the 5G SMF (Session Management Function) could select the same UPF or a new UPF for the UE per standard handover procedures described in 3GPP TS 23.501 and TS 23.502. If the UE is anchored to a new UPF-PSA when the handover process is complete, the packets to/from the UE is carried by a GTP tunnel to the new UPF-PSA. Per TS 23.501-h20 Section 5.8.2, the UE may maintain its IP address when anchored to a new UPF-PSA unless the new UPF-PSA belongs to different mobile operators. 5GC may maintain a path from the old UPF to the new UPF for a short time for the SSC [Session and Service Continuity] mode 3 to make the handover process more seamless.

1.2. 5G Edge Computing Network Properties

In this document, 5G Edge Computing Network refers to multiple Local IP Data Networks (LDN) in one region that interconnect the Edge Computing data centers. Those IP LDN networks are the N6 interfaces from 3GPP 5G perspective.

The ingress routers to the 5G Edge Computing Network are the routers directly connected to 5G UPFs. The egress routers to the 5G Edge Computing [EC] Network are the routers that have a direct link to the EC servers. The EC servers and the egress routers are co-located. Some of those Edge Computing Data centers may have virtual switches or Top of Rack [ToR] switches between the egress routers and the servers. But transmission delay between the egress routers and the EC servers is negligible, which is too small to be considered in this document.

When multiple EC servers are attached to one App Layer Load Balancer, only the IP addresses of the App Layer Load Balancer are visible to the 5G LDNs. How an App Layer Load balancer manages the individual servers is out of the scope of the network layer.

The 5G EC Services are registered premium services that require super-low latency and very high SLA. Most services by the UEs are not part of the registered 5G EC Services.

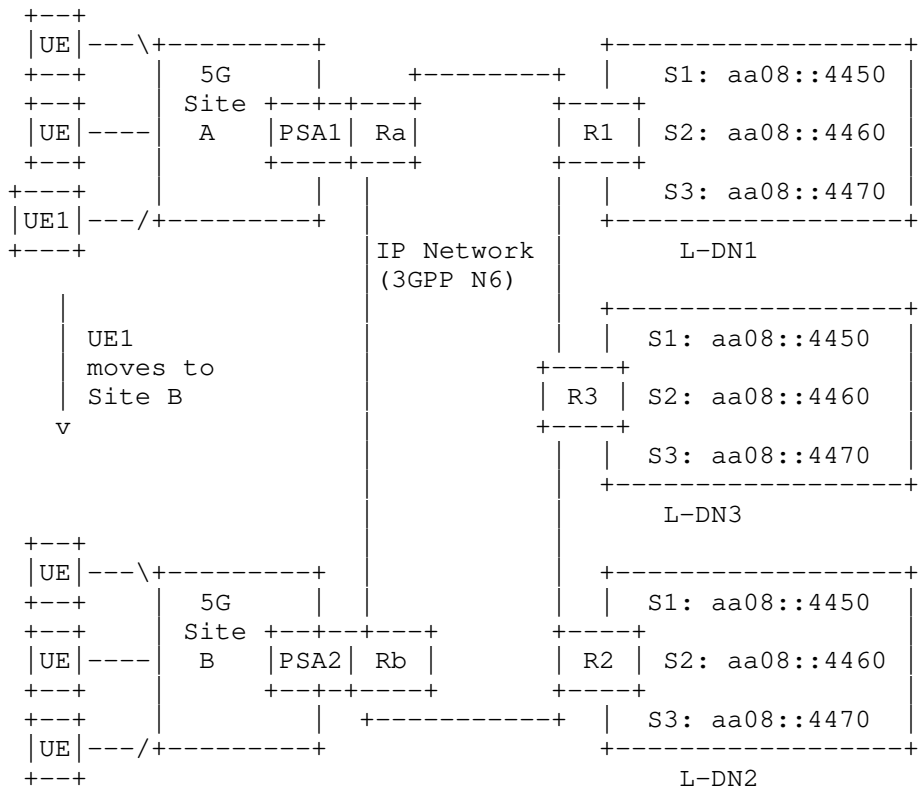


Figure 1: App Servers in different edge DCs

1.3. Problem#1: ANYCAST in 5G EC Environment

Increasingly, Anycast is used by various application providers and CDNs because Anycast provides better and faster resiliency to failover events than GEO database DNS-based load balancing, which relies on DNS to provide a different IP based on source address.

Anycast address leverages the proximity information present in the network (routing) layer. It eliminates the single point of failure and bottleneck at the DNS resolvers. Anycast address can be assigned to multiple app layer load balancers to leverage network condition for balanced forwarding. Another benefit of using the ANYCAST address is removing the dependency on UEs. Some UEs (or clients) might use their cached IP addresses for an extended period instead of querying DNS.

Client using Virtual IP address is a common practice in Cloud Native networking, e.g., Kubernetes, to scale dynamic changes of app servers' instantiations. Virtual IP requires the destination gateway node to perform address translation for return traffic, which is unsuitable for underlay network nodes with millions of flows passing by. The Cloud Native network can also leverage network condition to balance forwarding among multiple Cloud Gateway nodes by assigning the same virtual IP address (ANYCAST).

Having multiple locations of the same IP address in the 5G EC LDN can be problematic if path selection is solely based on routing cost as the routing cost differences to reach different egress routers can be very small. This list elaborates the issues in detail:

- a) Path Selection: When a new flow comes to an ingress node (Ra), how to avoid instability with Anycast flipping between paths to the same address. The problem also exists in the BGP multipath environment, with the optimal path selected based on routing cost metrics.

- b) Ingress node forwards the packets from one flow to the same ANYCAST server.

a.k.a. Flow Affinity, or Flow-based load balancing.

Almost all vendors have supported flow or session based ECMP load balancing and not per packet to avoid out of order packets

for decades. When a flow is hashed to an

ECMP path, the flow remains on that path for the life of the flow until the flow ends.

The ingress node, (Ra/Rb), can use Flow ID (in IPv6 header) or UDP/TCP port number combined with the source address to enforce packets in one flow being placed in

one tunnel to one Egress router. No new features are needed.

- c) When a UE moves to a new 5G site in the middle of a communication session with an EC server, a method is needed to stick the flow to the same EC server, which is required by 5G Edge Computing: 3GPP TR 23.748. [5g-edge-compute-sticky-service] describes several approaches to achieve stickiness in the IPv6 domain.

Note: most EC services have shorter sessions, e.g., shorter TCP sessions. Most likely, when a UE is moving to a new 5G site, the TCP session via the old UPF to an EC server is already finished. Only a very small percentage of registered EC services need to stick to the original server when handover to a new cell tower.

From BGP perspective, the multiple servers with the same IP address (ANYCAST) attached to different egress routers is the same as multiple next hops for the IP address.

This draft describes the BGP UPDATE to enable ingress routers to take the App Server load, the capacity index, and the location preference into consideration when computing the optimal path to the egress routers.

1.4. Problem #2: Unbalanced Anycast Distribution due to UE Mobility

Usually, higher capacity EC servers are placed in a metro data center to accommodate more UEs in the proximity needing the services, and fewer are placed in remote sites. When there is a special event occurring at a remote site for a short period, e.g., 1~2 days, the EC servers in the remote site might be heavily utilized. In contrast, the EC servers of the same app in the metro DC can be very underutilized. Since the condition can be short-lived, it might not make business sense to adjust EC capacity among DCs. Sometimes, UEs swarming to a specific site are not anticipated.

1.5. Problem 3: Application Server Relocation

When an EC server is added to, moved, or deleted from a 5G EC Data Center, the routing protocol needs to propagate the changes to 5G PSA or the PSA adjacent routers. After the change, the cost associated with the site might change as well.

Note: for ease of description, the Edge Application Server and Application Server are used interchangeably throughout this document.

2. Conventions used in this document

A-ER: Egress Router to an Application Server, [A-ER] is used to describe the last router that the Application Server is attached. For a 5G EC environment, the A-ER can be the gateway router to a (mini) Edge Computing Data Center.

Application Server: An application server is a physical or virtual server that hosts the software system for the application.

Application Server Location: Represent a cluster of servers at one location serving the same Application. One application may have a Layer 7 Load balancer, whose address(es) are reachable from an external IP network, in front of a set of application servers. From an IP network perspective, this whole group of servers is considered as the Application server at the location.

Edge Application Server: used interchangeably with Application Server throughout this document.

EC: Edge Computing

Edge Hosting Environment: An environment providing the support required for Edge Application Server's execution.

NOTE: The above terminologies are the same as those used in 3GPP TR 23.758

Edge DC: Edge Data Center, which provides the Edge Computing Hosting Environment. An Edge DC might host 5G core functions in addition to the frequently used application servers.

gNB next generation Node B

L-DN: Local Data Network

PSA: PDU Session Anchor (UPF)

SSC: Session and Service Continuity

UE: User Equipment

UPF: User Plane Function

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.

3. Usage of AppMetaData for 5G Edge Computing

AppMetaData consists of metrics about the running environment at the egress routers to which EC servers are directly attached.

3.1. Assumptions

From the IP Layer, the EC servers or their respective load balancers are identified by their IP addresses. Those IP addresses are the identifiers to the EC servers throughout this document. Here are some assumptions about the 5G EC services:

- Only the registered EC services, which are only a small portion of the services, need to incorporate the destination capacity metrics for optimal forwarding.

- The 5G EC controller or management system can send those EC service identifiers to relevant routers.
- The ingress routers' local BGP path compute algorithm includes a special plugin that can compute the path to the optimal Next Hop (egress router) based on the BGP AppMetaData TLV received for the registered EC services.

The proposed solution is for the egress routers, a.k.a. A-ERs in this document, that have direct links to the EC Servers to collect various measurements about the Servers' running status [5G-EC-Metrics] and advertise the metrics to other routers in 5G EC LDN (Local Data Network).

3.2. IP Layer Metrics to Gauge Application Behavior

[5G-EC-Metrics] describes the IP Layer Metrics that can gauge the application servers running status and environment:

- IP-Layer Metric for App Server Load Measurement:
The Load Measurement to an App Server is a weighted combination of the number of packets/bytes to the App Server and the number of packets/bytes from the App Server which are collected by the A-ER to which the App Server is directly attached.
The A-ER is configured with an ACL that can filter out the packets for the Application Server.
- Capacity Index
a numeric number, configured on all A-ERs in the domain consistently, is used to represent the capacity of the application server attached to an A-ER. At some sites, the IP address exposed to the A-ER is the App Layer Load balancer that have many instances attached. At other sites, the IP address exposed is the server instance itself.
- Site preference index:
is used to describe some sites are more preferred than others. For example, a site with higher bandwidth has a higher preference number than other.

In this document, the term "Application Server Egress Router" [A-ER] is used to describe the last router that an Application Server is attached. For the 5G EC environment, the A-ER can be

the gateway router to the EC DC where multiple Application servers are hosted.

3.3. AppMetaData Constrained Optimal Path Selection

The main benefit of using ANYCAST is to leverage the network layer conditions to select an optimal path to the application instantiated in multiple locations.

When the ingress routers to the 5G LDN are informed of the Load and Capacity Index of the App Servers at different EC data centers, they can incorporate those metrics with the network path conditions for path selection.

Here is an algorithm that computes the cost to reach the App Servers attached to Site-i relative to another site, say Site-b. When the reference site, Site-b, is plugged in the formula, the cost is 1. So, if the formula returns a value less than 1, the cost to reach Site-i is less than reaching Site-b.

$$\text{Cost-i} = (w * \frac{\text{CP-b} * \text{Load-i}}{\text{CP-i} * \text{Load-b}}) + (1-w) * (\frac{\text{Pref-b} * \text{Network-Delay-i}}{\text{Pref-i} * \text{Network-Delay-b}})$$

Load-i: Load Index at Site-i, it is the weighted combination of the total packets or/and bytes sent to and received from the Application Server at Site-i during a fixed time period.

CP-i: capacity index at Site-i, a higher value means higher capacity.

Delay-i: Network latency measurement (RTT) to the A-ER that has the Application Server attached at the site-i.

Pref-i: Preference index for the Site-i, a higher value means higher preference.

w: Weight for load and site information, which is a value between 0 and 1. If smaller than 0.5, Network latency and the site Preference have more influence; otherwise, Server load and its capacity have more influence.

4. BGP Protocol Extension to advertise Load & Capacity

The goal of the BGP extension is for egress routers to propagate the metrics about their running environment to ingress routers. Here are some examples of the metrics propagated by the egress routers:

- the Load Measurement Index for the attached EC Servers,
- the Capacity Index, and
- Site Preference Index.

This section specifies the Load Index Sub-TLV, Capacity Sub-TLV, and the Site Preference Sub-TLV that can be carried by the Tunnel Encap Path Attribute associated with the routes.

4.1. Ingress Node BGP Path Selection Behavior

4.1.1. AppMetaData Influenced BGP Path Selection

When an ingress router receives BGP updates for the same IP address from multiple egress routers, all those egress routers are considered the next hops for the IP address. For the selected EC services, the ingress router's BGP engine would call a Plugin function that can select paths based on the AppMetaData received. The Plugin function is called Load Compute Engine throughout this document.

Assume that both Ra and Rb in Figure-1 have BGP Multipath enabled. As a result, Dst Address: S1:aa08::4450 is resolved via multiple NextHop: R1, R2, R3.

Suppose the local BGP's Load Compute Engine identifies R1 as the optimal NextHop for the flow towards S1:aa08::4450. Then the Load Compute Engine can insert a higher weight for the path R1 so that BGP Best Path is locally influenced by the weight parameter based on the local decision.

4.1.2. Ingress Router Forwarding Behavior

When the ingress router receives a packet and lookup the FIB, it gets the destination prefix's whole path. It encapsulates the packet destined towards the optimal egress node.

For subsequent packets belonging to the same flow, the ingress router needs to forward them to the same egress router unless

the selected egress router is no longer reachable. Keeping packets from one flow to the same egress router, a.k.a. Flow Affinity, is supported by many commercial routers. Most registered EC services have relatively short flows.

How Flow Affinity is implemented is out of the scope for this document. Here is one example to illustrate how Flow Affinity can be achieved. This illustration is not to be standardized.

For the registered EC services, the ingress node keeps a table of

- Service ID (i.e., IP address)
- Flow-ID
- Sticky Egress ID (egress router loopback address)
- A timer

The Flow-ID in this table is to identify a flow, initialized to NULL. How Flow-ID is constructed is out of the scope for this document. Here is one example of constructing the Flow-ID:

- For IPv6, the Flow-ID can be the Flow-ID extracted from the IPv6 packet header with or without the source address.
- For IPv4, the Flow-ID can be the combination of the Source Address with or without the TCP/UDP Port number.

The Sticky Egress ID is the egress node address for the same flow. [5G-Sticky-Service] describes several methods to derive the Sticky Egress ID.

The Timer is always refreshed when a packet with the matching EC Service ID (IP address) is received by the node.

If there is no Stick Egress ID present in the table for the EC Service ID, the forwarding plane can select a NextHop influenced by the Load Compute Engine. The forwarding plane encapsulates the packet with a tunnel to the chosen NextHop. The chosen NextHop and the Flow ID are recorded in the EC Service table entry.

When the selected optimal NextHop (egress router) is no longer reachable, refer to Section 6 Soft Anchoring on how another path is selected.

4.1.3. Forwarding Behavior when UEs moving to new 5G Sites

When a UE moves to a new 5G eNB which is anchored to the same UPF, the packets from the UE traverse to the same ingress router. Path selection and forwarding behavior are same as before.

When the new eNB is anchored to a different UPF, the packets from the UE traverse a different ingress router. If the UE source IP address has been changed, indicating the new UPF might belong to a different administrative domain, the new ingress router treats the packets from the UE as a new flow and select the optimal path based on the configured policies. If the UE maintains the same IP address when anchored to a new UPF, the directly connected ingress router might use the pre-computed Egress Router, which is passed from a neighboring router. [5G-Edge-Sticky] describes methods for the ingress router connected to the UPF in the new site to consider the information passed from other ingress routers in selecting the optimal paths. The detailed algorithm is out of the scope of this document.

5. The Sub-TLVs for AppMetaData

The AppMetaData attribute is encoded in an optional subTLV within the Tunnel Encap [RFC9012] Path Attribute.

All values in the Sub-TLVs are unsigned 32 bits integers.

5.1. Load Measurement sub-TLV format

Two types of Load Measurement Sub-TLVs are specified. One is to carry the aggregated cost Index based on a weighted combination of the collected measurements; another one is to carry the raw measurements of packets/bytes to/from the App Server address. The raw measurement is useful when ingress routers have embedded analytics relying on the raw measurements.

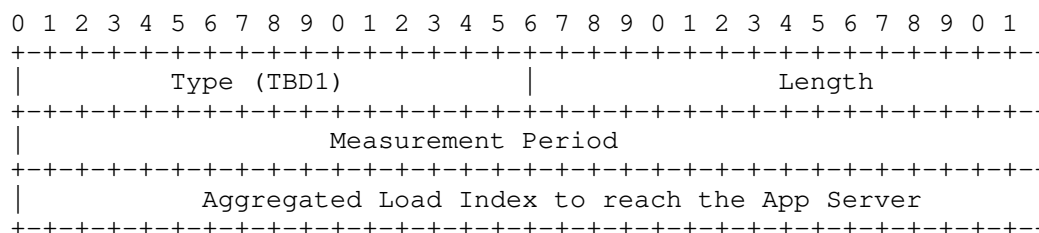


Figure 2: Aggregated Load Index Sub-TLV

Raw Load Measurement sub-TLV has the following format:

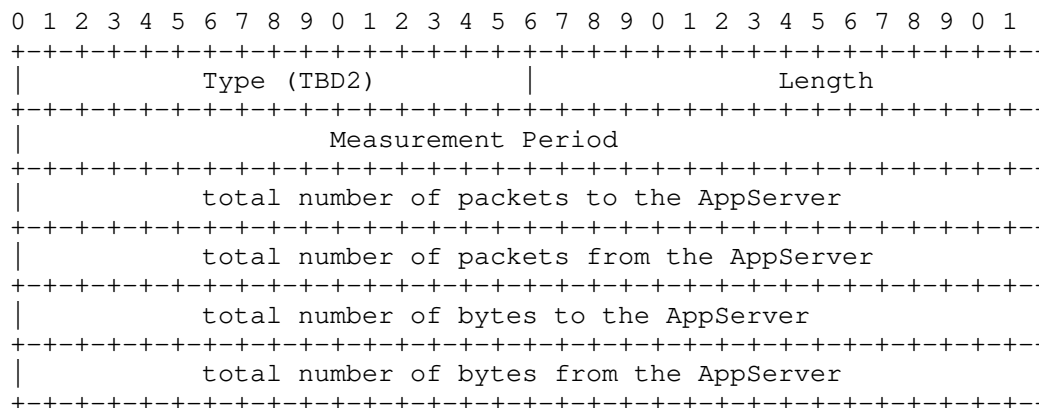


Figure 3: Raw Load Measurement Sub-TLV

Type =TBD1: Aggregated Load Measurement Index derived from the Weighted combination of bytes/packets sent to/received from the App server:

$$\text{Index} = w1 * \text{ToPackets} + w2 * \text{FromPackets} + w3 * \text{ToBytes} + w4 * \text{FromBytes}$$

Where w_i is a value between 0 and 1; $w1 + w2 + w3 + w4 = 1$.

Type= TBD2: Raw measurements of packets/bytes to/from the App Server address.

Measure Period: BGP Update period or user-specified period.

5.2. Capacity Index sub-TLV format

The Capacity Index sub-TLV has the following format:

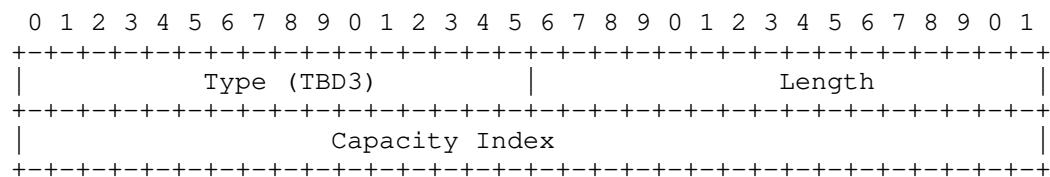


Figure 4: Capacity Index Sub-TLV

Note: "Capacity Index" can be more stable for each site. If those values are configured to nodes, they might not need to be included in every BGP UPDATE.

5.3. The Site Preference Index sub-TLV format

The site Preference Index is used to achieve Soft Anchoring [Section 5] an application flow from a UE to a specific location when the UE moves from one 5G site to another.

The Preference Index sub-TLV has the following format:

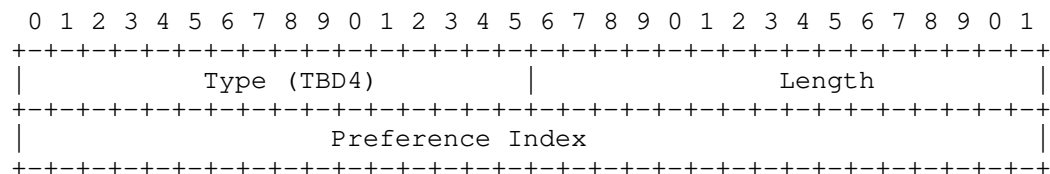


Figure 5: Preference Index Sub-TLV

Note: "Site Preference Index" can be more stable for each site. If those values are configured to nodes, they might not need to be included in every BGP UPDATE.

6. AppMetaData Propagation Scope

AppMetaData is only to be distributed to the relevant ingress nodes of the 5G EC local data networks. Only the ingress routers that are configured with the 5G EC services need to receive the AppMetaData for specific Service IDs.

For each registered EC service, a corresponding filter group can be formed on RR to represent the interested ingress routers that are interested in receiving the corresponding AppMetaData information.

7. Minimum Interval for Metrics Change Advertisement

As the metrics change can impact the path selection, the Minimum Interval for Metrics Change Advertisement is configured to control the update frequency to avoid route oscillations. Default is 30s.

Significant load changes at EC data centers can be triggered by short-term gatherings of UEs, like conventions, lasting a few hours or days, which are too short to justify adjusting EC server capacities among DCs. Therefore, the load metrics change rate can be in the magnitude of hours or days.

8. Soft Anchoring of an ANYCAST Flow

"Sticky Service" in the 3GPP Edge Computing specification (3GPP TR 23.748) is about flows from a UE sticking to a specific location when the UE moves from one 5G Site to another.

"Soft Anchoring" is a mechanism for ingress routers to apply preference to the path towards the previous server location when the UE is anchored to a new UPF and continue using its cached IP for the EC server.

Let's assume one application "App.net" is instantiated on four servers that are attached to four different routers R1, R2, R3, and R4 respectively. It is desired for packets to the "App.net" from UE-1 to stick with one server, say the App Server attached to R1, even when the UE moves from one 5G site to another. However, when there is a failure reaching R1 or the Application Server attached to R1, the packets of the flow "App.net" from UE-1 need to be forwarded to the Application Server attached to R2, R3, or R4.

We call this kind of sticky service "Soft Anchoring", meaning that anchoring to the site of R1 is preferred, but other sites can be chosen when the preferred site encounters a failure.

Here is a mechanism to achieve Soft Anchoring:

- Assign a group of ANYCAST addresses to one application. For example, "App.net" is assigned with 4 ANYCAST addresses, L1, L2, L3, and L4. L1/L2/L3/L4 represents the location preferred ANYCAST addresses.
- For the App.net Server attached to a router, the router has four Stub links to the same Server, L1, L2, L3, and L4 respectively. The cost to L1, L2, L3, and L4 is assigned differently for different egress routers. For example,
 - o When attached to R1, the L1 has the lowest cost, say 10, when attached to R2, R3, and R4, the L1 can have a higher cost, say 30.
 - o ANYCAST L2 has the lowest cost when attached to R2, higher cost when attached to R1, R3, R4 respectively.
 - o ANYCAST L3 has the lowest cost when attached to R3, higher cost when attached to R1, R2, R4 respectively, and
 - o ANYCAST L4 has the lowest cost when attached to R4, higher cost when attached to R1, R2, R3 respectively
- When a UE queries for the "App.net" for the first time, the DNS reply has the location preferred ANYCAST address, say L1, based on where the query is initiated.
- When the UE moves from one 5G site-A to Site-B, UE continues sending packets of the "App.net" to ANYCAST address L1. The routers will continue sending packets to R1 because the total cost for the App.net instance for ANYCAST L1 is lowest at R1. If any failure occurs making R1 not reachable, the packets of the "App.net" from UE-1 will be sent to R2, R3, or R4 (depending on the total cost to reach L1 attached to R2/R3/R4).

If the Application Server supports the HTTP redirect, more optimal forwarding can be achieved.

- When a UE queries for the "App.net" for the first time, the global DNS reply has the ANYCAST address G1, which has the same cost regardless of where the Application servers are attached.
- When the UE initiates the communication to G1, the packets from the UE will be sent to the Application Server that has the lowest cost, say the Server attached to R1. The Application server is instructed with HTTPs Redirect to reply with a location-specific URL, say App.net-Loc1. The client on the UE will query the DNS for App.net-Loc1 and get the response of ANYCAST L1. The subsequent packets from the UE-1 for App.net are sent to L1.

9. Manageability Considerations

To be added.

10. Security Considerations

To be added.

11. IANA Considerations

Here are new Sub-TLV types requiring IANA registration:

Type = TBD1: Aggregated Load Measurement Index derived from the Weighted combination of bytes/packets sent to/received from the App server.

Type = TBD2: Raw measurements of packets/bytes to/from the App Server address.

Type = TBD3: Capacity value sub-TLV

Type = TBD4: Site preference value sub-TLV

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4364] E. rosen, Y. Rekhter, "BGP/MPLS IP Virtual Private networks (VPNs)", Feb 2006.
- [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>>.
- [RFC8200] s. Deering R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", July 2017

12.2. Informative References

- [3GPP-EdgeComputing] 3GPP TR 23.748, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on enhancement of support for Edge Computing in 5G Core network (5GC)", Release 17 work in progress, Aug 2020.
- [5G-EC-Metrics] L. Dunbar, H. Song, J. Kaippallimalil, "IP Layer Metrics for 5G Edge Computing Service", draft-dunbar-ippm-5g-edge-compute-ip-layer-metrics-00, work-in-progress, Oct 2020.
- [5G-Edge-Sticky] L. Dunbar, J. Kaippallimalil, "IPv6 Solution for 5G Edge Computing Sticky Service", draft-dunbar-6man-5g-ec-sticky-service-00, work-in-progress, Oct 2020.

[RFC5521] P. Mohapatra, E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", April 2009.

[BGP-SDWAN-Port] L. Dunbar, H. Wang, W. Hao, "BGP Extension for SDWAN Overlay Networks", draft-dunbar-idr-bgp-sdwan-overlay-ext-03, work-in-progress, Nov 2018.

[SDWAN-EDGE-Discovery] L. Dunbar, S. Hares, R. Raszuk, K. Majumdar, "BGP UPDATE for SDWAN Edge Discovery", draft-dunbar-idr-sdwan-edge-discovery-00, work-in-progress, July 2020.

[Tunnel-Encap] E. Rosen, et al "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10, Aug 2018.

13. Acknowledgments

Acknowledgements to Donald Eastlake for their review and contributions.

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

Authors' Addresses

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

Kausik Majumdar
CommScope
350 W Java Drive, Sunnyvale, CA 94089
Email: kausik.majumdar@commscope.com

Haibo Wang
Huawei
Email: rainsword.wang@huawei.com

Gyan Mishra
Verizon
Email: gyan.s.mishra@verizon.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: 8 August 2022

S. Hares
Hickory Hill Consulting
D. Eastlake
Futurewei Technologies
C. Yadlapalli
ATT
S. Maduscke
Verizon
4 February 2022

BGP Flow Specification Version 2
draft-hares-idr-flowspec-v2-05

Abstract

BGP flow specification version 1 (FSv1), defined in RFC 8955, RFC 8956, and RFC 9117 describes the distribution of traffic filter policy (traffic filters and actions) distributed via BGP. Multiple applications have used BGP FSv1 to distribute traffic filter policy. These applications include the following: mitigation of denial of service (DoS), enabling traffic filtering in BGP/MPLS VPNs, centralized traffic control of router firewall functions, and SFC traffic insertion.

During the deployment of BGP FSv1 a number of issues were detected due to lack of consistent TLV encoding for rules for flow specifications, lack of user ordering of filter rules and/or actions, and lack of clear definition of interaction with BGP peers not supporting FSv1. Version 2 of the BGP flow specification (FSv2) protocol addresses these features. In order to provide a clear demarcation between FSv1 and FSv2, a different NLRI encapsulates FSv2.

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 8 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	4
1.1. Definitions and Acronyms	5
1.2. RFC 2119 language	6
2. Flow Specification	6
2.1. Flow Specification v1 (FSv1) Overview	7
2.2. Flow Specification v2 (FSv2) Overview	9
3. FSv2 Filters and Actions	12
3.1. IP header SubTLV (type=1)	14
3.1.1. IP Destination Prefix (type = 1)	16
3.1.2. IP Source Prefix (type = 2)	16
3.1.3. IP Protocol (type = 3)	17
3.1.4. Port (type = 4)	17
3.1.5. Destination Port (type = 5)	18
3.1.6. Source Port (type = 6)	18
3.1.7. ICMP Type (type = 7)	18
3.1.8. ICMP Code (type = 8)	19
3.1.9. TCP Flags (type = 9)	19
3.1.10. Packet length (type = 10 (0x0A))	19
3.1.11. DSCP (Differentiated Services Code Point) (type = 11 (0x0B))	20
3.1.12. Fragment (type = 12 (0x0C))	20
3.1.13. Flow Label (type = 13 (0x0D))	21
3.1.14. TTL (type=14 (0x0E))	21
3.1.15. Parts of SID (type = 15 (0xF))	21
3.1.16. MPLS Label Match1 (type=16, 0x10)	24
3.1.17. MPLS Label Match 2: Experimental bits match on top label (Type=17 (0x11))	25
3.2. Encoding of FSV2 Actions (type=2)	26
3.2.1. Action Chain operation (ACO) (1, 0x01)	28
3.2.2. Traffic Actions per interface set (TAIS) (2, 0x02)	29

3.2.3.	Traffic rate limited by bytes (TRB) (6, 0x06)	30
3.2.4.	Traffic Action (TA) (7, 0x07)	30
3.2.5.	Redirect to IPv4 (RDIPv4) (8, 0x08)	31
3.2.6.	Traffic marking (TM) (9, 0x09)	32
3.2.7.	Traffic rate limited by packets (TRP) (12, 0xC)	33
3.2.8.	Traffic redirect to IPv6 (RDIPv6) (13, 0xD)	33
3.2.9.	Traffic insertion in SFC (TISFC) (14, 0xE)	34
3.2.10.	Flow Specification Redirect to Indirection-ID (RDIID) (15, 0x0F)	35
3.2.11.	MPLS Label Action (MPLSLA) (16, 0x10)	36
3.2.12.	VLAN action (VLAN) (22, 0x16)	37
3.2.13.	TPID action (TPID) (23, 0x17)	39
3.3.	Extended Community vs. Action SubTLV formats	39
3.4.	L2 Traffic Rules	42
3.5.	SFC Traffic Rules	43
3.6.	BGP/MPLS VPN IP Traffic Rules	45
3.7.	BGP/MPLS VPN L2 Traffic Rules	45
3.8.	Encoding of Actions passed in Wide Communities	46
4.	Validation of FSv2 NLRI	47
4.1.	Validation of FS NLRI (FSv1 or FSv2)	47
4.2.	Validation of Flow Specification Actions	49
4.3.	Error handling and Validation	50
5.	Ordering for Flow Specification v2 (FSv2)	50
5.1.	Ordering of FSv2 NLRI Filters	50
5.2.	Ordering of the Actions	52
5.2.1.	Action Chain Operation (ACO)	52
5.2.2.	Summary of FSv2 ordering	55
6.	Ordering of FS filters for BGP Peers support FSv1 and FSv2	56
7.	Scalability and Aspirations for FSv2	58
8.	Optional Security Additions	59
8.1.	BGP FSv2 and BGPSEC	59
8.2.	BGP FSv2 with ROA	60
9.	IANA Considerations	60
9.1.	Flow Specification V2 SAFIs	60
9.2.	BGP Capability Code	61
9.3.	Filter IP Component types	61
9.4.	FSV2 NLRI TLV Types	62
9.5.	Wide Community Assignments	63
10.	Security Considerations	64
11.	References	64
11.1.	Normative References	64
11.2.	Informative References	67
	Authors' Addresses	68

1. Introduction

Modern IP routers have the capability to forward traffic and to classify, shape, rate limit, filter, or redirect packets based on administratively defined policies. These traffic policy mechanisms allow the operator to define match rules that operate on multiple fields within header of an IP data packet. The traffic policy allows actions to be taken upon a match to be associated with each match rule. These rules can be more widely defined as "event-condition-action" (ECA) rules where the event is always the reception of a packet.

BGP ([RFC4271]) flow specification as defined by [RFC8955], [RFC8956], [RFC9117] specifies the distribution of traffic filter policy (traffic filters and actions) via BGP to a mesh of BGP peers (IBGP and EBGP peers). The traffic filter policy is applied when packets are received on a router with the flow specification function turned on. The flow specification protocol defined in [RFC8955], [RFC8956], and [RFC9117] will be called BGP flow specification version 1 (BGP FSv1) in this draft.

Some modern IP routers also include the abilities of firewalls which can match on a sequence of packet events based on administrative policy. These firewall capabilities allow for user ordering of match rules and user ordering of actions per match.

Multiple deployed applications currently use BGP FSv1 to distribute traffic filter policy. These applications include: 1) mitigation of Denial of Service (DoS), 2) traffic filtering in BGP/MPLS VPNS, and 3) centralized traffic control for networks utilizing SDN control of router firewall functions, 4) classifiers for insertion in an SFC, and 5) filters for SRv6 (segment routing v6).

During the deployment of BGP flow specification v1, the following issues were detected:

- * lack of consistent TLV encoding prevented extension of encodings,
- * inability to allow user defined order for filtering rules,
- * inability to order actions to provide deterministic interactions or to allow users to define order for actions, and
- * no clearly defined mechanisms for BGP peers which do not support flow specification v1.

Networks currently cope with some of these issues by limiting the type of traffic filter policy sent in BGP. Current Networks do not have a good workaround/solution for applications that receive but do not understand FSv1 policies.

This document defines version 2 of the BGP flow specification protocol to address these shortcomings in BGP FSv1. Version 2 of BGP flow specification will be denoted as BGP FSv2.

BGP FSv1 as defined in [RFC8955], [RFC8956], and [RFC9117] specified 2 SAFIs (133, 134) to be used with IPv4 AFI (AFI = 1) and IPv6 AFI (AFI=2).

This document specifies 2 new SAFIs (TBD1, TBD2) for FSv2 to be used with 5 AFIs (1, 2, 6, 25, and 31) to allow user-ordered lists of traffic match filters for user-ordered traffic match actions encoded in Communities (Wide or Extended).

FSv1 and FSv2 use different AFI/SAFIs to send flow specification filters. Since BGP route selection is performed per AFI/SAFI, this approach can be termed "ships in the night" based on AFI/SAFI.

FSv1 is a critical component of deployed applications. Therefore, this specification defines how FSv2 will interact with BGP peers that support either FSv2, FSv1, FSv2 and FSv1, or neither of them. It is expected that a transition to FSv2 will occur over time as new applications require FSv2 extensibility and user-defined ordering for rules and actions or network operators tire of the restrictions of FSv1 such as error handling issues and restricted topologies.

Section 2 contains the definition of Flow specification, a short review of FSv1 and an overview of FSv2. Section 3 contains the encoding rules for FSv2 and user-based encoding sent via BGP. Section 4 describes how to validate FSv2 NLRI. Section 5 discusses how to order FSv2 rules. Section 6 covers combining FSv2 user-ordered match rules and FSv1 rules. Section 6 also discusses how to combine user-ordered actions, FSv1 actions, and default actions. Sections 7-10 address an alternate security mechanism, considerations for IANA, security in deployments, and scalability aspirations.

1.1. Definitions and Acronyms

AFI - Address Family Identifier

AS - Autonomous System

BGPSEC - secure BGP [RFC8205] updated by [RFC8206]

BGP Session ephemeral state - state which does not survive the loss of BGP peer session.

Configuration state - state which persist across a reboot of software module within a routing system or a reboot of a hardware routing device.

DDOs - Distributed Denial of Service.

Ephemeral state - state which does not survive the reboot of a software module, or a hardware reboot. Ephemeral state can be ephemeral configuration state or operational state.

FSv1 - Flow Specification version 1 [RFC8955] [RFC8956]

FSv2 - Flow Specification version 2 (this document)

NETCONF - The Network Configuration Protocol [RFC6241].

RESTCONF - The RESTCONF configuration Protocol [RFC8040]

RIB - Routing Information Base.

ROA - Route Origin Authentication [RFC6482]

RR - Route Reflector.

SAFI - Subsequent Address Family Identifier

1.2. RFC 2119 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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals as shown here.

2. Flow Specification

A BGP Flow Specification is an n-tuple containing one or more match criteria that can be applied to IP traffic, traffic encapsulated in IP traffic or traffic associated with IP traffic. The following are examples of such traffic: IP packet or an IP packet inside a L2 packet (Ethernet), an MPLS packet, and SFC flow.

A given Flow Specification NLRI may be associated with a set of path attributes depending on the particular application, and attributes within that set may or may not include reachability information

(e.g., NEXT_HOP). Extended Community or Wide Community attributes (well-known or AS-specific) MAY be used to encode a set of pre-determined actions.

A particular application is identified by a specific AFI/SAFI (Address Family Identifier/Subsequent Address Family Identifier) and corresponds to a distinct set of RIBs. Those RIBs should be treated independently of each other in order to assure noninterference between distinct applications.

BGP processing treats the NLRI as a key to entries in AFI/SAFI BGP databases. Entries that are placed in the Loc-RIB are then associated with a given set of semantics which are application dependent. Standard BGP mechanisms such as update filtering by NLRI or by attributes such as AS_PATH or large communities apply to the BGP Flow Specification defined NLRI-types.

Network operators can control the propagation of BGP routes by enabling or disabling the exchange of routes for a particular AFI/SAFI pair on a particular peering session. As such, the Flow Specification may be distributed to only a portion of the BGP infrastructure.

2.1. Flow Specification v1 (FSv1) Overview

The FSv1 NLRI defined in [RFC8955] and [RFC8956] include 13 match conditions encoded for the following AFI/SAFIs:

- * IPv4 traffic: AFI:1, SAFI:133
- * IPv6 Traffic: AFI:2, SAFI:133
- * BGP/MPLS IPv4 VPN: AFI:1, SAFI: 134
- * BGP/MPLS IPv6 VPN: AFI:2, SAFI: 134

If one considers the reception of the packet as an event, then BGP FSv1 describes a set of Event-MatchCondition-Action (ECA) policies where:

- * event is the reception of a packet,
- * condition stands for "match conditions" defined in the BGP NLRI as an n-tuple of component filters, and
- * the action is either: the default condition (accept traffic), or a set of actions (1 or more) defined in Extended BGP Community values [RFC4360].

The flow specification conditions and actions combine to make up FSv1 specification rules. Each FSv1 NLRI must have a type 1 component (destination prefix). Extended Communities with FSv1 actions can be attached to a single NLRI or multiple NLRIs in a BGP message

Within an AFI/SAFI pair, FSv1 rules are ordered based on the components in the packet (types 1-13) ordered from left-most to right-most and within the component types by value of the component. Rules are inserted in the rule list by component-based order where an FSv1 rule with existing component type has higher precedence than one missing a specific component type,

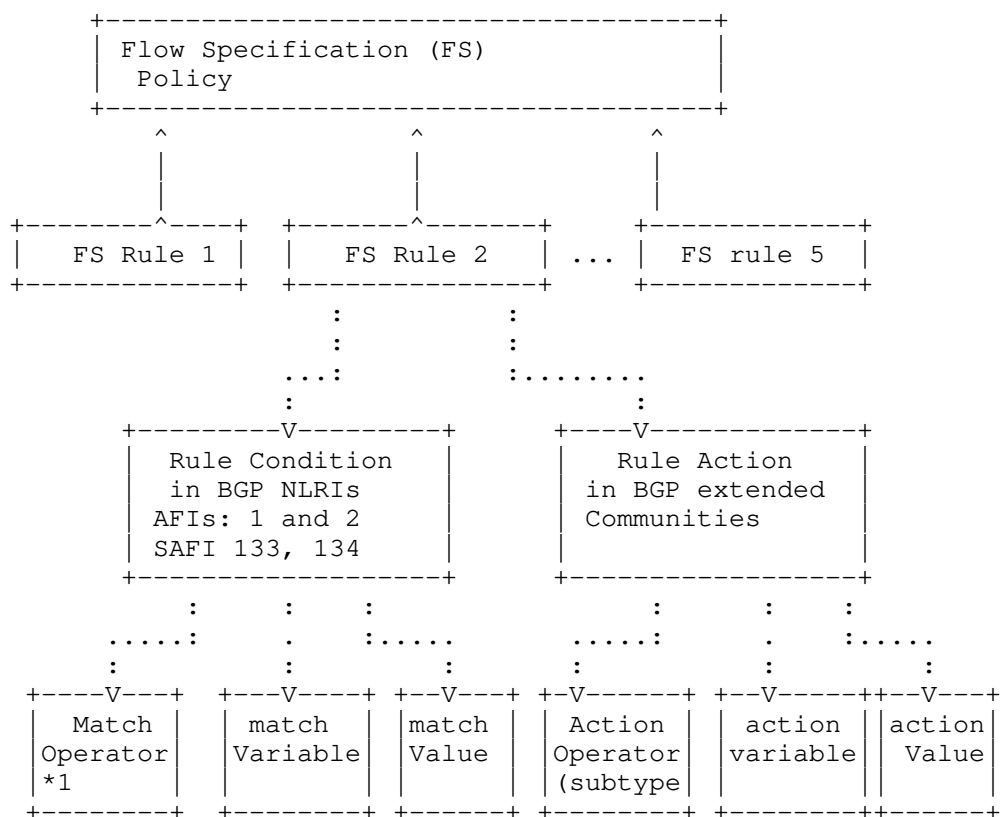
Since FSv1 specifications ([RFC8955], [RFC8956], and [RFC9117]) specify that the FSv1 NLRI MUST have a destination prefix (as component type 1) embedded in the flow specification, the FSv1 rules with destination components are ordered by IP Prefix comparison rules for IPv4 ([RFC8955]) and IPv6 ([RFC8956]). [RFC8955] specifies that more specific prefixes (aka longest match) have higher precedence than that of less specific prefixes and that for prefixes of the same length the lower IP number is selected (lowest IP value). [RFC8955] specifies that if the offsets within component 1 are the same, then the longest match and lowest IP comparison rules from [RFC8955] apply. If the offsets are different, then the lower offset has precedence.

These rules provide a set of FSv1 rules ordered by IP Destination Prefix by longest match and lowest IP address. [RFC8955] also states that the requirement for a destination prefix component "MAY be relaxed by explicit configuration" Since the rule insertions are based on comparing component types between two rules in order, this means the rules without destination prefixes are inserted after all rules which contain destination prefix component.

The actions specified in FSv1 are:

- * accept packet (default),
- * traffic flow limitation by bytes (0x6),
- * traffic-action (0x7),
- * redirect traffic (0x8),
- * mark traffic (0x9), and
- * traffic flow limitation by packets (12, 0xC)

Figure 1 shows a diagram of the FSv1 logical data structures with 5 rules. If FSv1 rules have destination prefix components (type=1) and FSv1 rule 5 does not have a destination prefix, then FSv1 rule 5 will be inserted in the policy after rules 1-4.



*1 match operator may be complex.

Figure 2-1: BGP Flow Specification v1 Policy

2.2. Flow Specification v2 (FSv2) Overview

Flow Specification v2 allows the user to order the flow specification rules and the actions associated with a rule. Each FSv2 rule may have one or more match conditions and one or more associated actions.

This FSv2 specification supports the components and actions for the following:

- * IPv4 (AFI=1, SAFI=TBD1),
- * IPv6 (AFI=2, SAFI=TBD2),
- * L2 (AFI=6, SAFI=TBD1),
- * BGP/MPLS IPv4 VPN: (AFI=1, SAFI=TBD2),
- * BGP/MPLS IPv6 VPN: (AFI=2, SAFI=TBD2),
- * BGP/MPLS L2VPN (AFI=25, SAFI=TBD2),
- * SFC: (AFI=31, SAFI=TBD1), and
- * SFC VPN (AFI=31, SAFI=TBD2).

The FSv2 specification for tunnel traffic is outside the scope of this specification. The FSv1 specification for tunneled traffic is in [I-D.ietf-idr-flowspec-nvo3].

FSv2 operates in the ships-in-the night model with FSv1 so network operators can manipulate which the distribution of FSv2 and FSv1 using configuration parameters. Since the lack of deterministic ordering was an FSv1 problem, this specification provides rules and protocol features to keep filters in a deterministic order between FSv1 and FSv2.

The basic principles regarding ordering of flow specification filter rules are:

- 1) Rule-0 (zero) is defined to be 0/0 with the "permit-all" action.
- 2) FSv2 rules are ordered based on user-specified order.
 - The user-specified order is carried in the FSv2 NLRI and a numerical lower value takes precedence over a numerically higher value. For rules received with the same order value, the FSv1 rules apply (order by component type and then by value of the components).
- 3) FSv2 rules are added starting with Rule 1 and FSv1 rules are added after FSv2 rules
 - For example, BGP Peer A has FSv2 data base with 10 FSv2 rules (1-10). FSv1 user number is configured to start at 301 so 10 FSv1 rules are added at 301-310.

4) An FSv2 peer may receive BGP NLRI routes from a FSv1 peer or a BGP peer that does not support FSv1 or FSv2. The capabilities sent by a BGP peer indicate whether the AFI/SAFI can be received (FSv1 NLRI or FSv2 NLRI).

5) Associate a chain of actions to rules based on user-defined action number (1-n). (optional)

- If no actions are associated with a filter rule, the default is to drop traffic the filter rules match
- An action chain of 1-n actions can be associated with a set of filter rules can via Extended Communities or Wide Communities. Only Wide Communities can associate a user-defined order for the actions. Extended Community actions occur after actions with a user specified order (see section 5.2 for details).

Figure 2-2 provides a logical diagram of the FSv2 structure

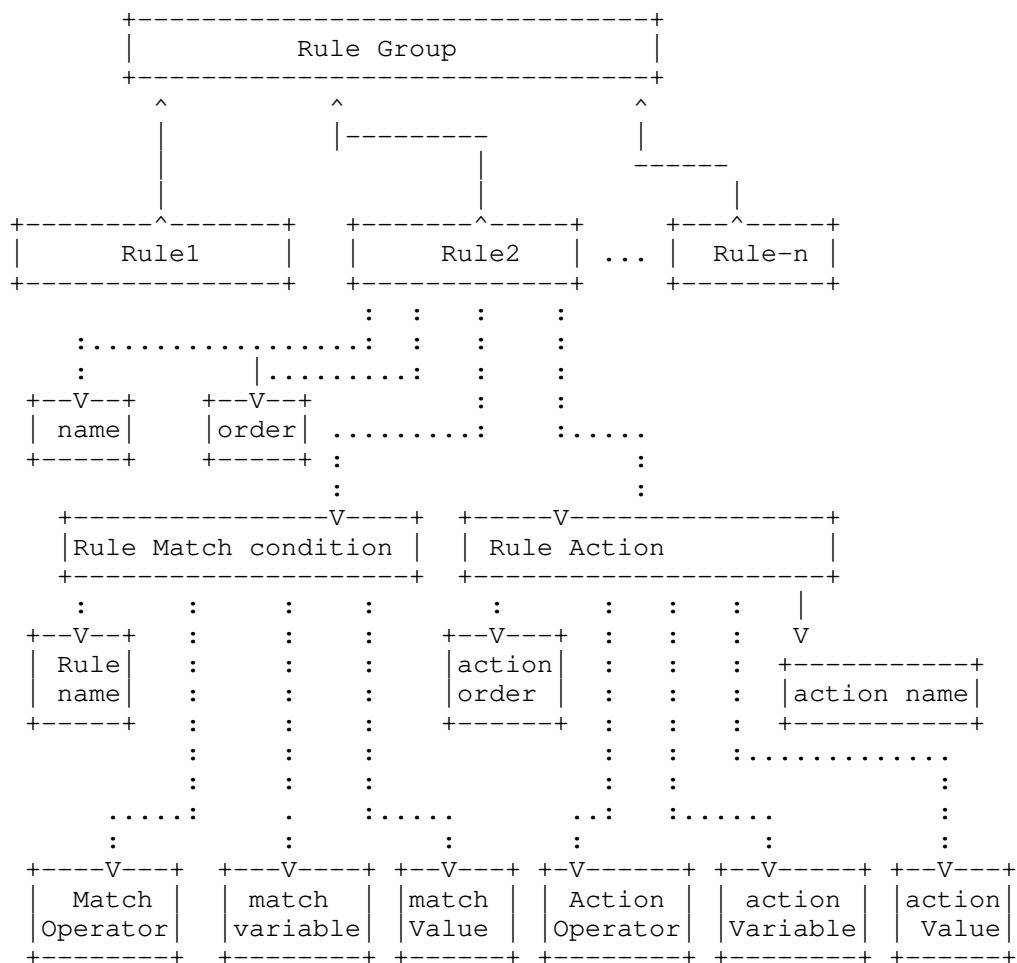


Figure 2-2: BGP FSv2 Data storage

3. FSv2 Filters and Actions

The BGP FSv2 uses an NRLI with the format for AFIs for IPv4 (AFI = 1), IPv6 (AFI = 2), L2 (AFI = 6), L2VPN (AFI=25), and SFC (AFI=31) with two following SAFIs to support transmission of the flow specification which supports user ordering of traffic filters and actions for IP traffic and IP VPN traffic.

This NLRI information is encoded using MP_REACH_NLRI and MP_UNREACH_NLRI attributes defined in [RFC4760]. When advertising FSv2 NLRI, the length of the Next-Hop Network Address MUST be set to 0. Upon reception, the Network Address in the Next-Hop field MUST be ignored.

Implementations wishing to exchange flow specification rules MUST use BGP's Capability Advertisement facility to exchange the Multiprotocol Extension Capability Code (Code 1) as defined in [RFC4760], and indicate a capability for FSv1, FSv2 (Code TBD3), or both.

The AFI/SAFI NLRI for BGP Flow Specification version 2 (FSv2) has the format:

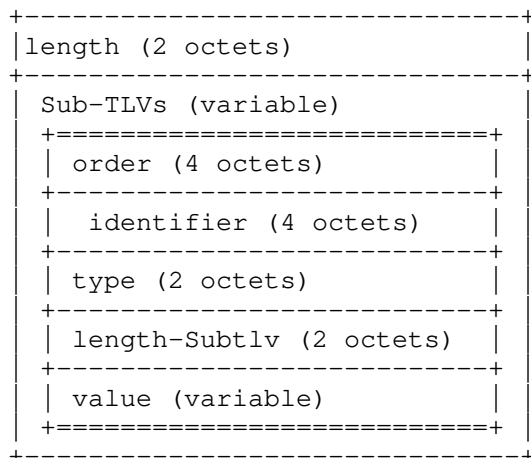


Figure 3-1: FSv2 format

where:

- * length: length of field including all SubTLVs in octets.
 - The combined lengths of any FSv2 NLRI in the MP_REACH_NLRI or MP_UNREACH_NLRI. The BGP NLRI length must be less than the packet size minus the other fields (BGP header, BGP Path Attributes, and NLRI).
- * order: flow-specification global rule order number (4 octets).
- * identifier: identifier for the rule (used for NM/Logging) (4 octets)

- * type: contains a type for FSv2 TLV format of the NRLI (2 octets) which can be:
 - 0 - reserved,
 - 1 - IP Traffic Rules
 - 2- L2 traffic rules
 - 3- SFC Traffic rules
 - 4- SFC VPN Traffic rules
 - 5 - BGP/MPLS VPN IP Traffic Rules
 - 6 - BGP/MPLS VPN L2 Traffic Rules
- * length-Subtlv: is the length of the value part of the Sub-TLV,
- * value: value depends on the subTLV (see sections below).

3.1. IP header SubTLV (type=1)

The format of the IP header TLV value field is shown in figure 4. The AFI/SAFI field includes the AFI (2 octets), SAFI (1 octet). The AFI will be 1 (IPv4) or 2 (IPv6) and the SAFI will be TBD1 or, for the VPN case, TBD2. The IP header for the VPN case is specified in section 3.5.

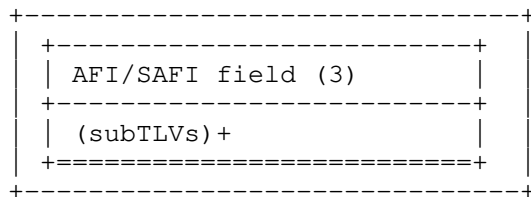


Figure 3-2 - IP Header TLV

Where: AFI is 1 (IPv4) or 2 (IPv6) and SAFI is TBD1.

Each SubTLV has the format:

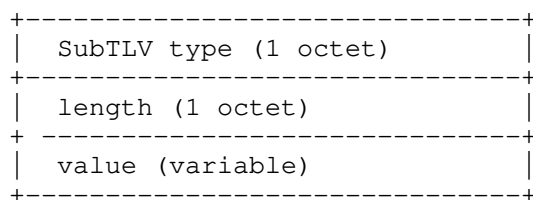


Figure 3-3 IP header SubTLV format

Where:

SubTLV type: component values are defined in the "Flow Specification Component types" registry for IPv4 and IPv6 by [RFC8955], [RFC8956], and [I-D.ietf-idr-flowspec-srv6]

length: length of SubTLV (varies depending on SubTLV type).

value: dependent on the subTLV

- For descriptions of value portions for components 1-13 see [RFC8955] and [RFC8956]. For component 14 see [I-D.ietf-idr-flowspec-srv6].

Many of the components use the operators [numeric_op] and [bitmask_op] defined in [RFC8955]

The list of valid SubTLV types appears in Table 2.

Table 2 IP SubTLV Types for IP Filters

SubTLV-type	Definition
=====	=====
1	IP Destination prefix
2	IP Source prefix
3	IPv4 Protocol / IPv6 Upper Layer Protocol
4	Port
5	Destination Port
6	Source Port
7	ICMPv4 type / ICMPv6 type
8	ICMPv4 code / ICPv6 code
9	TCP Flags
10	Packet length
11	DSCP (Differentiated Services Code Point)
12	Fragment
13	Flow Label
14	- TTL
15	Parts of SID
16	- MPLS Match 1: Label in Label stack
17	- MPLS Match 2: EXP bits in top Label

Ordering within the TLV in FSv2: The transmission of SubTLVs within a flow specification rule MUST be sent ascending order by SubTLV type. If the SubTLV types are the same, then the value fields are compared using mechanisms defined in [RFC8955] and [RFC8956] and MUST be in ascending order. NLRIs having TLVs which do not follow the above ordering rules MUST be considered as malformed by a BGP FSv2 propagator. This rule prevents any ambiguities that arise from the multiple copies of the same NLRI from multiple BGP FSv2 propagators. A BGP implementation SHOULD treat such malformed NLRIs as "Treat-as-withdraw" [RFC7606].

See [RFC8955], [RFC8956], and [I-D.ietf-idr-flowspec-srv6]. for specific details.

3.1.1. IP Destination Prefix (type = 1)

IPv4 Name: IP Destination Prefix (reference: [RFC8955])

IPv6 Name: IPv6 Destination Prefix (reference: [RFC8956])

IPv4 length: Prefix length in bits

IPv4 value: IPv4 Prefix (variable length)

IPv6 length: length of value

IPv6 value: [offset (1 octet)] [pattern (variable)]
[padding(variable)]

If IPv6 length = 0 and offset = 0, then component matches every address. Otherwise, length must be offset "less than" length "less than" 129 or component is malformed.

3.1.2. IP Source Prefix (type = 2)

IPv4 Name: IP Source Prefix (reference: [RFC8955])

IPv6 Name: IPv6 Source Prefix (reference: [RFC8956])

IPv4 length: Prefix length in bits

IPv4 value: Source IPv4 Prefix (variable length)

IPv6 length: length of value

IPv6 value: [offset (1 octet)] [pattern
(variable)] [padding(variable)]

If IPv6 length = 0 and offset = 0, then component matches every address. Otherwise, length must be offset < length < 129 or component is malformed.

3.1.3. IP Protocol (type = 3)

IPv4 Name: IP Protocol IP Source Prefix (reference: [RFC8955])

IPv6 Name: IPv6 Upper-Layer Protocol: (reference: [RFC8956])

IPv4 length: variable

IPv4 value: [numeric_op, value]+

IPv6 length: variable

IPv6 value: [numeric_op, value]+

where the value following each numeric_op is a single octet.

3.1.4. Port (type = 4)

IPv4/IPv6 Name: Port (reference: [RFC8955]), [RFC8956])

Filter defines: a set of port values to match either destination port or source port.

IPv4 length: variable

IPv4 value: [numeric_op, value]+

IPv6 length: variable

IPv6 value: [numeric_op, value]+

where the value following each numeric_op is a single octet.

Note-1: (from FSV1) In the presence of the port component (destination or source port), only a TCP (port 6) or UDP (port 17) packet can match the entire flow specification. If the packet is fragmented and this is not the first fragment, then the system may not be able to find the header. At this point, the FSv2 filter may fail to detect the correct flow. Similarly, if other IP options or the encapsulating security payload (ESP) is present, then the node may not be able to describe the transport header and the FSv2 filter may fail to detect the flow.

The restriction in note-1 comes from the inheritance of the FSv1 filter component for port. If better resolution is desired, a new FSv2 filter should be defined.

Note-2: FSv2 component only matches the first upper layer protocol value.

3.1.5. Destination Port (type = 5)

IPv4/IPv6 Name: Destination Port (reference: [RFC8955]), [RFC8956])

Filter defines: a list of match filters for destination port for TCP or UDP within a received packet

Length: variable

Component Value format: [numeric_op, value]+

3.1.6. Source Port (type = 6)

IPv4/IPv6 Name: Source Port (reference: [RFC8955]), [RFC8956])

Filter defines: a list of match filters for source port for TCP or UDP within a received packet

IPv4/IPv6 length: variable

IPv4/IPv6 value: [numeric_op, value]+

3.1.7. ICMP Type (type = 7)

IPv4: ICMP Type (reference: [RFC8955])

Filter defines: Defines: a list of match criteria for ICMPv4 type

IPv6: ICMPv6 Type (reference: [RFC8956])

Filter defines: a list of match criteria for ICMPv6 type.

IPv4/IPv6 length: variable

IPv4/IPv6 value: [numeric_op, value]+

3.1.8. ICMP Code (type = 8)

IPv4: ICMP Type (reference: [RFC8955])

Filter defines: a list of match criteria for ICMPv4 code.

IPv6: ICMPv6 Type (reference: [RFC8956])

Filter defines: a list of match criteria for ICMPv6 code.

IPv4/IPv6 length: variable

IPv4/IPv6 value: [numeric_op, value]+

3.1.9. TCP Flags (type = 9)

IPv4/IPv6: TCP Flags Code (reference: [RFC8955])

Filter defines: a list of match criteria for TCP Control bits

IPv4/IPv6 length: variable

IPv4/IPv6 value: [bitmask_op, value]+

Note: a 2 octets bitmask match is always used for TCP-Flags

3.1.10. Packet length (type = 10 (0x0A))

IPv4/IPv6: Packet Length (reference: [RFC8955], [RFC8956])

Filter defines: a list of match criteria for length of packet (excluding L2 header but including IP header).

IPv4/IPv6 length: variable

IPv4/IPv6 value: [numeric_op, value]+

Note:[RFC8955] uses either 1 or 2 octet values.

3.1.11. DSCP (Differentiated Services Code Point) (type = 11 (0x0B))

IPv4/IPv6: DSCP Code (reference: [RFC8955], [RFC8956])

Filter defines: a list of match criteria for DSCP code values to match the 6-bit DSCP field.

IPv4/IPv6 length: variable

IPv4/IPv6 value: [numeric_op, value]+

Note: This component uses the Numeric Operator (numeric_op) described in [RFC8955] in section 4.2.1.1. Type 11 component values MUST be encoded as single octet (numeric_op len=00).

The six least significant bits contain the DSCP value. All other bits SHOULD be treated as 0.

3.1.12. Fragment (type = 12 (0x0C))

IPv4/IPv6: Fragment (reference: [RFC8955], [RFC8956])

Filter defines: a list of match criteria for specific IP fragments.

Length: variable

Component Value format: [bitmask_op, value]+

Bitmask values are:

0	1	2	3	4	5	6	7
+	+	+	+	+	+	+	+
	0		0		0		0
	LF		FF		IsF		DF
+	+	+	+	+	+	+	+

Figure 3-4

Where:

DF (don't fragment): match If IP header flags bit 1 (DF) is 1.

IsF(is a fragment other than first: match if IP header fragment offset is not 0.

FF (First Fragment): Match if [RFC0791] IP Header Fragment offset is zero and Flags Bit-2 (MF) is 1.

LF (last Fragment): Match if [RFC7091] IP header Fragment is not 0
And Flags bit-2 (MF) is 0

0: MUST be sent in NLRI encoding as 0, and MUST be ignored during reception.

3.1.13. Flow Label(type = 13 (0x0D))

IPv4/IPv6: Fragment (reference: [RFC8956])

Filter defines: a list of match criteria for 20-bit Flow Label in the IPv6 header field.

Length: variable

Component Value format: [numeric_op, value]+

3.1.14. TTL (type=14 (0x0E))

TTL: Defines matches for 8-bit TTL field in IP header

Encoding: <[numeric_op, value]+>

where: value is a 1 octet value for TTL.

ordering: by full value of number_op concatenated with value

conflict: none

reference: draft-bergeon-flowspec-ttl-match-00.txt

3.1.15. Parts of SID (type = 15 (0xF))

IPv6: Service Identifier Matches (reference:
[I-D.ietf-idr-flowspec-srv6])

Filter defines: a list of match bit match criteria for some combinations of the LOC (location), FUNCT (function) and ARG (arguments) fields in the SID or or whole SID.

Length: variable

Component Value format: [type, LOC-Len, FUNCT-Len, ARG-Len, [op, value]+]

where:

- * type (1 octet): This indicates the new component type (TBD1, which is to be assigned by IANA).
- * LOC-Len (1 octet): This indicates the length in bits of LOC in SID.
- * FUNCT-Len (1 octet): This indicates the length in bits of FUNCT in SID.
- * ARG-Len (1 octet): This indicates the length in bits of ARG in SID.
- * [op, value]+: This contains a list of {operator, value} pairs that are used to match some parts of SID.

The total of three lengths (i.e., LOC length + FUNCT length + ARG length) MUST NOT be greater than 128. If it is greater than 128, an error occurs and it is treated as a withdrawal [RFC7606] and [RFC4760].

The operator (op) byte is encoded as:

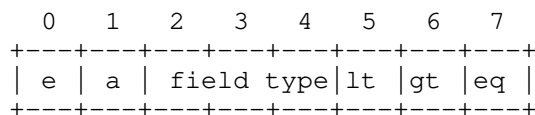


Figure 3-5

where:

where the behavior of each operator bit has clear similarity with that of [RFC8955]'s Numeric Operator field.

e (end-of-list bit): Set in the last {op, value} pair in the sequence.

a - AND bit: If unset, the previous term is logically ORed with the current one. If set, the operation is a logical AND. It should be unset in the first operator byte of a sequence. The AND operator has higher priority than OR for the purposes of evaluating logical expressions.

field type:

- 000: SID's LOC
- 001: SID's FUNCT

- 010: SID's ARG
- 011: SID's LOC:FUNCT (the concatenation of the LOC and FUNCTION fields)
- 100: SID's FUNCT:ARG (the concatenation of the FUNCTION and ARG fields)
- 101: SID's LOC:FUNCT:ARG (the concatenation of the FUNCTION and ARG fields)

Note: For an unknown field type, Error Handling is to "treat as withdrawal" [RFC7606] and [RFC4760].

lt: less than comparison between data' and value'.

gt: greater than comparison between data' and value'.

eq: equality between data' and value'.

The data' and value' used in lt, gt and eq are indicated by the field type in an operator and the value field following the operator.

The length of the value field depends on the field type and is the length of the SID parts being matched (see Table 3, Figure 3-6) in bytes, rounded up if that length is not a multiple of 8.

Table 3 - SID Parts fields

Field Type	Value
SID's LOC	value of LOC bits
SID's FUNCT	value of FUNCT bits
SID's ARG	value of ARG bits
SID's LOC:FUNCT	value of LOC:FUNCT bits
SID's FUNCT:ARG	value of FUNCT:ARG bits
SID's LOC:FUNCT:ARG	value of LOC:FUNCT:ARG bits

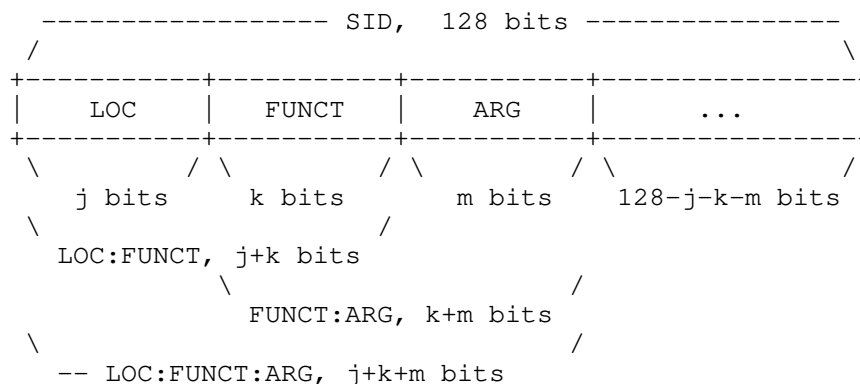


Figure 3-6

3.1.16. MPLS Label Match1 (type=16, 0x10)

Function: This match1 applies to MPLS Label field on the label stack.

reference: [I-D.ietf-idr-flowspec-mpls-match]

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

It contains a set of {operator, value} pairs that are used for the matching filter.

The operator byte is encoded as:

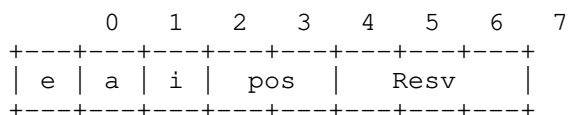


Figure 3-7

where:

e - end of list bit: Set in the last {op, value} pair in the list.

a - AND bit: If unset, the previous term is logically ORed with the current one. If set, the operation is a logical AND. It should be unset in the first operator byte of a sequence. The AND operator has higher priority than OR for the purposes of evaluating logical expressions.

i - before bit: If unset, apply matching filter before MPLS label

data plane action; if set, apply matching filter after MPLS label data plane action.

pos - the label position indication bits: whose meaning for various values is shown below:

00: any position on the label stack - the presented label value is used to match any label on the label stack. When applying it, at least one label on the stack MUST match the value

01: top label indication- the presented label value MUST be used to match the top label on the label stack.

10: bottom label indication- the presented label value MUST match the bottom label on the label stack. When it is clear, the present label value can match to any label on the label stack

11: reserved value - - This value is reserved and MUST not be sent in the packet.

The value field is encoded as:

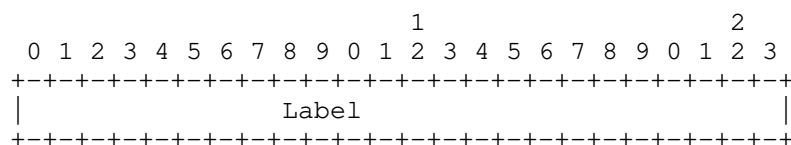


Figure 3-8

reference:

3.1.17. MPLS Label Match 2: Experimental bits match on top label (Type=17 (0x11))

Function: MPLS Match2 applies to MPLS Label experiment bits (EXP) on the top label in the label stack.

reference: [I-D.ietf-idr-flowspec-mpls-match]

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

- [op,value] - Defines a list of {operation, value} pairs used to match 3-bit exp field on the top label of packets [RFC3032].

- Values are encoded using a single byte, where the five most significant bits are zero and the three least significant bits contain the exp value.

3.2. Encoding of FSV2 Actions (type=2)

The FSv2 actions may be sent in an Extended Community or a Wide Community.

The Extended Community encodes the Flow Specification actions in the Extended Community format [RFC4360].

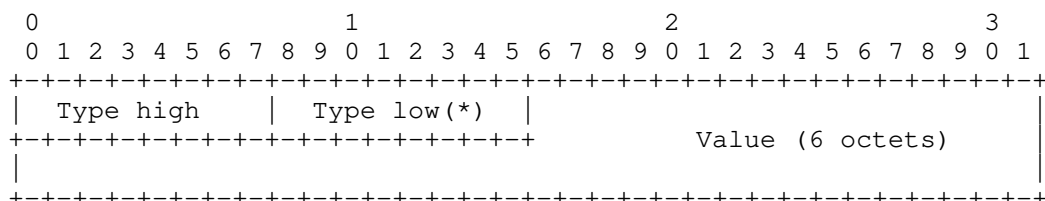


Figure 3-9

The Wide Community definition for FSv2 actions is as follows:

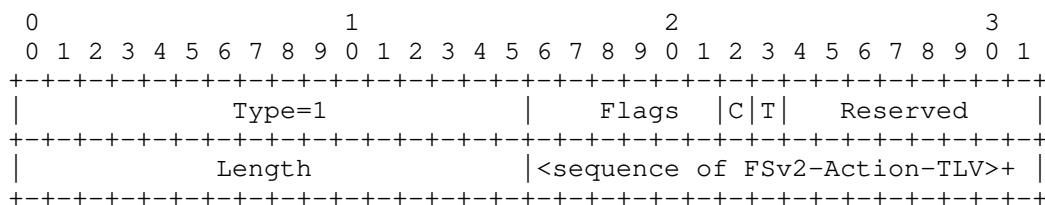


Figure 3-10

where FSv2-Action-TLV is defined as:

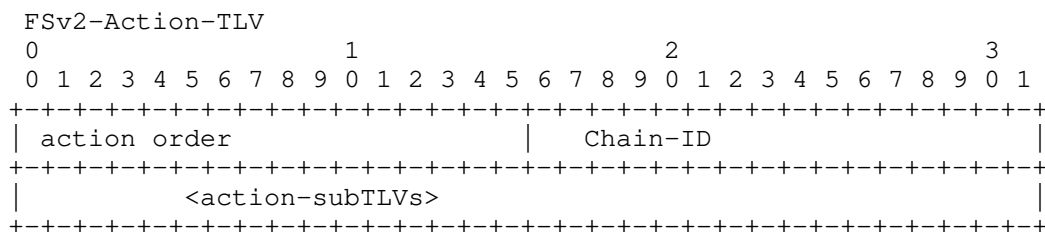


Figure 3-11

Where action-SubTLVs have the format:

action-SubTLVs

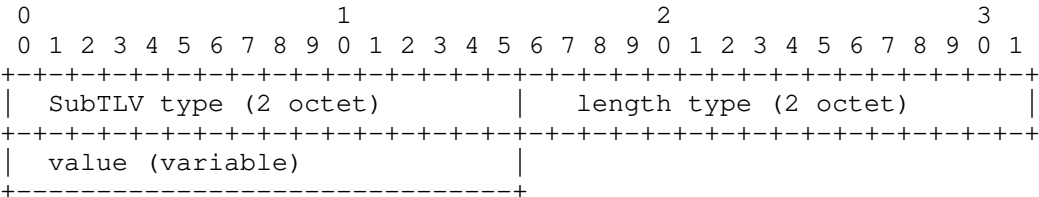


Figure 3-12

where:

- action-order: is the user defined order of the action within the list
- chain ID: is a 2-byte identifier for an action chain
- length - is the length of the TLV
- value - contains a sequence of action SubTLVs.

Each Action SubTLV has the format:

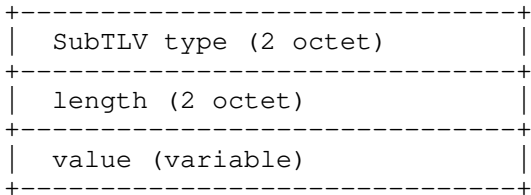


Figure 3-14

Where:

- * SubTLV type: values are action type values shown in Table 4 below.
- * length: is the length of the action SubTLV
- * Value is specific to the SubTLV

Table 4 FSv2 Action types

Action	Description
=====	=====
00	reserved
01	ACO: action chain operation
02	TAIS: traffic actions per interface group
06	TRB: traffic rate limited by bytes
07	TA: traffic action (terminal/sample)
08	RDIP: Redirect IPv4
09	TM: mark DSCP value
10	TBA (to be assigned)
11	TBA (to be assigned)
12	TRP: traffic rate limited by packets
13	RDIPv6: redirect to IPv6
14	TISFC: SFC Classifier Info (moved from OD to OE)
15	RDIID: redirect to Indirection-id (move from 0x00)
16	MPLSLA: MPLS label action
17-21	TBA (to be assigned)
22	VLAN: VLAN-Action (0x16) [draft-ietf-idr-flowspec-l2vpn-17]
23	TPID: TPID-Action (0x17) [draft-ietf-idr-flowspec-l2vpn-17]
24-254	TBA (to be assigned)
255	reserved

Figure 3-15

Ordering of actions within a rule:

The actions are first stored in user-defined order. If multiple actions exist for a single action order value, then the actions will be ordered by action type followed by value.

Action specifications must include descriptions of order comparison for the values within the action.

3.2.1. Action Chain operation (ACO) (1, 0x01)

SubTLV: 0x01

Length: variable

Value:

AC-failure-type - byte that determines the action on failure

AC-failure-value - variable depending on AC-failure-type.

Actions may succeed or fail and an Action chain must deal with it. The default value stored for an action chain that does not have this action chain is "stop on failure".

where:

AC-Failure types are:

- 0x00 - default - stop on failure
- 0x01 - continue on failure (best effort on actions)
- 0x02 - conditional stop on failure - depending on AC-Failure-value
- 0x03 - rollback - do all or nothing - depending in AC-Failure-value

AC-Failure values: TBD

Interactions with other actions: Interactions with all other Actions

Ordering within Action type: By AC-Failure type

3.2.2. Traffic Actions per interface set (TAIS) (2, 0x02)

SubTLV: 0x02

Length: 8 octets (6 in extended community)

Value field: [4-octet-AS] [GroupID 2-octet] [action 2-octet]

where:

Group-ID: identifier for group in 2 octets (14 lower bits)

- Note: Extended Community format will have 2 bits for action.

Action: determines inbound or outbound action where:

- Outbound(0x1): FSv2 rule MUST be applied in outbound Direction to interface set identified by Group-ID.
- Inbound (0x2): FSv2 rule MUST be applied in inbound Direction to interface set identified by Group-ID.

Value ordering: AS, then Group ID, then Action bytes.

Conflict: with any bi-direction action such as

1. traffic rate limited by bytes, or
2. traffic rate limited by packets.

Reference: [I-D.ietf-idr-flowspec-interfaceset]

3.2.3. Traffic rate limited by bytes (TRB) (6, 0x06)

SubTLV: 0x06

Length: 8 octets

Value field: [4-octet-AS] [float (4 bytes)]

where:

[4-octet-AS]: 4 byte AS number

- If FSV1 passes the lower 2 bytes of 4 byte AS number, use [TBD6] as higher 2 bytes to identify.
- Open issue : TBD6 can be either be chosen to match the common 2-byte to 4-byte or a unique value. Feedback is needed from WG and authors.

Float: maximum byte rate in IEEE 32-bit floating point [IEEE.754.19895 format] in bytes per second.

- A value of 0 should result in all traffic for the particular flow to be discarded.
- On encoding the traffic-rate-packets MUST NOT be negative.
- On decoding, negative values MUST BE treated as zero (discard all traffic).

Value ordering: AS then float value

Action Conflict: traffic-rate-packets

reference: [RFC8955]

3.2.4. Traffic Action (TA) (7, 0x07)

SubTLV: 0x07

Length: 1

Value field: [1-octet action]

where the traffic action values are:

- 1 = Terminal flow specification action
- 2 = Sample - enables sampling and logging
- 3 = Terminal action + sample

Value ordering: By traffic action values

Conflicts/Interactions: duplication of packets also occurs in:

- Redirect to IPv4 (action 0x08),
- Redirect to IPv6 (action 0x0D (13)),
- Redirect to SFC (action 0x0E (14))
- Redirect to Indirection-ID (action 0xF (15))

3.2.5. Redirect to IPv4 (RDIPv4) (8,0x08)

SubTLV: 0x08

Length: 12 octets

Value field:

[4-byte-AS] [IPv4 address (4 octets)] [ID (4 octets)] [Flag (1 octet)]

where:

4-octet-AS - is a 4-byte AS in a Route Target

IPv4 address - is an IP Address in RT

ID - the 4-octet value set by user

Flag is 1 octet value with the following definitions:

- 0 - reserved
- 1 - copy and redirect copy

Value ordering: 4-octet AS, then IP address, then ID (lowest to highest) with:

No AS specified uses AS value of zero.

No IP specified uses IP value of zero.

No ID specified uses ID value of zero.

Conflicts/Interactions: Any redirection or traffic sampling found in:

Traffic Action (action 0x07),

Redirect to IPv6 (action 0x0D (13)),

Redirect to SFC (action 0x0E (14))

Redirect to Indirection-ID (action 0xF (15))

reference: [RFC8955], draft-ietf-idr-flowspec-ip-02.txt

3.2.6. Traffic marking (TM) (9, 0x09)

SubTLV: 0x09

Length: 1 octet

Value: DSCP field with the 2 left most bits zero

The DSCP field format is:

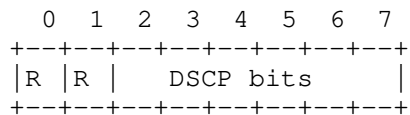


Figure 3-16

where:

R - reserved bits (set to zero to send, ignored upon reception and set to zero).

DSCP - 6 bits of DSCP values

Ordering within Value: Based on DSCP value

Conflicts: none

reference: [RFC8955]

3.2.7. Traffic rate limited by packets (TRP) (12, 0xC)

SubTLV: 12 (0xC)

Length: 8

Value field: [4-octet-AS] [float (4 octet)]

Where:

4-octet AS - is the AS setting this value

Float - specifies maximum rate in IEEE 32-bit format
[IEEE.754.185] in packets per second.

- A traffic rate of zero should result in all packets being discard.
- On encoding the traffic-rate-packets MUST NOT be negative.
- On decoding, negative values MUST BE treated as zero (discard all traffic).

Ordering within Value: Based on DSCP value

Conflicts: Traffic rate limited by bytes (0x06)

reference: [RFC8955]

3.2.8. Traffic redirect to IPv6 (RDIPv6) (13, 0xD)

SubTLV: 13 (0xD)

Length: 24 octets

Value field: [4-octet-as] [IPv6-address (16 octets)] [local administrator (2 octets)] [Flag (1 octets)]

where:

4-octet-AS - is the AS requesting action in 4-byte AS format,

IPv6-address - is the redirection IPv6 address

Local administrator - 2 bytes assigned by network administrator.

lag (1 octet) with the following definitions:

- 0 - reserved
- 1 - copy and redirect copy

Ordering within Value: AS, then IPv6, the flag (low to high)

Conflicts/Interactions: Any redirection or traffic sampling found in:

Traffic Action (action 0x07) ,

Redirect to IPv4 (action 0x08 (8)),

Redirect to SFC (action 0x0E (14))

Redirect to Indirection-ID (action 0xF (15))

3.2.9. Traffic insertion in SFC (TISFC) (14, 0xE)

SubTLV:14 (0xE)

Note: replace IANA 0xD FSv1 with FSv2 0xE.

Length: 6 octets

Value field: [SPI (3 octets)][SI (1 octet)][SFT (2 octet)]

where:

SPI - is the service path identifier

SI - is the service index

SFT - is the service function type.

Value ordering: SPI, then SI, then SFT (lowest to highest)

Conflicts/Interactions: Any redirection or traffic sampling found in:

Traffic Action (action 0x07) ,

Redirect to IPv4 (action 0x08 (8)),

Redirect to IPv6 (action 0x0D (13))

Redirect to Indirection-ID (action 0xF (15))

Reference: [RFC9015]

3.2.10. Flow Specification Redirect to Indirection-ID (RDIID) (15, 0x0F)

SubTLV: 15 (0x0F)

note: current value is 0x00 for FSv1

Length: 6 octets

Value field:

[Flags (1 octet)] [ID-Type (1 octet)] [Generalized-ID (4 octets)]

Figure 3-17

where:

Flags: are defined as:

- [S-ID]: sequence number for indirection IDs (3 bits).
 - o Value of zero means sequence is not set and all other S-ID values should be ignored
- [C] - copy packets matching this ID

ID-Type: type of indirection ID with following values:

- 0 - localized ID
- 1 - Node with SID/index in MPLS SR
- 2 - Node with SID/label in MPLS SR
- 3 - Node with Binding Segment ID with SID/Index
- 4 - Node with Binding Segment ID with SID/Label
- 5 - Tunnel ID

Generalized-ID (G-ID): indirection value

Value Ordering: first indirection ID, then Generalized ID

Action Value ordering: ID-Type by value (lowest to highest)

Conflicts/Interactions: Any redirection or traffic sampling found in:

Traffic Action (action 0x07),
 Redirect to IPv4 (action 0x08 (8)),
 Redirect to IPv6 (action 0x0D, (13))
 Redirect to SFC (action 0x0E (14))

reference: [I-D.ietf-idr-flowspec-path-redirect]

3.2.11. MPLS Label Action (MPLSLA) (16, 0x10)

Function: MPLS Label actions

SubTLV: 16 (0x10)

Length: 6 octets

Value:

[action (1 octet)]
 [order (1 octet)]
 [Label Stack Entry (4 octets)]

where Action:

Action	Function
0	Push the MPLS tag
1	Pop the outermost MPLS tag in the packet
2	Swap the MPLS tag with the outermost MPLS tag in the packet
3~15	Reserved

Figure 3-18

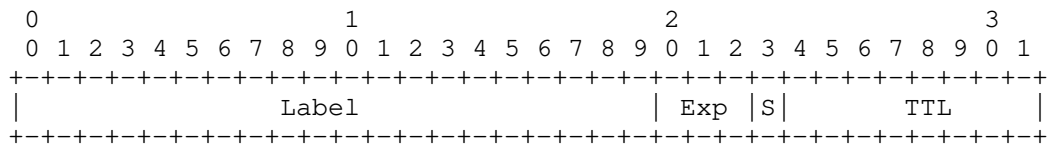


Figure 3-19 - Label Stack Entry

Action Value ordering: ID-Type, then value (lowest to highest)

Value Ordering: order, action, label, Exp

Conflicts/Interactions: Any redirection for IP before MPLS

Traffic Action (action 0x07),

Redirect to IPv4 (action 0x08 (8)),

Redirect to IPv6 (action 0x0D, (13))

Redirect to SFC (action 0x0E (14))

reference: [I-D.ietf-idr-bgp-flowspec-label]

3.2.12. VLAN action (VLAN) (22, 0x16)

Function: Rewrite inner or outer VLAN header

SubTLV: 22 (0x16)

Length: 6 octets

Value:

[Rewrite-actions (2 octets)]

[vlan-PCP-DE-1 (2 octets)]

[vlan-PCP-DE-2 (2 octets)]

where:

Rewrite-actions - is as follows:

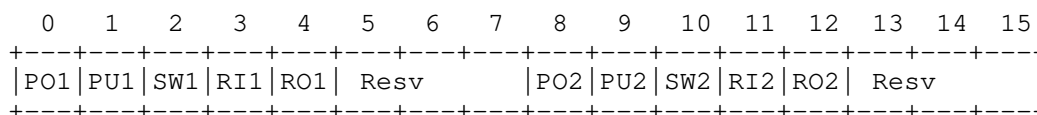


Figure 3-20

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

PU1: Push action. If PU1 is one, it indicates VLAN ID1 will be added, the associated Priority Code Point (PCP and Drop Eligibility Indicator (DEI) are PCP1 and DEI1.

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

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

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

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

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

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

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

Value ordering: rewrite-actions, VLAN1, VLAN2, PCP-DE1, PCP-DE2

Conflicts: TIPD Action

reference: [I-D.ietf-idr-flowspec-l2vpn]

3.2.13. TPID action (TPID) (23, 0x17)

Function: Replace Inner or outer TP

SubTLV: 23 (0x17)

Length: 6 octets

Value:

[Rewrite-actions (2 octets)]

[TP-ID-1 (2 octets)]

[TP-ID-2 (2 octets)]

Where: rewrite-actions are bitmask (2 octets) with 2 actions as follows:

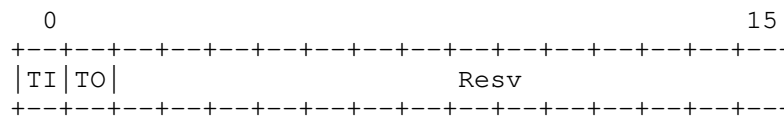


Figure 3-21

TI: Mapping inner Tag Protocol (TP) ID (typically a VLAN) action. If the TI flag is one, it indicates the inner TP ID should be replaced by a new TP ID, the new TP ID is TP ID1.

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

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

Value Ordering: rewrite-actions, TP-ID-1, TP-ID-2

Conflicts: VLAN action

reference:[I-D.ietf-idr-flowspec-l2vpn]

3.3. Extended Community vs. Action SubTLV formats

The SubTLV format is used for the Wide communities and for the action subTLVs in the NLRI.

Sub-TLV type	Action Name	Action format	SubTLV	Extended Community format
=====	=====	=====		=====
1	ACO	type: 1 (0x01) length:variable		not applicable (n/a)
2	TAIS	type: 2 (0x02) length:8 [4-octet-as] [group-3-octet] [flags-1-octet]		type: 0x0702 or 0x4702 length: 6 [4-octet-AS] [flags-group] (2 octets)
3-5	reserved			

Sub-TLV type	Action Name	Action format	SubTLV	Extended Community format
=====	=====	=====		=====
6	TRB	type:6 (0x06) length:8 [4-byte-AS] [float (4 octets)]		type:8006 length: 6 octets [2-byte-AS] [float (4 octets)]
7	TA	type:7 length:1 flags: (1 octet)		type:8007 length:6 octets flags (6 octets)
8	RDIPv4	type:8 length: 12 [4-byte-AS] [IPv4-address]		type:8008 length: 6 octets [AS-2-octets] [IPv4 address] type:8108 length: 6 octets [IPv4 address] [ID-2 octets] type:8208 length: 6 octets [AS-4-octets] [ID-2-octets]
9	TM	type:9 length:1 DSCP: 1 octet		type:8009 length: 6 octets DSCP: 1 octet

10		type:10 (0X0A)	TBA
11		type:11 (0x0B)	TBA
12	TRP	type:12 (0x0C) length: 8 octets [4-byte-AS] [float-4-octet]	type: 0x800C length: 6 octets [2-byte-AS] [float-4-octet]
13	RDIPv6	type:13 (0x0D) length:22 [4-byte-AS] [IPv6-address (16)] [local-admin (2)]	type:0x000C length: 18 octets [IPv6-address (16)] [local-admin (2)]

Sub-TLV type =====	Action Name =====	Action format =====	SubTLV	Extended Community format =====
14	TISFC	type:14 (0x0E) length:6 SPI (3 octets) SI (1 octet) SFT (2 octets)		type: 0xD (FSv1) type: 0xE (FSv2) length:6 SPI (3 octets) SI (1 octet) SFT (2 octets)
15	RDIID	type:15 (0x0F) length: 6 flags (1) ID-type (1) G-ID (4 octets)		type: 0900 (FSv1) length 6 flags (1) ID type (1) G-ID (4-octets)
16	MPLSLA	type:16 (0x10)		
16-21	TBA -			
22	VLAN	type:22 (0x16) length:6 [rewrite-action(2)] [vlan-pcp-de-1 (2)] [vlan-pcp-de-2 (2)]		Type: (TBD) length:6 [rewrite-actions (2)] [vlan-pcp-de-1 (2)] [vlan-pcp-de-2 (2)]
23	TPID	type:23 (0x17) length:6		Type: (TBD) length:6


```

[rewrite-action(2)] [rewrite-actions (2)]
[TP-ID-1 (2)]      [TP-ID-1 (2)]
[TP-ID-2 (2)]      [TP-ID-2 (2)]

```

3.4. L2 Traffic Rules

The format of the L2 header TLV value field is shown in Figure 3-22. The AFI/SAFI field includes the AFI (2 octets), SAFI (1 octet).

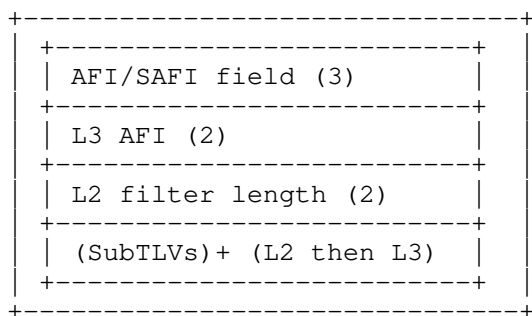


Figure 3-22 -L2 Header TLV value

Where:

AFI/SAFI field has AFI is 6 (IEEE 802) and SAFI is TBD1.

L3 AFI is zero if the filter test only L2 fields, otherwise it is or 2 depending on whether the filter L3 tests after the L2 header are for IPv4 or IPv6.

L2 filter length is the length of the L2 SubTLVs in bytes. These are followed by the L3 SubTLVs is the L3 AFI field is non-zero.

Each L2 SubTLV has the format shown in Figure 3-23. (The L3 SubTLVs are as defined in Section 4.1.)

Each SubTLV has the format:

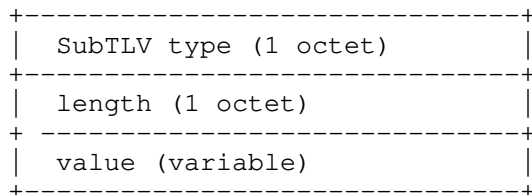


Figure 3-23

SubTLV type: A component type value defined in the "L2 Flow Specification Component Types" registry for L2 by [draft-ietf-idr-flowspec-l2vpn].

Where the SubTLVs have the following component types:

Component Types Table

Component type	Description
=====	=====
1	EtherType
2	Source MAC
3	Destination MAC
4	DSAP (destination service access point)
5	SSAP (source service access point)
6	control field in LLC
7	SNAP
8	VLAN ID
9	VPAN PCP
10	Inner VLAN ID
11	Inner VLAN PCP
12	VLAN DEI
13	VLAN DEI
14	Source MAC special bits
15	Destination MAC special bits

Table 4 L2 VPN components

See [I-D.ietf-idr-flowspec-l2vpn] for the details on the format and value fields for each component.

Value ordering: Ordering of L2 FSv2 rules will be by user-defined order of the rule. For FSv2 filters within the same rule, the ordering will be by component number and then by value within the component. See [I-D.ietf-idr-flowspec-l2vpn] for the ordering of the values within the component.

L2 VPN filtering using SAFI TBD2 is specified in section 3.6.

reference: [I-D.ietf-idr-flowspec-l2vpn]

3.5. SFC Traffic Rules

The FSv2 filters allow for filtering of the SFC NLRI family of routes. The traffic NLRIs filtered are from SFC AFI/SAFI (AFI = 31, SAFI=9).

The FSv2 filters provide this filtering with SFC AFI (AFI=31) and SAFI for FSv2 filters (SAFI = TB1).

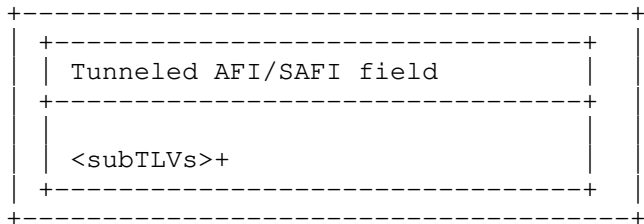


Figure 3-24

Each SubTLV has the format:

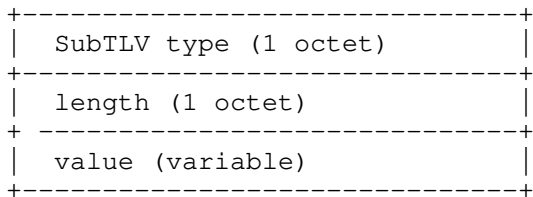


Figure 3-25 Tunneled SubTLV format

The components listed are:

- 1 = SFIR RD Type (types 1, 2, 3)
- 2 = SFIR RD Value
- 3 = SFIR Pool ID
- 4 = SFIR MPLS context/label
- 5 = SFPR SPI
- 6 = SPF attribute fields

Table 6 SFC Filter types

Ordering is by: User-defined rule order, component number, and then value within component.

reference: [RFC9015], [TBD]

3.6. BGP/MPLS VPN IP Traffic Rules

The format of the match filter for BGP/MPLS VPN IP traffic is very similar to the format for non-VPN IP traffic as defined in Section 3.1 except that the SAFI is TBD2 and the initial NLRI header has an 8-byte Route Distinguisher added to it as shown in Figure 3-26. The SubTLV format and filter components formats remain the same.

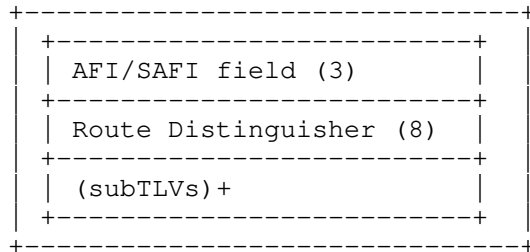


Figure 3-26: VPN IP Filter Header

3.7. BGP/MPLS VPN L2 Traffic Rules

The format of the match filter for BGP/MPLS VPN IP traffic is very similar to the format for non-VPN L2 traffic as defined in Section 3.4 except that the SAFI is TBD2 and the initial NLRI header has an 8-byte Route Distinguisher added to it right after the AFI/SAFI as shown in Figure 3-27. The SubTLV format and filter components formats remain the same.

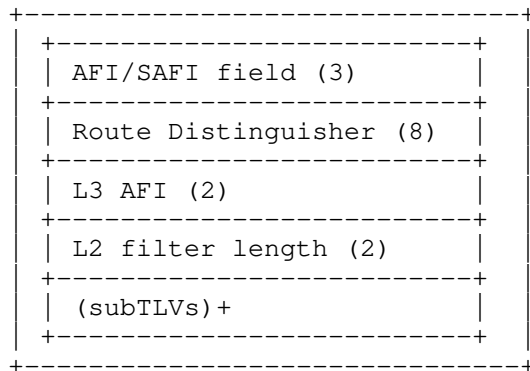


Figure 3-27: VPN L2 Filter Header

3.8. Encoding of Actions passed in Wide Communities

The BGP FSv2 actions are passed in a Wide Community attribute with a BGP Wide Community container (type 01) [I-D.ietf-idr-wide-bgp-communities] with community of FSv2 Actions (TBD4) and Wide Community attributes of Target TLV, Exclude TLVs, and Parameter TLVs. The Parameter MUST contain an FSv2 Atom which contains a sequence of Action TLVs.

BGP Wide Community Container
with FSv2 actions

Community: FSv2-actions (community value = TBD4)
Source AS number
Context AS number
Target or Exclude TLVs (optional)
Parameter TLV with FSv2 atom

figure 3-28 - BGP

FSv2 Actions atom-id
length (2 octets)
<Action-Sub-TLVs>+

Figure 3-29 - Flow Specification
with IDs for Wide Community Actions

where:

Atom-id: (TBD5)

length: variable depending on SubTLVS

Action Sub-TLVs as defined above

4. Validation of FSv2 NLRI

The validation of FSv2 NLRI adheres to the combination of rules for general BGP FSv1 NLRI found in [RFC8955], [RFC8956], [RFC9117], and the specific additions made for SFC NLRI [RFC9015], and L2VPN NLRI [I-D.ietf-idr-flowspec-l2vpn].

To provide clarity, the full validation process for flow specification routes (FSv1 or FSv2) is described in this section rather than simply referring to the relevant portions of these RFCs. Validation only occurs after BGP UPDATE message reception and the FSv2 NLRI and the path attributes relating to FSv2 (Extended community and Wide Community) have been determined to be well-formed. Any MALFORMED FSv2 NLRI is handled as a "TREAT as WITHDRAW" [RFC7606].

4.1. Validation of FS NLRI (FSv1 or FSv2)

Flow specifications received from a BGP peer that are accepted in the respective Adj-RIB-In are used as input to the route selection process. Although the forwarding attributes of the two routes for the same prefix may be the same, BGP is still required to perform its path selection algorithm in order to select the correct set of attributes to advertise.

The first step of the BGP Route selection procedure (section 9.1.2 of [RFC4271]) is to exclude from the selection procedure routes that are considered unfeasible. In the context of IP routing information, this is used to validate that the NEXT_HOP Attribute of a given route is resolvable.

The concept can be extended in the case of the Flow Specification NLRI to allow other validation procedures.

The FSv2 validation process validates the FSv2 NLRI with following unicast routes received over the same AFI (1 or 2) but different SAFIs:

- * Flow specification routes (FSv1 or FSv2) received over SAFI=133 will be validated against SAFI=1,
- * Flow Specification routes (FSv1 or FSv2) received over SAFI=134 will be validated against SAFI=128, and
- * Flow Specification routes (FSv1 or FSv2) [AFI =1, 2] received over SAFI=77 will be validated using only the Outer Flow Spec against SAFI = 133.

The FSv2 validates L2 FSv2 NLRI with the following L2 routes received over the same AFI (25), but a different SAFI:

- * Flow specification routes (FSv1 or FSv2) received over SAFI=135 are validated against SAFI=128.

In the absence of explicit configuration, a Flow specification NLRI (FSv1 or FSv2) MUST be validated such that it is considered feasible if and only if all of the conditions are true:

- a) A destination prefix component is embedded in the Flow Specification,
- b) One of the following conditions holds true:
 - 1. The originator of the Flow Specification matches the originator of the best-match unicast route for the destination prefix embedded in the flow specification (this is the unicast route with the longest possible prefix length covering the destination prefix embedded in the flow specification).
 - 2. The AS_PATH attribute of the flow specification is empty or contains only an AS_CONFED_SEQUENCE segment [RFC5065].
 - o 2a. This condition should be enabled by default.
 - o 2b. This condition may be disabled by explicit configuration on a BGP Speaker,
 - o 2c. As an extension to this rule, a given non-empty AS_PATH (besides AS_CONFED_SEQUENCE segments) MAY be permitted by policy].
- c) There are no "more-specific" unicast routes when compared with the flow destination prefix that have been received from a different neighbor AS than the best-match unicast route, which has been determined in rule b.

However, part of rule a may be relaxed by explicit configuration, permitting Flow Specifications that include no destination prefix component. If such is the case, rules b and c are moot and MUST be disregarded.

By "originator" of a BGP route, we mean either the address of the originator in the ORIGINATOR_ID Attribute [RFC4456] or the source address of the BGP peer, if this path attribute is not present.

A BGP implementation MUST enforce that the AS in the left-most position of the AS_PATH attribute of a Flow Specification Route (FSv1 or FSv2) received via the Exterior Border Gateway Protocol (eBGP) matches the AS in the left-most position of the AS_PATH attribute of the best-match unicast route for the destination prefix embedded in the Flow Specification (FSv1 or FSv2) NLRI.

The best-match unicast route may change over time independently of the Flow Specification NLRI (FSv1 or FSv2). Therefore, a revalidation of the Flow Specification MUST be performed whenever unicast routes change. Revalidation is defined as retesting rules against as described above.

4.2. Validation of Flow Specification Actions

Flow Specifications may be mapped to actions using Extended Communities or a Wide Communities. The FSv2 actions in Extended Communities and Wide communities can be associated with large number of NLRIs.

The ordering of precedence for these actions in the case when the user-defined order is the same follows the precedence of the FSv2 NLRI action TLV values (lowest to highest). User-defined order is the same when the order value for action is the same. All Extended Community actions MUST be translated to the user-defined order data format for internal comparison. By default, all Extended Community actions SHOULD be translated to a single value.

Actions may conflict, duplicate, or complement other actions. An example of conflict is the packet rate limiting by byte and by packet. An example of a duplicate is the request to copy or sample a packet under one of the redirect functions (RDIPv4, RDIPv6, RDIID,) Each FSv2 actions in this document defines the potential conflicts or duplications. Specifications for new FSv2 actions outside of this specification MUST specify interactions or conflicts with any FSv2 actions (that appear in this specification or subsequent specifications).

Well-formed syntactically correct actions should be linked to a filtering rule in the order the actions should be taken. If one action in the ordered list fails, the default procedure is for the action process for this rule to stop and flag the error via system management. By explicit configuration, the action processing may continue after errors.

Implementations MAY wish to log the actions taken by FS actions (FSv1 or FSv2).

4.3. Error handling and Validation

The following two error handling rules must be followed by all BGP speakers which support FSv2:

- * FSv2 NLRI having TLVs which do not have the correct lengths or syntax must be considered MALFORMED.
- * FSv2 NLRIs having TLVs which do not follow the above ordering rules described in section 4.1 MUST be considered as malformed by a BGP FSv2 propagator.

The above two rules prevent any ambiguity that arises from the multiple copies of the same NLRI from multiple BGP FSv2 propagators.

A BGP implementation SHOULD treat such malformed NLRIs as 'Treat-as-withdraw' [RFC7606]

An implementation for a BGP speaker supporting both FSv1 and FSv2 MUST support the error handling for both FSv1 and FSv2.

5. Ordering for Flow Specification v2 (FSv2)

Flow Specification v2 allows the user to order flow specification rules and the actions associated with a rule. Each FSv2 rule has one or more match conditions and one or more actions associated with that match condition.

This section describes how to order FSv2 filters received from a peer prior to transmission to another peer. The same ordering should be used for the ordering of forwarding filtering installed based on only FSv2 filters.

Section 7.0 describes how a BGP peer that supports FSv1 and FSv2 should order the flow specification filters during the installation of these flow specification filters into FIBs or firewall engines in routers.

The BGP distribution of FSv1 NLRI and FSv2 NLRI and their associated path attributes for actions (Wide Communities and Extended Communities) is "ships-in-the-night" forwarding of different AFI/SAFI information. This recommended ordering provides for deterministic ordering of filters sent by the BGP distribution.

5.1. Ordering of FSv2 NLRI Filters

The basic principles regarding ordering of rules are simple:

- 1) Rule-0 (zero) is defined to be 0/0 with the "permit-all" action
 - BGP peers which do not support flow specification permit traffic for routes received. Rule-0 is defined to be "permit-all" for 0/0 which is the normal case for filtering for routes received by BGP.
 - By configuration option, the "permit-all" may be set to "deny-all" if traffic rules on routers used as BGP must have a "route" AND a firewall filter to allow traffic flow.
- 2) FSv2 rules are ordered based on the user-defined order numbers specified in the FSv2 NLRI (rules 1-n).
- 3) If multiple FSv2 NLRI have the same user-defined order, then the filters are ordered by type of FSv2 NLRI filters (see Table 1, section 4) with lowest numerical number have the best precedence.
 - For the same user-defined order and the same value for the FSv2 filters type, then the filters are ordered by FSv2 the component type for that FSv2 filter type (see Tables 3-6) with the lowest number having the best precedence.
 - For the same user-defined order, the same value of FSv2 Filter Type, and the same value for the component type, then the filters are ordered by value within the component type. Each component type defines value ordering.
 - For component types inherited from the FSv1 component types, there are the following two types of comparisons:
 - o FSv1 component value comparison for the IP prefix values, compares the length of the two prefixes. If the length is different, the longer prefix has precedence. If the length is the same, the lower IP number has precedence.
 - o For all other FSv1 component types, unless specified, the component data is compared using the memcmp() function defined by [ISO_IEC_9899]. For strings with the same length, the lowest string memcmp() value has precedence. For strings of different lengths, the common prefix is compared. If the common string prefix is not equal, then the string with the lowest string prefix has higher precedence. If the common prefix is equal, the longest string is considered to have higher precedence

Notes:

- * Since the user can define rules that re-order these value comparisons, this order is arbitrary and set to provide a deterministic default.

5.2. Ordering of the Actions

The FSv2 specification allows for actions to be associated by:

- a) a Wide Community path attribute, or
- b) an Extended Community path attribute.

Actions may be ordered by user-defined action order number from 1-n (where n is $2^{16}-2$ and the value $2^{16}-1$ is reserved).

By default, extended community actions are associated with default order number 32768 [0x8000] or a specific configured value for the FSv2 domain.

Action user-order number zero is defined to have an Action type of "Set Action Chain operation" (ACO) (value 0x01) that defines the default action chain process. For details on "set action chain operation" see section 3.2.1 or section 5.2.1 below.

If the user-defined action number for two actions are the same, then the actions are ordered by FSv2 action types (see Table 3 for a list of action types). If the user-defined action number and the FSv2 action types are the same, then the order must be defined by the FSv2 action.

5.2.1. Action Chain Operation (ACO)

The "Action Chain Operation" (ACO) changes the way the actions after the current action in an action chain are handled after a failure. If no action chain operations are set, then the default action of "stop upon failure" (value 0x00) will be used for the chain.

5.2.1.1. Example 1 - Default ACO

Use Case 1: Rate limit to 600 packets per second

Description: The provider will support 600 packets per second All Packets sampled for reporting purposes and packet streams over 600 packets per second will be dropped.

Suppose BGP Peer A has a

- * a Wide Community action with user-defined order 10 with Traffic Sampling
- * a Wide Community action with user-defined order 11 from AS 2020 that limits packet-based rate limit of 600 packets per second.
- * an Extended Community from AS 2020 that does limits packet-based rate limit of 50 packets per second.

The FSV2 data base would store the following action chain:

- * at user-defined action order 10
 - A user action of type 7 (traffic action) with values of Sampling and logging.
- * at user-defined action order 11
 - a user action type of 12 (packet-based rate limit) with values of AS 2020 and float value for 600 packets per second (pps)
- * at user-defined action order 32768 (0x8000) with type 12 and values of A user action of type 12 with values of AS 2020 and float value of 50 packets/second.

Normal action:

The match on the traffic would cause a sample of the traffic (probably with packet rate saved in logging) followed by a rate limit to 600 pps. The Extended community action would further limit the rate to 50 packets per second.

When does the action chain stop?

The default process for the action chain is to stop on failure. If there is no failure, then all three actions would occur. This is probably not what the user wants.

If there is failure at action 10 (sample and log), then there would be no rate limiting per packet (actions 11 and action 32768).

If there is failure at action 11 (rate limit to packet 600), then there would be no rate limiting per packet (action 32768).

The different options for Action chain ordering (ACO) have been worked on with NETCONF/RESTCONF configuration and actions.

5.2.1.2. Example 2: Redirect traffic over limit to processing via SFC

Use case 2: Redirect traffic over limit to processing via SFC.

Description: The normal function is for traffic over the limit to be forwarded for offline processing and reporting to a customer.

Suppose we have the following 4 actions defined for a match:

- * Sent Redirect to indirection ID (0x01) with user-defined match 2 attached in wide community,
- * Traffic rate limit by bytes (0x07) with user-defined match 1 attached in wide community,
- * Traffic sample (0x07) sent in extended community, and
- * SF classifier Info (0x0E) sent in extended community.

These 4 filters rate limit a potential DDoS attack by: a) redirect the packet to indirection ID (for slower speed processing), sample to local hardware, and forward the attack traffic via a SFC to a data collection box.

The FSV2 action list for the match would look like this

Action 0: Operation of action chain (0x01) (stop upon failure)

Action 1: Traffic Rate limit by byte (0x07)

Action 2: Redirect to Redirection ID (0x0F)

Action 32768 (0x8000) Traffic Action (0x07) Sample

Action 32768 (0x8000) SFC Classifier: (0xE)

If the redirect to a redirection ID fails, then Traffic Sample and sending the data to an SFC classifier for forwarding via SFC will not happen. The traffic is limited, but not redirect away from the network and a sample sent to DDOS processing via a SFC classifier.

Suppose the following 5 actions were defined for a FSV2 filter:

- * Set Action Chain Operation (ACO) (0x01) to continue on failure (0x01) at user-order 2 attached in wide community,
- * redirect to indirection ID (0x0F) at user-order 2 attached in wide community,

- * traffic rate limit by bytes (0x07) with user-order 1 attached in wide community,
- * Traffic sample (0x07) attached via extended community, and
- * SFC classifier Info (0x0E) attached in extended community.

The FSv2 action list for the match would look like this:

Action 00: Operation of action chain (0x01) (stop upon failure)

Action 01: Traffic Rate limit by byte (0x07)

Action 02: Set Action Chain Operation (ACO) (0x01) (continue on failure)

Action 02: Redirect to Redirection ID (0F)

Action 32768 (0x8000): Traffic Action (0x07) Sample

Action 32768 (0x8000): SFC classifier (0x0E) forward via SFC [to DDoS classifier]

If the redirect to a redirection ID fails, the action chain will continue on to sample the data and enact SFC classifier actions.

5.2.2. Summary of FSv2 ordering

Operators should use user-defined ordering to clearly specify the actions desired upon a match. The FSv2 actions default ordering is specified to provide deterministic order for actions which have the same user-defined order and same type.

FS Action (lowest value to highest)	Value Order (lowest to highest)
=====	=====
0x01: ACO: Action chain operation	Failure flag
0x02: TAIS: Traffic actions per Interface group	AS, then Group-ID, then Action ID
0x03-0x05 to be assigned	TBD
0x06: TRB: Traffic rate limit by bytes	AS, then float value
0x07: TA: Traffic Action	traffic action value
0x08: RDIP: Redirect to IP	AS, then IP Address, then ID
0x09: TM: Traffic Marking	DSCP value (lowest to highest)
0x0A: AL2: Associated L2 Info.	TBD
0x0B: AET: Associated E-tree Info.	TBD
0x0C: TRP: Traffic Rate limit by bytes	AS, then float value
0x0D: RDIPv6: Traffic Redirect to IPv6	AS, IPv6 value, then local Admin
0x0E: TISFC: Traffic insertion to SFC	SPI, then SI, the SFT
0x0F: Redirect to Indirection-ID	ID-type, then Generalized-ID
0x10: MPLSLA: MPLS Label stack	order, action, label, Exp
0x16 VLAN action	rewrite-actions, VALN1, VLAN2, PCP-DE1, PCP-DE2
0x17 TPID action	rewrite actions, TP-ID-1, TP-ID-2

Figure 6-1

6. Ordering of FS filters for BGP Peers support FSv1 and FSv2

FSv2 allows the user to order flow specification rules and the actions associated with a rule. Each FSv2 rule has one or more match conditions and one or more actions associated with each rule.

FSv1 and FSv2 filters are sent as different AFI/SAFI pairs so FSv1 and FSv2 operate as ships-in-the-night. Some BGP peers in an AS may support both FSv1 and FSv2. Other BGP peers may support FSv1 or FSv2. Some BGP will not support FSv1 or FSv2. A coherent flow specification technology must have consistent best practices for ordering the FSv1 and FSv2 filter rules.

One simple rule captures the best practice: Order the FSv1 filters after the FSv2 filter by placing the FSv1 filters after the FSv2 filters.

To operationally make this work, all flow specification filters should be included the same data base with the FSv1 filters being assigned a user- defined order beyond the normal size of FSv2 user-ordered values. A few examples, may help to illustrate this best practice.

Example 1: User ordered numbering - Suppose you might have 1,000 rules for the FSv2 filters. Assign all the FSv1 user defined rules to 1,001 (or better yet 2,000). The FSv1 rules will be ordered by the components and component values.

Example 2: Storage of actions - All FSv1 actions are defined ordered actions in FSv2. Translate your FSv1 actions into FSv2 ordered actions for storing in a common FSv1-FSv2 flow specification data base.

Example 3: Mixed Flow Specification Support -

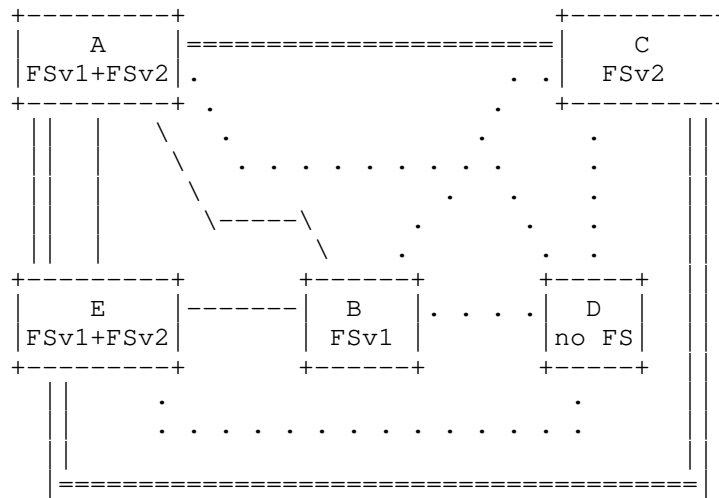
Suppose an FSv2 peer (BGP Peer A) has the capability to send either FSv1 or FSv2. BGP Peer A peers with BGP Peers B, C, D and E.

BGP Peer B can only send FSv1 routes (NLRI + Extended Community). BGP Peer C can send FSv2 routes (NLRI + path attributes (wide community or extended community or none)). BGP Peer D cannot send any FS routes. BGP E can send FSv2 and FSv1 routes

BGP Peer A sends FSv1 routes in its databases to BGP B. Since the FSv2 NLRI cannot be sent to the FSv1 peer, only the FSv1 NLRI is sent. BGP Peer A sends to BGP C the FSv2 routes in its database (configured or received).

BGP peer A would not send the FSv1 NLRI or FSv2 NLRI to BGP Peer D. The BGP Peer D does not support for these NLRI.

BGP Peer A sends the NLRI for both FSv1 and FSv2 to BGP Peer E.



Double line = FSv2

Single line = FSv1

Dotted line = BGP peering with no FlowSpec

Figure 6-2: FSv1 and FSv2 Peering

7. Scalability and Aspirations for FSv2

Operational issues drive the deployment of BGP flow specification as a quick and scalable way to distribute filters. The early operations accepted the fact validation of the distribution of filter needed to be done outside of the BGP distribution mechanism. Other mechanisms (NETCONF/RESTCONF or PCEP) have reply-request protocols.

These features within BGP have not changed. BGP still does not have an action-reply feature.

NETCONF/RESTCONF latest enhancements provide action/response features which scale. The combination of a quick distribution of filters via BGP and a long-term action in NETCONF/RESTCONF that ask for reporting of the installation of FSv2 filters may provide the best scalability.

The combination of NETCONF/RESTCONF network management protocols and BGP focuses each protocol on the strengths of scalability.

FSv2 will be deployed in webs of BGP peers which have some BGP peers passing FSv1, some BGP peers passing FSv2, some BGP peers passing FSv1 and FSv2, and some BGP peers not passing any routes.

The TLV encoding and deterministic behaviors of FSv2 will not deprecate the need for careful design of the distribution of flow specification filters in this mixed environment. The needs of networks for flow specification are different depending on the network topology and the deployment technology for BGP peers sending flow specification.

Suppose we have a centralized RR connected to DDoS processing sending out flow specification to a second tier of RR who distribute the information to targeted nodes. This type of distribution has one set of needs for FSv2 and the transition from FSv1 to FSv2

Suppose we have Data Center with a 3-tier backbone trying to distribute DDoS or other filters from the spine to combinational nodes, to the leaf BGP nodes. The BGP peers may use RR or normal BGP distribution. This deployment has another set of needs for FSv2 and the transition from FSv1 to FSv2.

Suppose we have a corporate network with a few AS sending DDoS filters using basic BGP from a variety of sites. Perhaps the corporate network will be satisfied with FSv1 for a long time.

These examples are given to indicate that BGP FSv2, like so many BGP protocols, needs to be carefully tuned to aid the mitigation services within the network. This protocol suite starts the migration toward better tools using FSv2, but it does not end it. With FSv2 TLVs and deterministic actions, new operational mechanisms can start to be understood and utilized.

This FSv2 specification is merely the start of a revolution of work - not the end.

8. Optional Security Additions

This section discusses the optional BGP Security additions for BGP-FS v2 relating to BGPSEC [RFC8205] and ROA [RFC6482].

8.1. BGP FSv2 and BGPSEC

Flow specification v1 ([RFC8955] and [RFC8956]) do not comment on how BGP Flow specifications to be passed BGPSEC [RFC8205] BGP Flow Specification v2 can be passed in BGPSEC, but it is not required.

FSv1 and FSv2 may be sent via BGPSEC.

8.2. BGP FSv2 with ROA

BGP FSv2 can utilize ROAs in the validation. If BGP FSv2 is used with BGPSEC and ROA, the first thing is to validate the route within BGPSEC and second to utilize BGP ROA to validate the route origin.

The BGP-FS peers using both ROA and BGP-FS validation determine that a BGP Flow specification is valid if and only if one of the following cases:

- * If the BGP Flow Specification NLRI has a IPv4 or IPv6 address in destination address match filter and the following is true:
 - A BGP ROA has been received to validate the originator, and
 - The route is the best-match unicast route for the destination prefix embedded in the match filter; or
- * If a BGP ROA has not been received that matches the IPv4 or IPv6 destination address in the destination filter, the match filter must abide by the [RFC8955] and [RFC8956] validation rules as follows:
 - The originator match of the flow specification matches the originator of the best-match unicast route for the destination prefix filter embedded in the flow specification", and
 - No more specific unicast routes exist when compared with the flow destination prefix that have been received from a different neighboring AS than the best-match unicast route, which has been determined in step A.

The best match is defined to be the longest-match NLRI with the highest preference.

9. IANA Considerations

This section complies with [RFC7153].

9.1. Flow Specification V2 SAFIs

IANA is requested to assign two SAFI Values in the registry at <https://www.iana.org/assignments/safi-namespace> from the Standard Action Range as follows:

Value	Description	Reference
TBD1	BGP FSv2	[this document]
TBD2	BGP FSv2 VPN	[this document]

9.2. BGP Capability Code

IANA is requested to assign a Capability Code from the registry at <https://www.iana.org/assignments/capability-codes/> from the IETF Review range as follows:

Value	Description	Reference	Controller
TBD3	Flow Specification V2	[this document]	IETF

9.3. Filter IP Component types

IANA is requested to indicate [this draft] as a reference on the following assignments in the Flow Specification Component Types Registry:

Value	Description	Reference
1	Destination filter	[RFC8955] [RFC8956] [this document]
2	Source Prefix	[RFC8955] [RFC8956] [this document]
3	IP Protocol	[RFC8955] [RFC8956] [this document]
4	Port	[RFC8955] [RFC8956] [this document]
5	Destination Port	[RFC8955] [RFC8956] [this document]
6	Source Port	[RFC8955] [RFC8956] [this document]
7	ICMP Type [v4 or v6]	[RFC8955] [RFC8956] [this document]
8	ICMP Code [v4 or v6]	[RFC8955] [RFC8956] [this document]
9	TCP Flags [v4]	[RFC8955] [RFC8956] [this document]
10	Packet Length	[RFC8955] [RFC8956] [this document]
11	DSCP marking	[RFC8955] [RFC8956] [this document]
12	Fragment	[RFC8955] [RFC8956] [this document]
13	Flow Label	[RFC8956] [this document]
14	TTL	[this document]
15	Partial SID	[draft-ietf-idr-flowspec-srv6] [this document]
16	MPLS Label Match 1	[this document] [draft-ietf-idr-flowspec-mpls-match]
17	MPLS Label Match 2	[this document] [draft-ietf-idr-flowspec-mpls-match]

9.4. FSV2 NLRI TLV Types

IANA is requested to create the following two new registries on a new "Flow Specification v2 TLV Types" web page.

Name: BGP FSv2 TLV types

Reference: [this document]

Registration Procedures: 0x01-0x3FFF Standards Action.

Type	Use	Reference
-----	-----	-----
0x00	Reserved	[this document]
0x01	IP traffic rules	[this document]
0x02	FSv2 Actions	[this document]
0x03	L2 traffic rules	[this document]
0x04	tunnel traffic rules	[this document]
0x05	SFC AFI filter rules	[this document]
0x06	BGP/MPLS VPN IP traffic rules	[this document]
0x07	BGP/MPLS VPN L2 traffic rules	[this document]
0x08-0x3FFF	Unassigned	[this document]
0x4000-0x7FFF	Vendor specific	[this document]
0x8000-0xFFFF	Reserved	[this document]

Name: BGP FSv2 Action types

Reference: [this document]

Registration Procedure: 0x01-0x3FFF Standards Action.

Type	Use	Reference
0x00	Reserved	[this document]
0x01	ACO: Action Chain Operation	[this document]
0x02	TAIS: Traffic actions per interface group	[this document]
0x03	Unassigned	[this document]
0x04	Unassigned	[this document]
0x05	Unassigned	[this document]
0x06	TRB: traffic rate limited by bytes	[this document]
0x07	TA: Traffic action (terminal/sample)	[this document]
0x08	RDIPv4: redirect IPv4	[this document]
0x09	TM: traffic marking (DSCP)	[this document]
0x0A	AL2: associate L2 Information	[this document]
0x0B	AET: associate E-Tree information	[this document]
0x0C	TRP: traffic rate limited by packets	[this document]
0x0D	RDIPv6: Redirect to IPv6	[this document]
0x0E	TISFC: Traffic insertion to SFC	[this document]
0x0F	RDIID: Redirect to indirection-iD	[this document]
0x10	MPLS Label Action	[this document]
0x11	unassigned	[this document]
0x12	unassigned	[this document]
0x13	unassigned	[this document]
0x14	unassigned	[this document]
0x15	unassigned	[this document]
0x16	VLAN action	[this document]
0x17	TIPD action	[this document]
0x18-		
0x3fff	Unassigned	[this document]
0x4000-		
0x7fff	Vendor assigned	[this document]
0x8000-		
0xFFFF	Reserved	[this document]

9.5. Wide Community Assignments

IANA is requested to assign values in the BGP Community Container Atom Type Registry

Name -----	Type value -----
FSv2 action atom	TBD5

IANA is requested to assign values from the Registered Type 1 BGP Wide Community Types:

Name -----	type Value -----
FSv2 Actions	TBD4

10. Security Considerations

The use of ROA improves on [RFC8955] by checking to see of the route origination. This check can improve the validation sequence for a multiple-AS environment.

>The use of BGPSEC [RFC8205] to secure the packet can increase security of BGP flow specification information sent in the packet.

The use of the reduced validation within an AS [RFC9117] can provide adequate validation for distribution of flow specification within a single autonomous system for prevention of DDoS.

Distribution of flow filters may provide insight into traffic being sent within an AS, but this information should be composite information that does not reveal the traffic patterns of individuals.

11. References

11.1. Normative References

[I-D.ietf-idr-bgp-flowspec-label]
Liang, Q., Hares, S., You, J., Raszuk, R., and D. Ma,
"Carrying Label Information for BGP FlowSpec", Work in
Progress, Internet-Draft, draft-ietf-idr-bgp-flowspec-
label-01, 6 December 2016,
<[https://www.ietf.org/archive/id/draft-ietf-idr-bgp-
flowspec-label-01.txt](https://www.ietf.org/archive/id/draft-ietf-idr-bgp-flowspec-label-01.txt)>.

- [I-D.ietf-idr-flowspec-interfaceset]
Litkowski, S., Simpson, A., Patel, K., Haas, J., and L. Yong, "Applying BGP flowspec rules on a specific interface set", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-interfaceset-05, 18 November 2019, <<https://www.ietf.org/archive/id/draft-ietf-idr-flowspec-interfaceset-05.txt>>.
- [I-D.ietf-idr-flowspec-l2vpn]
Hao, W., Eastlake, D. E., Litkowski, S., and S. Zhuang, "BGP Dissemination of L2 Flow Specification Rules", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-l2vpn-18, 24 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-idr-flowspec-l2vpn-18.txt>>.
- [I-D.ietf-idr-flowspec-mpls-match]
Yong, L., Hares, S., Liang, Q., and J. You, "BGP Flow Specification Filter for MPLS Label", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-mpls-match-01, 6 December 2016, <<https://www.ietf.org/archive/id/draft-ietf-idr-flowspec-mpls-match-01.txt>>.
- [I-D.ietf-idr-flowspec-nvo3]
Eastlake, D., Weiguo, H., Zhuang, S., Li, Z., and R. Gu, "BGP Dissemination of Flow Specification Rules for Tunneled Traffic", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-nvo3-14, 15 August 2021, <<https://www.ietf.org/internet-drafts/draft-ietf-idr-flowspec-nvo3-14.txt>>.
- [I-D.ietf-idr-flowspec-path-redirect]
Velde, G. V. D., Patel, K., and Z. Li, "Flowspec Indirection-id Redirect", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-path-redirect-11, 26 May 2020, <<https://www.ietf.org/archive/id/draft-ietf-idr-flowspec-path-redirect-11.txt>>.
- [I-D.ietf-idr-flowspec-srv6]
Li, Z., Li, L., Chen, H., Loibl, C., Mishra, G. S., Fan, Y., Zhu, Y., Liu, L., and X. Liu, "BGP Flow Specification for SRv6", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-srv6-00, 8 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-idr-flowspec-srv6-00.txt>>.

- [I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S.,
and P. Jakma, "BGP Community Container Attribute", Work in
Progress, Internet-Draft, draft-ietf-idr-wide-bgp-
communities-06, 10 January 2022,
<[https://www.ietf.org/archive/id/draft-ietf-idr-wide-bgp-
communities-06.txt](https://www.ietf.org/archive/id/draft-ietf-idr-wide-bgp-communities-06.txt)>.
- [RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791,
DOI 10.17487/RFC0791, September 1981,
<<https://www.rfc-editor.org/info/rfc791>>.
- [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>>.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y.,
Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack
Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001,
<<https://www.rfc-editor.org/info/rfc3032>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A
Border Gateway Protocol 4 (BGP-4)", RFC 4271,
DOI 10.17487/RFC4271, January 2006,
<<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended
Communities Attribute", RFC 4360, DOI 10.17487/RFC4360,
February 2006, <<https://www.rfc-editor.org/info/rfc4360>>.
- [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>>.
- [RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous
System Confederations for BGP", RFC 5065,
DOI 10.17487/RFC5065, August 2007,
<<https://www.rfc-editor.org/info/rfc5065>>.
- [RFC6482] Lepinski, M., Kent, S., and D. Kong, "A Profile for Route
Origin Authorizations (ROAs)", RFC 6482,
DOI 10.17487/RFC6482, February 2012,
<<https://www.rfc-editor.org/info/rfc6482>>.

- [RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <<https://www.rfc-editor.org/info/rfc7153>>.
- [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>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8955] Loibl, C., Hares, S., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", RFC 8955, DOI 10.17487/RFC8955, December 2020, <<https://www.rfc-editor.org/info/rfc8955>>.
- [RFC8956] Loibl, C., Ed., Raszuk, R., Ed., and S. Hares, Ed., "Dissemination of Flow Specification Rules for IPv6", RFC 8956, DOI 10.17487/RFC8956, December 2020, <<https://www.rfc-editor.org/info/rfc8956>>.
- [RFC9015] Farrel, A., Drake, J., Rosen, E., Uttaro, J., and L. Jalil, "BGP Control Plane for the Network Service Header in Service Function Chaining", RFC 9015, DOI 10.17487/RFC9015, June 2021, <<https://www.rfc-editor.org/info/rfc9015>>.
- [RFC9117] Uttaro, J., Alcaide, J., Filsfils, C., Smith, D., and P. Mohapatra, "Revised Validation Procedure for BGP Flow Specifications", RFC 9117, DOI 10.17487/RFC9117, August 2021, <<https://www.rfc-editor.org/info/rfc9117>>.

11.2. Informative References

- [RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011, <<https://www.rfc-editor.org/info/rfc6241>>.
- [RFC8040] Bierman, A., Bjorklund, M., and K. Watsen, "RESTCONF Protocol", RFC 8040, DOI 10.17487/RFC8040, January 2017, <<https://www.rfc-editor.org/info/rfc8040>>.
- [RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", RFC 8205, DOI 10.17487/RFC8205, September 2017, <<https://www.rfc-editor.org/info/rfc8205>>.

[RFC8206] George, W. and S. Murphy, "BGPsec Considerations for Autonomous System (AS) Migration", RFC 8206, DOI 10.17487/RFC8206, September 2017, <<https://www.rfc-editor.org/info/rfc8206>>.

Authors' Addresses

Susan Hares
Hickory Hill Consulting
7453 Hickory Hill
Saline, MI 48176
United States of America

Phone: +1-734-604-0332
Email: shares@endzh.com

Donald Eastlake
Futurewei Technologies
2386 Panoramic Circle
Apopka, FL 32703
United States of America

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

Chaitanya Yadlapalli
ATT
United States of America

Email: cy098d@att.com

Sven Maduschke
Verizon
Germany

Email: sven.maduschke@de.verizon.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 8 October 2022

H. Bidgoli, Ed.
Nokia
D. Voyer
Bell Canada
A. Stone
Nokia
R. Parekh
S. Krier
S. Agrewal
Cisco System, Inc.
6 April 2022

Advertising p2mp policies in BGP
draft-hb-idr-sr-p2mp-policy-05

Abstract

SR P2MP policies are set of policies that enable architecture for P2MP service delivery.

A P2MP policy consists of candidate paths that connects the Root of the Tree to a set of Leaves. The P2MP policy is composed of replication segments. A replication segment is a forwarding instruction for a candidate path which is downloaded to the Root, transit nodes and the leaves.

This document specifies a new BGP SAFI with a new NLRI in order to advertise P2MP policy from a controller to a set of nodes.

This document introduces three new route types within this NLRI, one for P2MP policy and its candidate paths that need to be programmed on the Root node, one for the replication segment incoming SID which uniquely will identify the cross connect and another for each outgoing interface that the packets get replicated to. The last two route types are forwarding instructions that needs to be programmed on the Root, and optionally on Transit and Leaf nodes.

It should be noted that this document does not specify how the Root and the Leaves are discovered on the controller, it only describes how the P2MP Policy and Replication Segments are programmed from the controller to the nodes.

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 8 October 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
2. Conventions used in this document	4
3. P2MP Policy and Replication Segment Encoding	4
3.1. P2MP Policy SAFI and NLRI	4
3.1.1. P2MP Policy Route - Route Type TBD1	5
3.1.2. Replication segment Route Binding SID- Route type TBD 2	6
3.1.3. Replication segment Route OIF- Route type TBD 3	8
3.2. Tunnel Encapsulation Attribute	9
3.2.1. SR P2MP policy encoding	9
3.2.2. Replication segment Binding SID encoding	10
3.2.3. Replication segment OIF encoding	10
3.3. P2MP Policy Sub-TLVs	11
3.3.1. preference Sub-TLV	11
3.3.2. leaf-list Sub-TLV	11
3.3.3. path-instance Sub-TLV	12
3.3.3.1. active instance-id Sub-TLV	12
3.3.3.2. instance-id Sub-TLV	13
3.4. Replication segment Sub-TLVs	13
3.4.1. Segment list Sub-TLV	14

3.4.2.	Weight sub-tlv	14
3.4.3.	Protection sub-tlv	14
3.4.4.	Segment Sub-TLV	15
4.	P2MP Policy Operation	15
4.1.	Configuration and advertisement of P2MP Policies	16
4.2.	Reception of an P2MP Policy NLRI	16
4.3.	Global Optimization for P2MP LSPs	16
5.	IANA Consideration	17
6.	Security Considerations	17
7.	Acknowledgments	17
8.	References	17
8.1.	Normative References	17
8.2.	Informative References	17
	Authors' Addresses	18

1. Introduction

The draft [draft-ietf-pim-sr-p2mp-policy] defines a variant of the SR Policy [draft-ietf-spring-segment-routing-policy] for constructing a P2MP segment to support multicast service delivery.

A Point-to-Multipoint (P2MP) Policy contains a set of candidate paths and identifies a Root node and a set of Leaf nodes in a Segment Routing Domain. The draft also defines a Replication segment, which corresponds to the state of a P2MP segment on a particular node. The Replication segment is the forwarding instruction for a P2MP LSP at the Root, Transit and Leaf nodes.

For a P2MP segment, a controller may be used to compute a tree from a Root node to a set of Leaf nodes, optionally via a set of replication nodes. A packet is replicated at the root node and optionally on Replication nodes towards each Leaf node.

We define two types of a P2MP segment: Ingress Replication (aka Spray) and Downstream Replication (aka TreeSID).

A Point-to-Multipoint service delivery could be via Ingress Replication (aka Spray in some SR context), i.e., the root unicasts individual copies of traffic to each leaf. The corresponding P2MP segment consists of replication segments only for the root and the leaves.

A Point-to-Multipoint service delivery could also be via Downstream Replication (aka TreeSID in some SR context), i.e., the root and some downstream replication nodes replicate the traffic along the way as it traverses closer to the leaves.

It should be noted that two replication nodes can be connected directly, or they can be connected via unicast SR segment or a segment list.

The leaves and the root of a p2mp policy can be discovered via the multicast protocols or procedures like NG-MVPN [RFC6513] or manually configured on the PCC (CLI) or the PCE.

Based on the discovered root and leaves, the controller builds a P2MP policy and advertise it to the head-end router (i.e. the root of the P2MP Tree). The advertisement uses BGP extensions defined in this document. The controller also calculates the tree path and builds the replication segments on each segment of the tree, Root, Transit and Leaf nodes and downloads the forwarding instructions to the nodes via BGP extensions defined in this document.

SR p2mp policy is a variant of the SR policy and as such it reuses the concept of a candidate path. This draft reuses some of the concepts and TLVs mentioned in [draft-ietf-idr-segment-routing-te-policy]

A candidate path within the P2MP policy can contain multiple path-instances. A path-instance can be viewed as a P2MP LSP. For candidate path global optimization purposes, two or more path-instances can be used to execute make before break procedures.

Each path-instance is a P2MP LSP as such each path-instance needs a set of replication segments to construct its forwarding instructions.

2. Conventions used in this document

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

3. P2MP Policy and Replication Segment Encoding

3.1. P2MP Policy SAFI and NLRI

This document defines a new BGP NLRI, called the P2MP-POLICY NLRI.

A new SAFI is defined: the SR P2MP Policy SAFI, (Codepoint tbd assigned by IANA). The following is the format of the P2MP-POLICY NLRI:

route type	1 octet

length	1 octet

route type specific (variable)	

- * The Route type field defines the encoding of the rest of the P2MP-POLICY NLRI.
- * The length field indicates the length in octets of the route type specific data, excluding route type and length
- * This document defines the following route types:
 - P2MP Policy route: TBD1, this is the actually P2MP policy on the root which contains the candidate paths, its preference and path instances.
 - Replication Segment Binding SID: TBD2, this is part of the replication segment and it is used for programming the incoming SID used to identify a P2MP cross connect.
 - Replication Segment OIF: TBD3, this is a single Outgoing Interface for the P2MP cross connect. It also contains the outgoing SID.

The NLRI containing the SR P2MP 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 "TBD" (assigned by IANA from the "Subsequent Address Family Identifiers (SAFI) Parameters" registry).

All other recommendations of [draft-ietf-idr-segment-routing-te-policy] section SR Policy SAFI and NLRI, should be taken into account for P2MP policy.

3.1.1. P2MP Policy Route - Route Type TBD1

Root-ID Length	1 octets
Root-ID	4 or 16 octets (ipv4/ipv6)
Tree-ID	4 octets
Distinguisher	4 octets

- * Root-ID: IPv4/IPv6 address of the head-end (Root) of the p2mp tree, based on AFI.
- * Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end (root)router.
- * Distinguisher: 4-octets value uniquely identifying the policy in the context of <Tree-ID, Originating Router's IP> tuple. The distinguisher has no semantic value and is solely used by the SR P2MP Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR P2MP Policy.

3.1.2. Replication segment Route Binding SID- Route type TBD 2

There can be two type of replication segment, shared and non-shared. A shared replication segment can carry multiple MVPN services or it can be used for Facility Fast reroute protecting multiple P2MP trees. A non-shared tree is used when the label field of the PMSI Tunnel Attribute (PTA) is set to 0 as per [draft-ietf-bess-mvpn-evpn-sr-p2mp]. The Binding SID route type Programs the incoming replication SID on the replication node. Since a replication cross connect has a single incoming replication SID with a set of Outgoing Interfaces, this route type can be used to download the replication SID once for the cross connect.

	Root-ID Length		1 octets
~	Root-ID	~	4 or 16 octets (ipv4/ipv6)
	Tree-ID		4 octets
	Distinguisher		4 octets
	instance-ID		2 octets
	Node-ID Length		1 octets
~	Node-ID	~	4 or 16 octets
	Replication SID Length		1 octets
~	Replication SID	~	4 or 16 octets

- * Root-ID: IPv4/IPv6 address of the head-end (Root) of the p2mp tree based on AFI.
- * Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end router (Root)
- * instance-id, identifies the path-instance with in the p2mp-policy. Each candidate path can have one, two or more path-instance. Path-instance is used for global optimization of the candidate path via make before break procedures. Instance-ID can be used
- * Distinguisher: 4-octets value uniquely identifying the policy in the context of <Root-ID, Tree-ID> tuple. The distinguisher has no semantic value and is solely used by the SR P2MP Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR P2MP Policy.
- * Node-ID: This Node's IPv4/IPv6 address
- * Replication SID: the incoming replication SID used to identify this replication point (MPLS or SRv6). Note the replication SID is not part of the NLRI key.

3.1.3. Replication segment Route OIF- Route type TBD 3

This route type is used to identify and program each out going interface individually for a replication cross connect. Downloading each OIF individually ensures easier modification and programming and will keep the programming of each OIF in par with [draft-ietf-idr-segment-routing-te-policy] . Note: this route type can be used for shared and non-shared replication segment as it was explained in previous sections.

Root-ID Length	1 octets
Root-ID	4 or 16 octets (ipv4/ipv6)
Tree-ID	4 octets
Distinguisher	4 octets
instance-ID	2 octets
Node-ID Length	1 octets
Node-ID	4 or 16 octets
Downstream-Node Length	1 octets
Downstream-Node	4 or 16 octets
Outgoing-TreeSID Length	1 octets
Outgoing-TreeSID	4 or 16 octets

- * Root-ID: IPv4/IPv6 address of the head-end (Root) of the p2mp tree based on AFI.
- * Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end router (Root)
- * instance-id, identifies the path-instance with in the p2mp-policy. Each candidate path can have one, two or more path-instance. Path-instance is used for global optimization of the candidate path via make before break procedures. Instance-ID can be used

- * Distinguisher: 4-octets value uniquely identifying the policy in the context of <Root-ID, Tree-ID> tuple. The distinguisher has no semantic value and is solely used by the SR P2MP Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR P2MP Policy.
- * Node-ID: Node's IPv4/IPv6 address
- * Downstream Node: Downstream Node Identifier
- * Outgoing TreeSID: The outgoing SID for this branch (MPLS or SRv6). Note the outgoing-TreeSID is not part of the NLRI Key.

3.2. Tunnel Encapsulation Attribute

The content of this new NLRI is encoded in the tunnel Encapsulation Attribute originally defined in [RFC9012] using two new Tunnel-Type TLV (codepoint is TBD, assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry) one for P2MP Policy and another for Replication segment.

3.2.1. SR P2MP policy encoding

SR P2MP Policy SAFI NLRI: <route-type p2mp-policy>

Attributes:

```
Tunnel Encaps Attribute (23)
  Tunnel Type: (TBD, P2MP-Policy)
    Preference
    Policy Name
    Policy Candidate Path Name
    leaf-list (optional)
      remote-end point
      remote-end point
      ...
    path-instance
      active-instance-id
      instance-id
      instance-id
      ...
```

- * Relevant only at the Root.
- * SR P2MP-POLICY NLRI and P2MP Policy route type.
- * Tunnel Encapsulation Attribute is defined in [RFC9012]

- * Tunnel-Type is set to P2MP-Policy Tunnel-Type TBD (assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).
- * Policy Name, Policy Candidate Path Name are defined in [draft-ietf-idr-segment-routing-te-policy]
- * Preference, leaf-list, remote-end point and path- instance, instance-ids are defined in this document.
- * Additional sub-TLVs may be defined in the future.

3.2.2. Replication segment Binding SID encoding

```
replication segment Binding SID SAFI NLRI:
    <route-type non-sahred/shared
      tree replication-segment-binding-sid>
```

This route type has no additional sub-TLVs, and it is only meant to download the incoming SID for the replication cross connect.

3.2.3. Replication segment OIF encoding

```
replication segment SAFI NLRI: <route-type non-sahred/shared
                                tree replication-segment-oif>
```

Attributes:

```
Tunnel Encaps Attribute (23)
  Tunnel Type: (TBD Replication-Segment-oif)
    segment-list
      weight (optional)
      protection (optional, must be present when protection flag is enabled
for downstream-nodes)
      segment
      segment
      ...
    segment-list
      weight (optional)
      protection (optional, must be present when protection flag is enabled
for downstream-nodes)
      segment
      segment
      ...
    segment-list (protection segment list)
      protection (protecting the first segment list, can't have weight sub-
tlv)
      segment
      segment
      ...
    ...
  ...
```

- * SR P2MP-POLICY NLRI and non-shared tree Replication segment route type or shared tree Replication segment route type.
- * Tunnel Encapsulation Attribute is defined in [RFC9012].
- * Tunnel-Type is set to Replication Segment OIF Tunnel Type, TBD (assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).
- * segment-list are defined in this document.
- * Additional sub-TLVs may be defined in the future.

3.3. P2MP Policy Sub-TLVs

EACH P2MP policy NLRI represents a candidate path for a P2MP policy. A P2MP policy can have multiple candidate paths and would need multiple P2MP policy NLRI to download all the candidate paths.

3.3.1. preference Sub-TLV

As defined in preference Sub-TLV section in [draft-ietf-idr-segment-routing-te-policy] the candidate path with highest preference is the active candidate path.

3.3.2. leaf-list Sub-TLV

The leaf list sub-tlv identifies a set of leaves for the tree. Each leaf is a remote endpoint as defined in [RFC9012] The leaf-list sub-tlv is optional. The PCE can choose to download the leaf list every time it is configured or learns a new leaf. If the PCE chooses to download this optional sub-tlv it should download the entire set of the end-points every time the endpoint list has been modified. The leaf list has informational value only hence why it is optional and it is not required for the root PE to operate. However, it must be noted that in some cases the end-points list can become very large with 100s of leaves.

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
Type										Length										RESERVED																			
// sub-TLVs //																																							

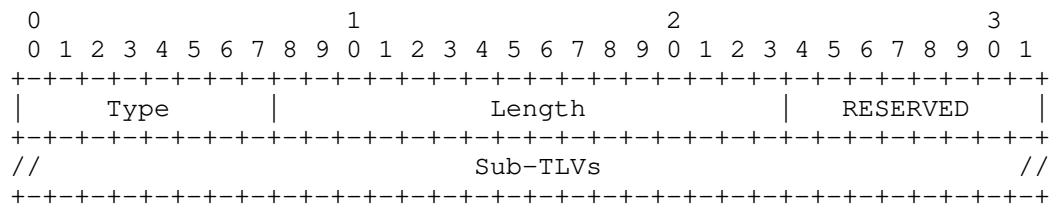
- * Type: TBD, 1 octet

- * Length: 2 octets, the total length (not including the Type and Length fields) of the sub-TLVs encoded within the leaf-list sub-TLV.
- * RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- * sub-TLVs: One or more remote endpoint sub-TLVs. Note the remote endpoint object is defined in [RFC9012]

3.3.3. path-instance Sub-TLV

The path instance sub-tlv contains a set of instance-ids (P2MP LSPs). These LSPs can be used for MBB procedure under a candidate path. Each LSP Instance-id has a unique id (4 octets) with in the <root node, P2MP policy>, in other word it is unique per <root node, tree-id>. The PCE SHOULD always download all instance-ids to the node. The active instance is identified via the active instance-id sub-tlv.

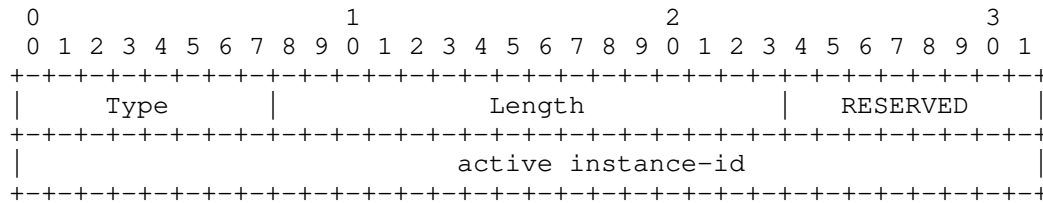
The P2MP LSP and its replication segments should be configured from root to the leaves first before the PCE switches that active instance-id to this new instance.



- * Type: TBD, 1 octet
- * Length: 2 octets, the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.
- * RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt
- * sub-TLVs: * active instance-id * one or more instance-id

3.3.3.1. active instance-id Sub-TLV

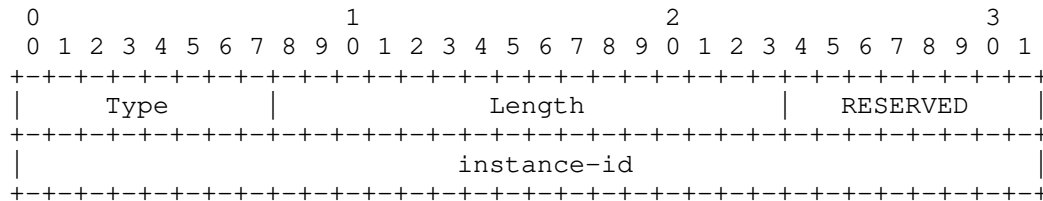
The Active instance-id is used to identify the P2MP LSP which should be active amongst the collection of instances.



- * Type: TBD.
- * Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.
- * RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- * active instant-id: The identifier of the active instance-id

3.3.3.2. instance-id Sub-TLV

Multiple Instance-ids can be programmed for a candidate path.



- * Type: TBD
- * Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.
- * RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- * instan-id: a 32 bit unique identifier. The instance-id is unique with in the context of the <root node, p2mp policy>

3.4. Replication segment Sub-TLVs

3.4.1. Segment list Sub-TLV

The segment list Sub-TLV is defined in [draft-ietf-spring-segment-routing-policy]. The segment-list Sub-TLV contains one or more segment Sub-TLVs. Two replication segments can be directly connected via a replication sid or can be connected via a unicast segment list and a replication sid. In the later case the replication sid needs to be at the bottom of the unicast segment list.

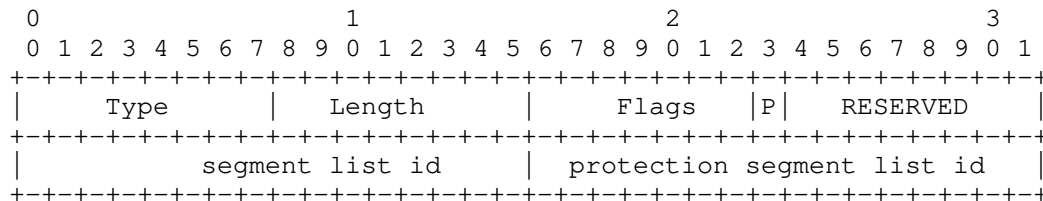
3.4.2. Weight sub-tlv

The Weight sub-TLV is optional and is as defined in [draft-ietf-idr-segment-routing-te-policy]. With in the downstream node sub-tlv, there can be one or more segment list used for ECMP. In this case the weight sub-tlv can provide weighted ECMP.

3.4.3. Protection sub-tlv

Protection sub-tlv is optional, if FRR is desired for the downstream node this sub-tlv can be used to identify the protection segment list. To identify protection segment list this sub-tlv provides a segment list identifier. If protection is desired under the endpoint all the segment lists should have this sub-tlv. A protection segment list can not have a weight sub-tlv and it can not participate in ECMP. That said a segment list that is being protected can have a weight sub-tlv and participate in ECMP.

In general protection segment list is used only if replication segments are directly connected and there is no unicast segment list connecting two replication segment. If there is a unicast replication segment connecting the two replication sid, then the unicast protection mechanism can be exercise and there is no need for this protection sub-tlv, hence why this sub-tlv is optional.



* Type : tbd, 1 octet.

* Length: 8

- * Flag: 1 octet, the P bit is set when this segment list is protected by another segment list for the downstream node
- * segment list id: the segment list id
- * protection segment list id: the segment list id that is being used as protection.

3.4.4. Segment Sub-TLV

The segment sub-Tlv is identified in [draft-ietf-idr-segment-routing-te-policy]. As it was mentioned before two replication segments can be connected directly to each other or via a segment list. If they are connected directly to each other then the segment list can be constructed via:

- * If the replication segment is steered via IPv4 or IPv6 nexthops or interface then the segment type E or G can be used with the new R flag set.
- * If the replication segment is steered via a SR Unicast node or adjacency SID then segment type A can be used with the new R flag set. Unicast SR segment types can also be configured for steering.

If they are connected via SR domain then the segment list can contain multiple different types of SIDs, such as Node, Adjacency or Binding SIDs. In this case the replication sid is at the bottom of the stack and of type A with the R flag set. The SR node/adjacency or binding sids steer the packet through a SR domain until it reaches another replication segment. where the bottom of the stack replication sid identifies the forwarding information on that replication segment.

It should be noted that the segment sub-TLV is only used to program the unicast SR Segment or outgoing interface for the replication SID outgoing interface. The outgoing tree SID it self is programmed in the appropriate route type.

4. P2MP Policy Operation

Inline with [draft-ietf-idr-segment-routing-te-policy] the consumer of an P2MP Policy is not the BGP process. The BGP process is used for distributing the P2MP policy NLRI and its route-types but its installation and use is outside the scope of BGP. The detail for P2MP Policy can be found in [draft-ietf-pim-sr-p2mp-policy]

4.1. Configuration and advertisement of P2MP Policies

The controller usually is connected to the receivers via a route reflector. As such one or more route-target SHOULD be attached to the advertisement of P2MP Policy NLRI and its route-type. Each route target identifies one head-end (root nodes) for P2MP Policy route or one or more head-end, transit and leaf nodes for the Non- Shared/ Shared Tree Replication Segment route, for the advertised P2MP Policy.

4.2. Reception of an P2MP Policy NLRI

When a BGP speaker receives an P2MP Policy NLRI the following rules apply:

- * The P2MP Policy update MUST have either the NO_ADVERTISE community or at least one route-target extended community in IPv4-address format. If a router supporting this document receives an P2MP Policy update with no route-target extended communities and no NO_ADVERTISE community, the update MUST NOT be processed. Furthermore, it SHOULD be considered to be malformed, and the "treat-as-withdraw" strategy of [RFC7606] is applied.
- * If one or more route-targets are present, then at least one route-target MUST match one of the BGP Identifiers of the receiver in order for the update to be considered usable. The BGP Identifier is defined in [RFC4271] as a 4 octet IPv4 address. Therefore the route- target extended community MUST be of the same format.
- * If one or more route-targets are present and no one matches any of the local BGP Identifiers, then, while the P2MP Policy NLRI is acceptable, it is not usable on the receiver node.

4.3. Global Optimization for P2MP LSPs

When a P2MP LSP needs to be optimized for any reason (i.e. it is taking on an FRR Path or new routers are added to the network) a global optimization is possible. Note that optimization works per candidate path. Each candidate path is capable of global optimization. To do so each candidate path contains two or more path- instances. Each path instance is a P2MP LSP, each P2MP LSP is identified via a path-instance-id (equivalent to an lsp-id [RFC3209]). After calculating an optimized P2MP LSP path the PCE will program the candidate path with a 2nd path instance and its set of replication segments for this path-instance on the root, transit and leaf nodes. After the optimized LSP replication segments are downloaded a MBB procedure is performed and the previous instance of the path instance is deleted and removed from head-end node and its

corresponding replication segments from head-end, transit and leaves.

5. IANA Consideration

- * A new SAFI is defined: the SR P2MP Policy SAFI, (Codepoint tbd assigned by IANA)
- * 3 new Route type field defines the encoding of the rest of the P2MP- POLICY SAFI
 - P2MP Policy Route
 - Replication Segment Binding Sid
 - Replication Segment OIF
- * Two new Tunnel type to be assigned by IANA
 - P2MP-Policy Tunnel-Type
 - Replication Segment OIF Tunnel Type

6. Security Considerations

TBD

7. Acknowledgments

8. References

8.1. Normative References

[RFC2119] "S. Bradner "Key Words for use in RFCs to Indicate Requirement levels"", October 2019.

8.2. Informative References

[draft-ietf-bess-mvpn-evpn-sr-p2mp]
"R. Parekh, C. Filsfils, A.V. Venkateswaran, H. Bidgoli, D. Voyer, Z. Zhang "Multicast and Ethernet VPN with Segment Routing P2MP"".

[draft-ietf-idr-segment-routing-te-policy]
"s. Previdi, C. Filsfils, K. Talaulikar, P. Mattes, D. Jain, S. Lin "Advertise Segment Routing Policies in BGP"".

- [draft-ietf-pim-sr-p2mp-policy]
"D. Voyer, C. Filsfils, R.Prekh, H.bidgoli, Z. Zhang,
"draft-ietf-pim-sr-p2mp-policy"", October 2019.
- [draft-ietf-spring-segment-routing-policy]
"C. Filsfils, K. Talaulikar, D. Voyer, A. Bogdanov, P.
Mattes "Segment Routing Policy Architecture".
- [RFC4271] "Y. Rekhter, T. Li, S. Hares "A Border Gateway Protocol 4
(BGP-4) "".
- [RFC4760] "T. Bates, R. Chandra, D. Katz, Y. Rekhter "Multiprotocol
Extensions for BGP-4"".
- [RFC6513] "E. Rosen, R. Aggarwal "Multicast in MPLS/BGP IP VPNs"".
- [RFC7606] "e. Chen, J. Scudder, P. Mohapatra, K. Patel "Revised
Error handling for BGP UPDATE Messages"".
- [RFC9012] "K. Patel, G. Van de Velde, S. Sangli, J. Scudder "The BGP
Tunnel Encapsulation Attribute"".

Authors' Addresses

Hooman Bidgoli (editor)
Nokia
Ottawa
Canada
Email: hooman.bidgoli@nokia.com

Daniel Voyer
Bell Canada
Montreal
Canada
Email: daniel.yover@bell.ca

Andrew Stone
Nokia
Ottawa
Canada
Email: andrew.stone@nokia.com

Rishabh Parekh
Cisco System, Inc.
San Jose,
United States of America
Email: riparekh@cisco.com

Serge Krier
Cisco System, Inc.
Rixensart
Belgium
Email: sekrier@cisco.com

Swadesh Agrewal
Cisco System, Inc.
San Jose,
United States of America
Email: swaagraw@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 28 October 2022

J. Head, Ed.
T. Przygienda
Juniper Networks
26 April 2022

BGP-LS Extensions for IS-IS Flood Reflectors
draft-head-idr-bgp-ls-isis-fr-01

Abstract

This document defines new BGP-LS (BGP Link-State) TLVs in order to carry IS-IS Flood Reflection information.

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 28 October 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	2
1.1. Requirements Language	2
2. BGP-LS Extensions for IS-IS Flood Reflectors	2
3. BGP-LS TLVs for IS-IS Flood Reflection	2
4. IANA Considerations	3
4.1. Requested TLV Entries	3
5. Security Considerations	3
6. Acknowledgements	4
7. References	4
7.1. Normative References	4
Authors' Addresses	4

1. Introduction

BGP Link-State RFC7752 [RFC7752] defines mechanisms to advertise information about the underlying IGP in BGP NLRI to an external entity (e.g. a controller). New BGP-LS TLVs are required in order to facilitate IS-IS Flood Reflection [IS-IS-FR] extensions.

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. BGP-LS Extensions for IS-IS Flood Reflectors

This document defines the following BGP-LS TLV code point value in accordance with RFC7752 rules:

TLV Code Point	Description	IS-IS TLV
TBD1	Flood Reflection TLV	TBD1 (161) [IS-IS-FR]

Table 1: BGP-LS Flood Reflection TLV Code Points

TLV formats are described in detail in subsequent subsections.

3. BGP-LS TLVs for IS-IS Flood Reflection

This TLV advertises Flood Reflector details. The semantics and values of the fields in the TLV are described in [IS-IS-FR].

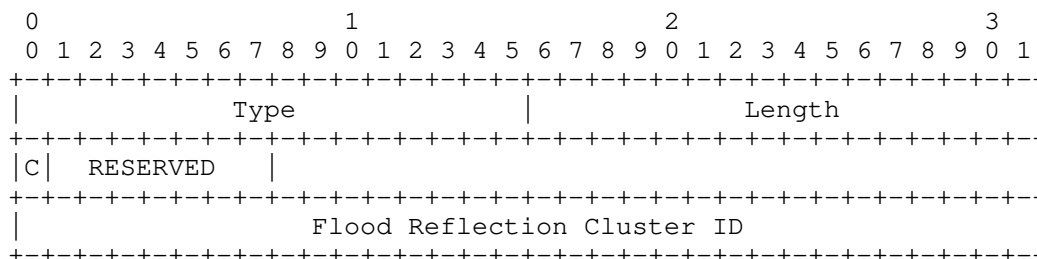


Figure 1: Flood Reflection TLVs

where:

Type: TBD1

Length: 5

4. IANA Considerations

This section requests entries from the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry for the following TLVs:

4.1. Requested TLV Entries

TLV Code Point	Description
TBD1	Flood Reflection TLV

Table 2: IANA Requests

5. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See the "Security Considerations" section of [RFC4271] for a discussion of BGP security. Also, refer to [RFC4272] and [RFC6952] for analyses of BGP security issues. Security considerations for acquiring and distributing BGP-LS information are discussed in [RFC7752].

The TLVs introduced in this document are used to propagate IS-IS Flood Reflection TLVs defined in [IS-IS-FR]. These TLVs represent IS-IS Flood Reflector state and are therefore assumed to support any/all of the required security and authentication mechanisms as described in [IS-IS-FR] to prevent any security issues when propagating the TLVs into BGP-LS.

6. Acknowledgements

7. References

7.1. Normative References

- [IS-IS-FR] Przygienda, T., Bowers, C., Lee, Y., Sharma, A., and R. White, "IS-IS Flood Reflection", October 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-lsr-isis-flood-reflection>>.
- [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., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", January 2006, <<https://www.rfc-editor.org/info/rfc4272>>.
- [RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", May 2013, <<https://www.rfc-editor.org/info/rfc6952>>.
- [RFC7752] Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", March 2016, <<https://www.rfc-editor.org/info/rfc7752>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.

Authors' Addresses

Jordan Head (editor)
Juniper Networks
1137 Innovation Way
Sunnyvale, CA
United States of America
Email: jhead@juniper.net

Tony Przygienda
Juniper Networks
1137 Innovation Way
Sunnyvale, CA
United States of America
Email: prz@juniper.net

Internet Engineering Task Force
Internet Draft
Intended status: Standards Track
Expiration Date: April 21, 2022

E. Chen
Palo Alto Networks
S. Sangli
Juniper Networks
October 20, 2021

Dynamic Capability for BGP-4
draft-ietf-idr-dynamic-cap-16.txt

Abstract

This document defines a new BGP capability termed "Dynamic Capability", which would allow the dynamic update of capabilities over an established BGP session. This capability would facilitate non-disruptive capability changes by BGP speakers.

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 21, 2022.

Copyright Notice

Copyright (c) 2021 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.

1. Introduction

Currently BGP capabilities [RFC5492] are only advertised in the BGP OPEN message [RFC4271] during the session initialization. In order to enable or disable a capability (such as the Address Family support [RFC4760]), an established session would need to be reset, which may disrupt other services running over the session. In addition, currently an advertised capability can not be updated on-demand over an established session. One example of such a requirement is for adjusting the "Restart Time" in the Graceful Restart Capability [RFC4724] when performing certain planned maintenance in a network.

This document defines a new BGP capability termed "Dynamic Capability", which would allow the dynamic update of capabilities over an established BGP session. This capability would facilitate non-disruptive capability changes by BGP speakers.

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. Dynamic Capability

The Dynamic Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is specified in the "IANA Considerations" section of this document. The Capability Value field consists of a list of capability codes (one-octet for each) that specify the capabilities that MAY be revised dynamically by the remote speaker.

By advertising the Dynamic Capability to a peer in the OPEN, a BGP speaker conveys to the peer that the speaker is capable of receiving and properly handling the CAPABILITY message (as defined in the next Section) from the peer after the BGP session has been established.

3. Capability Message

The CAPABILITY Message is a new BGP message type with type code 6. In addition to the fixed-size BGP header [RFC4271], the CAPABILITY message contains one or more of the following tuples of capability revisions:

Init/Ack (1 bit)
Ack Request (1 bit)
Reserved (5 bits)
Action (1 bit)
Sequence Number (4 octets)
Capability Code (1 octet)
Capability Length (2 octets)
Capability Value (variable)

The Init/Ack bit indicates whether a capability revision is being initiated (when set to 0), or being acknowledged (when set to 1).

The Ack Request bit indicates whether an acknowledgment is requested (when set to 1), or not (when set to 0) for a capability revision being initiated.

The Reserved bits should be set to zero by the sender and ignored by the receiver.

The Action bit is 0 for advertising a capability, and 1 for removing a capability.

The Sequence Number field can be used by a BGP speaker to match an acknowledgment with a capability revision that the speaker initiated previously.

Conceptually the triple <Capability Code, Capability Length, Capability Value> is the same as the one defined in [RFC5492], and it specifies a capability for which the "Action" shall be applied. The Capability Length field, though, is larger than the one specified in

[RFC5492].

If multiple capability instances (as described in [RFC5492]) are defined for the capability code, then each capability instance SHALL be revised individually. The triple <Capability Code, Capability Length, Capability Value> in the CAPABILITY message SHALL contain only one instance of the capability. The Multiprotocol Extensions Capability specified in [RFC4760] is an example of such a capability that has multiple instances defined.

If multiple capability instances (as described in [RFC5492]) are not defined for the capability code, then the "Action" specified applies to the whole capability identified by the capability code. Furthermore, if the "Action" is to remove a capability, then the Capability Length field SHOULD be set to zero by the sender and the Capability Value field MUST be ignored by the receiver even when the Capability Length field has a non-zero value.

If the "Action" is to remove a capability and the Capability Length field is zero, then the whole capability identified by the capability code is removed regardless whether multiple capability instances are defined for the capability code.

4. Operation

A BGP speaker that is willing to receive the CAPABILITY message (for one or more capability codes) from its peer SHOULD use the BGP Capabilities Advertisement [RFC5492] to advertise the Dynamic Capability for these capability codes.

A BGP speaker MAY send to its peer a CAPABILITY message to initiate revisions for one or more capability codes only if these capability codes are listed in the Dynamic Capability of the OPEN message received from its peer.

A CAPABILITY message MAY be received only in the Established state. Receiving a CAPABILITY message in any other state is a Finite State Machine Error as defined in [RFC4271]. A BGP speaker SHOULD reset the HoldTimer upon receiving a CAPABILITY message from its peer.

When a BGP speaker sends a CAPABILITY message to its peer to initiate a capability revision, the Init/Ack bit for the capability revision in the message MUST be set to 0. The setting of the Ack Request bit is capability specific. The assignment of the Sequence Number is a local matter, but MUST allow the BGP speaker to unambiguously identify a capability revision it initiated previously based on the Sequence Number carried in the acknowledgment from the peer.

If the Init/Ack bit is set to 1 for a capability revision in a CAPABILITY message received by a BGP speaker, then the BGP speaker SHALL treat the capability revision as an acknowledgment of the receipt of a capability revision initiated by the BGP speaker. The BGP speaker MUST ignore the Ack Request bit, and SHALL use the Sequence Number carried in the capability revision to match with the capability revision previously initiated. The BGP speaker SHALL ignore an acknowledgment for a capability revision in which an acknowledgment was not requested by the BGP speaker. If the Sequence Number carried in the capability revision does not match any of the the Sequence Numbers used in the capability revisions initiated by the BGP speaker, then the BGP speaker SHOULD send a NOTIFICATION message as specified in the Error Handling section.

If the Init/Ack bit is set to 0 for a capability revision in a CAPABILITY message received by a BGP speaker, then the BGP speaker SHOULD first validate the capability code in the message. If the capability code is not listed in the Dynamic Capability advertised by the speaker to the peer, the BGP speaker SHOULD send a NOTIFICATION message as specified in the Error Handling section. For a valid capability code, if the Ack Request bit is set to 1, the BGP speaker MUST first send a CAPABILITY message to acknowledge the receipt of the capability revision. The Init/Ack bit in the acknowledgment MUST be set to 1, and all the other fields in the capability revision MUST be kept unchanged.

After receiving a capability revision initiated by a peer, the BGP speaker SHALL update the capability previously received from that peer based on the Action bit in the message, and then function in accordance with the revised capability for the peer. The BGP speaker SHALL ignore such a capability revision that either results in no change to an existing capability, or removes a capability that was not advertised previously. The procedures specified in the "Error Handling" section SHOULD be followed when an error is detected in processing the CAPABILITY message.

In order to avoid ambiguities in sending and processing UPDATE messages, certain capability revisions may require close coordination between the BGP speaker (the Initiator) that initiates the capability revisions and another BGP speaker (the Receiver) that receives the capability revisions. The mechanism of acknowledgment defined in this document SHALL be used for the revision of such a capability. For the Initiator, the capability revision SHALL take effect (for the purpose of sending updates) immediately after the capability revision is sent, and the capability revision SHALL take effect (for the purpose of receiving updates) immediately after an acknowledgment is received from the Receiver. For the Receiver, the capability revision SHALL take effect (for the purpose of receiving updates)

immediately after the capability revision is received from the Initiator, and the capability revision SHALL take effect (for the purpose of sending updates) immediately after an acknowledgment is sent.

5. Error Handling

This document defines a new NOTIFICATION error code:

Error Code	Symbolic Name
7	CAPABILITY Message Error

The following error subcodes are defined as well:

Subcode	Symbolic Name
1	Unknown Sequence Number
2	Invalid Capability Length
3	Malformed Capability Value
4	Unsupported Capability Code

If a BGP speaker detects an error while processing a CAPABILITY message, it MUST send a NOTIFICATION message with Error Code CAPABILITY Message Error. If any of the defined error subcode is applicable, the Data field of the NOTIFICATION message MUST contain the tuple for the capability revision that causes the speaker to send the message.

If the Sequence Number carried in a capability revision marked as acknowledgment does not match any of the the Sequence Numbers used in the capability revisions initiated by the BGP speaker, then the error subcode is set to Unknown Sequence Number.

If the Capability Length field in the CAPABILITY message is incorrect for a Capability Code, then the error subcode is set to Invalid Capability Length.

If the Capability Value field in the CAPABILITY message is malformed (the definition of "malformed" depends on the Capability Code), then the error subcode is set to Malformed Capability Value.

If the Capability Code in the CAPABILITY message is not any of the capability codes advertised in the Dynamic Capability by the speaker, then the error subcode is set to Unsupported Capability Code.

6. Implementation Considerations

The extension specified in this document is designed for BGP capabilities in general. It can be used for a simple capability revision (e.g., a parameter change), as well as for a more complex revision that may involve changes to the encoding of BGP messages.

However, that does not mean all BGP capabilities warrant the support of dynamic revisions. For a given capability, one should carefully consider the tradeoffs between the complexities in its implementation and the potential benefits when deciding whether to support its dynamic revision. For example, the tradeoff considerations could be more favorable for the Address Family Capability [RFC4760] and the Graceful Restart Capability [RFC4724] than for the ADD-PATH Capability [RFC7911].

7. IANA Considerations

This document defines the CAPABILITY message type for BGP with type code 6, and a NOTIFICATION error code and subcodes for the errors in a CAPABILITY message.

This document uses a BGP capability code to indicate that a BGP speaker supports the Dynamic Capability. The capability code 67 has been assigned by IANA.

8. Security Considerations

The extension proposed in this document does not change the underlying security or confidentiality issues inherent in the existing BGP [RFC4271].

9. Acknowledgments

The authors would like to thank Yakov Rekhter, Ravi Chandra, Dino Farinacci, Pedro Marques, Chandrashekhhar Appanna, Derek Yeung, Bruno Rijsman, John Scudder, Jeffrey Haas and Heidi Ou for their review and comments.

10. References

10.1. 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., Rekhter, Y., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<http://www.rfc-editor.org/info/rfc4760>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<http://www.rfc-editor.org/info/rfc5492>>.
- [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>>.

10.2. Informative References

- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J. and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <<http://www.rfc-editor.org/info/rfc4724>>.
- [RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <<http://www.rfc-editor.org/info/rfc7911>>.

11. Authors' Addresses

Enke Chen
Palo Alto Networks, Inc.

Email: enchen@paloaltonetworks.com

Srihari R. Sangli
Juniper Networks, Inc.

Email: ssangli@juniper.net

IDR
Internet-Draft
Intended status: Standards Track
Expires: July 14, 2022

F. Qin
China Mobile
H. Yuan
UnionPay
T. Zhou
G. Fioccola
Y. Wang
Huawei
January 10, 2022

BGP SR Policy Extensions to Enable IFIT
draft-ietf-idr-sr-policy-ifit-03

Abstract

Segment Routing (SR) policy is a set of candidate SR paths consisting of one or more segment lists and necessary path attributes. It enables instantiation of an ordered list of segments with a specific intent for traffic steering. In-situ Flow Information Telemetry (IFIT) refers to network OAM data plane on-path telemetry techniques, in particular the most popular are In-situ OAM (IOAM) and Alternate Marking. This document defines extensions to BGP to distribute SR policies carrying IFIT information. So that IFIT methods can be enabled automatically when the SR policy is applied.

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

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

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

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

This Internet-Draft will expire on July 14, 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 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. Motivation	3
3. IFIT methods for SR Policy	4
4. IFIT Attributes in SR Policy	5
5. IFIT Attributes Sub-TLV	6
5.1. IOAM Pre-allocated Trace Option Sub-TLV	8
5.2. IOAM Incremental Trace Option Sub-TLV	9
5.3. IOAM Directly Export Option Sub-TLV	9
5.4. IOAM Edge-to-Edge Option Sub-TLV	10
5.5. Enhanced Alternate Marking (EAM) sub-TLV	11
6. SR Policy Operations with IFIT Attributes	12
7. IANA Considerations	12
8. Security Considerations	13
9. Acknowledgements	14
10. References	14
10.1. Normative References	14
10.2. Informative References	16
Appendix A.	16
Authors' Addresses	16

1. Introduction

Segment Routing (SR) policy [I-D.ietf-spring-segment-routing-policy] is a set of candidate SR paths consisting of one or more segment lists and necessary path attributes. It enables instantiation of an ordered list of segments with a specific intent for traffic steering.

In-situ Flow Information Telemetry (IFIT) denotes a family of flow-oriented on-path telemetry techniques (e.g. IOAM, Alternate

Marking), which can provide high-precision flow insight and real-time network issue notification (e.g., jitter, latency, packet loss). In particular, IFIT refers to network OAM (Operations, Administration, and Maintenance) data plane on-path telemetry techniques, including In-situ OAM (IOAM) [I-D.ietf-ippm-ioam-data] and Alternate Marking [RFC8321]. It can provide flow information on the entire forwarding path on a per-packet basis in real time.

An automatic network requires the Service Level Agreement (SLA) monitoring on the deployed service. So that the system can quickly detect the SLA violation or the performance degradation, hence to change the service deployment. For this reason, the SR policy native IFIT can facilitate the closed loop control and enable the automation of SR service.

This document defines extensions to Border Gateway Protocol (BGP) to distribute SR policies carrying IFIT information. So that IFIT behavior can be enabled automatically when the SR policy is applied.

This BGP extension allows to signal the IFIT capabilities together with the SR-policy. In this way IFIT methods are automatically activated and running. The flexibility and dynamicity of the IFIT applications are given by the use of additional functions on the controller and on the network nodes, but this is out of scope here.

IFIT is a solution focusing on network domains according to [RFC8799] that introduces the concept of specific domain solutions. A network domain consists of a set of network devices or entities within a single administration. As mentioned in [RFC8799], for a number of reasons, such as policies, options supported, style of network management and security requirements, it is suggested to limit applications including the emerging IFIT techniques to a controlled domain. Hence, the IFIT methods MUST be typically deployed in such controlled domains.

2. Motivation

IFIT Methods are being introduced in multiple protocols and below is a proper picture of the relevant documents for Segment Routing. Indeed the IFIT methods are becoming mature for Segment Routing over the MPLS data plane (SR-MPLS) and Segment Routing over IPv6 data plane (SRv6), that is the main focus of this draft:

IOAM: the reference documents for the data plane are
[I-D.ietf-ippm-ioam-ipv6-options] for SRv6 and
[I-D.gandhi-mpls-ioam-sr] for SR-MPLS.

Alternate Marking: the reference documents for the data plane are [I-D.ietf-6man-ipv6-alt-mark] for SRv6 and [I-D.ietf-mpls-rfc6374-sfl], [I-D.gandhi-mpls-rfc6374-sr] for SR-MPLS.

The definition of these data plane IFIT methods for SR-MPLS and SRv6 imply requirements for various routing protocols, such as BGP, and this document aims to define BGP extensions to distribute SR policies carrying IFIT information. This allows to signal the IFIT capabilities so IFIT methods are automatically configured and ready to run when the SR Policy candidate paths are distributed through BGP.

It is to be noted that, for PCEP (Path Computation Element Communication Protocol), [I-D.chen-pce-pcep-ifit] proposes the extensions to PCEP to distribute paths carrying IFIT information and therefore to enable IFIT methods for SR policy too.

3. IFIT methods for SR Policy

In-situ Operations, Administration, and Maintenance (IOAM) [I-D.ietf-ippm-ioam-data] records operational and telemetry information in the packet while the packet traverses a path between two points in the network. In terms of the classification given in RFC 7799 [RFC7799] IOAM could be categorized as Hybrid Type 1. IOAM mechanisms can be leveraged where active OAM do not apply or do not offer the desired results. When SR policy enables the IOAM, the IOAM header will be inserted into every packet of the traffic that is steered into the SR paths.

The Alternate Marking [RFC8321] technique is an hybrid performance measurement method, per RFC 7799 [RFC7799] classification of measurement methods. Because this method is based on marking consecutive batches of packets. It can be used to measure packet loss, latency, and jitter on live traffic.

This document aims to define the control plane. While the relevant documents for the data plane application of IOAM and Alternate Marking are respectively [I-D.ietf-ippm-ioam-ipv6-options] and [I-D.ietf-6man-ipv6-alt-mark] for Segment Routing over IPv6 data plane (SRv6), [I-D.ietf-mpls-rfc6374-sfl], [I-D.gandhi-mpls-rfc6374-sr] and [I-D.gandhi-mpls-ioam-sr] for Segment Routing over the MPLS data plane (SR-MPLS).

4. IFIT Attributes in SR Policy

As defined in [I-D.ietf-idr-segment-routing-te-policy], a new SAFI is defined (the SR Policy SAFI with codepoint 73) as well as a new NLRI. The NLRI contains the SR Policy candidate path and, according to [I-D.ietf-idr-segment-routing-te-policy], the content of the SR Policy Candidate Path is encoded in the Tunnel Encapsulation Attribute defined in [I-D.ietf-idr-tunnel-encaps] using a new Tunnel-Type called SR Policy Type with codepoint 15. 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

 SRv6 Binding SID

 Preference

 Priority

 Policy Name

 Policy Candidate Path Name

 Explicit NULL Label Policy (ENLP)

 Segment List

 Weight

 Segment

 Segment

 ...

 ...

A candidate path includes multiple SR paths, each of which is specified by a segment list. IFIT can be applied to the candidate path, so that all the SR paths can be monitored in the same way. The new SR Policy encoding structure is expressed as below:

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>

Attributes:

 Tunnel Encaps Attribute (23)

 Tunnel Type: SR Policy

 Binding SID

 SRv6 Binding SID

 Preference

 Priority

 Policy Name

 Policy Candidate Path Name

 Explicit NULL Label Policy (ENLP)

 IFIT Attributes

 Segment List

 Weight

 Segment

 Segment

 ...

 ...

IFIT attributes can be attached at the candidate path level as sub-TLVs. There may be different IFIT tools. The following sections will describe the requirement and usage of different IFIT tools, and define the corresponding sub-TLV encoding in BGP.

Once the IFIT attributes are signalled, if a packet arrives at the headend and, based on the types of steering described in [I-D.ietf-spring-segment-routing-policy], it may get steered into an SR Policy where IFIT methods are applied. Therefore it will be managed consequently with the corresponding IOAM or Alternate Marking information according to the enabled IFIT methods.

Note that the IFIT attributes here described can also be generalized and included as sub-TLVs for other SAFIs and NLRIs.

5. IFIT Attributes Sub-TLV

The format of the IFIT Attributes Sub-TLV is defined as follows:

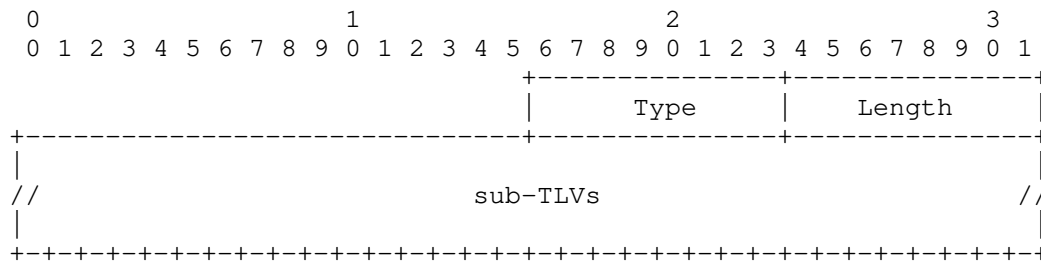


Fig. 1 IFIT Attributes Sub-TLV

Where:

Type: to be assigned by IANA.

Length: the total length of the value field not including Type and Length fields.

sub-TLVs currently defined:

- * IOAM Pre-allocated Trace Option Sub-TLV,
- * IOAM Incremental Trace Option Sub-TLV,
- * IOAM Directly Export Option Sub-TLV,
- * IOAM Edge-to-Edge Option Sub-TLV,
- * Enhanced Alternate Marking (EAM) sub-TLV.

The presence of the IFIT Attributes Sub-TLV implies support of IFIT methods (IOAM and/or Alternate Marking). It is worth mentioning that IOAM and Alternate Marking can be activated one at a time or can coexist; so it is possible to have only IOAM or only Alternate Marking enabled as Sub-TLVs. The sub-TLVs currently defined for IOAM and Alternate Marking are detailed in the next sections.

In case of empty IFIT Attributes Sub-TLV, i.e. no further IFIT sub-TLV and Length=0, IFIT methods will not be activated. If two conflicting IOAM sub-TLVs are present (e.g. Pre-allocated Trace Option and Incremental Trace Option) it means that they are not usable and none of the two methods will be activated. The same applies if there is more than one instance of the sub-TLV of the same type. Anyway the validation of the individual fields of the IFIT Attributes sub-TLVs are handled by the SRPM (SR Policy Module).

The process of stopping IFIT methods can be done by setting empty IFIT Attributes Sub-TLV, while, for modifying IFIT methods parameters, the IFIT Attributes Sub-TLVs can be updated accordingly. Additionally the backward compatibility is guaranteed, since an implementation that does not understand IFIT Attributes Sub-TLV can simply ignore it.

5.1. IOAM Pre-allocated Trace Option Sub-TLV

The IOAM tracing data is expected to be collected at every node that a packet traverses to ensure visibility into the entire path a packet takes within an IOAM domain. The preallocated tracing option will create pre-allocated space for each node to populate its information.

The format of IOAM pre-allocated trace option sub-TLV is defined as follows:

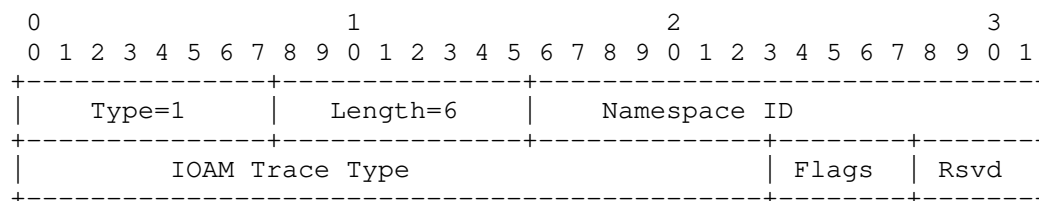


Fig. 2 IOAM Pre-allocated Trace Option Sub-TLV

Where:

Type: 1 (to be assigned by IANA).

Length: 6, it is the total length of the value field (not including Type and Length fields).

Namespace ID: A 16-bit identifier of an IOAM-Namespace. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

IOAM Trace Type: A 24-bit identifier which specifies which data types are used in the node data list. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

Flags: A 4-bit field. The definition is the same as described in [I-D.ietf-ippm-ioam-flags] and section 4.4 of [I-D.ietf-ippm-ioam-data].

Rsvd: A 4-bit field reserved for further usage. It MUST be zero and ignored on receipt.

5.2. IOAM Incremental Trace Option Sub-TLV

The incremental tracing option contains a variable node data fields where each node allocates and pushes its node data immediately following the option header.

The format of IOAM incremental trace option sub-TLV is defined as follows:

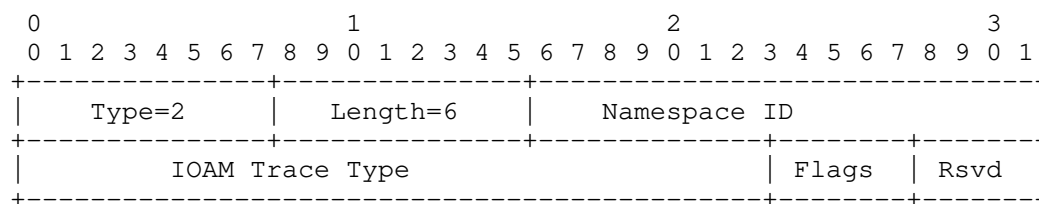


Fig. 3 IOAM Incremental Trace Option Sub-TLV

Where:

Type: 2 (to be assigned by IANA).

Length: 6, it is the total length of the value field (not including Type and Length fields).

All the other fields definition is the same as the pre-allocated trace option sub-TLV in section 4.1.

5.3. IOAM Directly Export Option Sub-TLV

IOAM directly export option is used as a trigger for IOAM data to be directly exported to a collector without being pushed into in-flight data packets.

The format of IOAM directly export option sub-TLV is defined as follows:

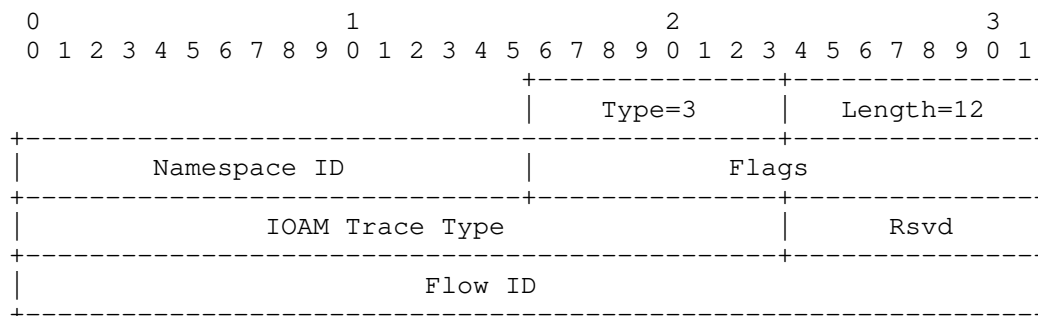


Fig. 4 IOAM Directly Export Option Sub-TLV

Where:

Type: 3 (to be assigned by IANA).

Length: 12, it is the total length of the value field (not including Type and Length fields).

Namespace ID: A 16-bit identifier of an IOAM-Namespace. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

Flags: A 16-bit field. The definition is the same as described in section 3.2 of [I-D.ietf-ippm-ioam-direct-export].

IOAM Trace Type: A 24-bit identifier which specifies which data types are used in the node data list. The definition is the same as described in section 4.4 of [I-D.ietf-ippm-ioam-data].

Rsvd: A 4-bit field reserved for further usage. It MUST be zero and ignored on receipt.

Flow ID: A 32-bit flow identifier. The definition is the same as described in section 3.2 of [I-D.ietf-ippm-ioam-direct-export].

5.4. IOAM Edge-to-Edge Option Sub-TLV

The IOAM edge to edge option is to carry data that is added by the IOAM encapsulating node and interpreted by IOAM decapsulating node.

The format of IOAM edge-to-edge option sub-TLV is defined as follows:

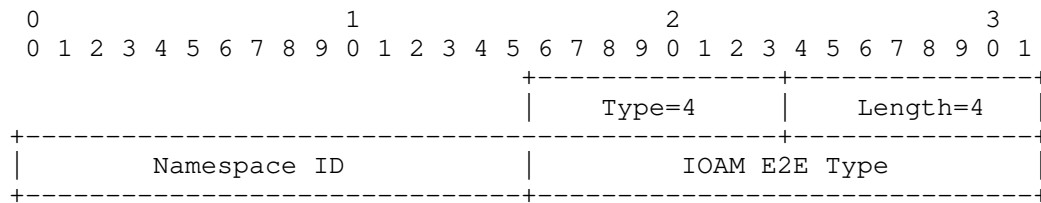


Fig. 5 IOAM Edge-to-Edge Option Sub-TLV

Where:

Type: 4 (to be assigned by IANA).

Length: 4, it is the total length of the value field (not including Type and Length fields).

Namespace ID: A 16-bit identifier of an IOAM-Namespaces. The definition is the same as described in section 4.6 of [I-D.ietf-ippm-ioam-data].

IOAM E2E Type: A 16-bit identifier which specifies which data types are used in the E2E option data. The definition is the same as described in section 4.6 of [I-D.ietf-ippm-ioam-data].

5.5. Enhanced Alternate Marking (EAM) sub-TLV

The format of Enhanced Alternate Marking (EAM) sub-TLV is defined as follows:

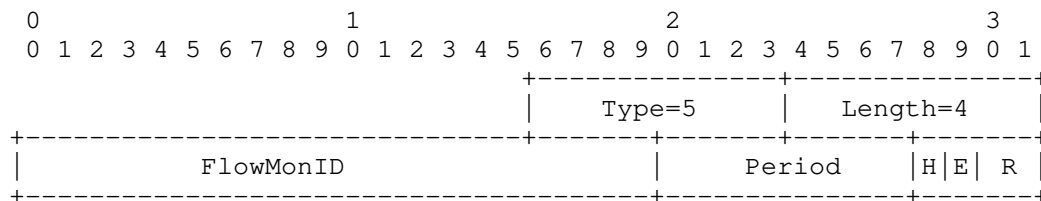


Fig. 6 Enhanced Alternate Marking Sub-TLV

Where:

Type: 5 (to be assigned by IANA).

Length: 4, it is the total length of the value field (not including Type and Length fields).

FlowMonID: A 20-bit identifier to uniquely identify a monitored flow within the measurement domain. The definition is the same as described in section 5.3 of [I-D.ietf-6man-ipv6-alt-mark].

Period: Time interval between two alternate marking period. The unit is second.

H: A flag indicating that the measurement is Hop-By-Hop.

E: A flag indicating that the measurement is end to end.

R: A 2-bit field reserved for further usage. It MUST be zero and ignored on receipt.

6. SR Policy Operations with IFIT Attributes

The details of SR Policy installation and use are specified in [I-D.ietf-spring-segment-routing-policy]. This document complements SR Policy Operations described in [I-D.ietf-idr-segment-routing-te-policy] by adding the IFIT Attributes.

The operations described in [I-D.ietf-idr-segment-routing-te-policy] are always valid. The only difference is the addition of IFIT Attributes Sub-TLVs for the SR Policy NLRI, that can affect its acceptance by a BGP speaker, but the implementation MAY provide an option for ignoring the unrecognized or unsupported IFIT sub-TLVs. SR Policy NLRIs that have been determined acceptable, usable and valid can be evaluated for propagation, including the IFIT information.

The error handling actions are also described in [I-D.ietf-idr-segment-routing-te-policy], indeed A BGP Speaker MUST perform the syntactic validation of the SR Policy NLRI to determine if it is malformed, including the TLVs/sub-TLVs. In case of any error detected, either at the attribute or its TLV/sub-TLV level, the "treat-as-withdraw" strategy MUST be applied.

The validation of the IFIT Attributes sub-TLVs introduced in this document MUST be performed to determine if they are malformed or invalid. The validation of the individual fields of the IFIT Attributes sub-TLVs are handled by the SRPM (SR Policy Module).

7. IANA Considerations

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

Codepoint	Description	Reference
TBD1	IFIT Attributes Sub-TLV	This document

This document requests creation of a new registry called "IFIT Attributes Sub-TLVs". The allocation policy of this registry is "Specification Required" according to RFC 8126 [RFC8126].

The following initial Sub-TLV codepoints are assigned by this document:

Value	Description	Reference
1	IOAM Pre-allocated Trace Option Sub-TLV	This document
2	IOAM Incremental Trace Option Sub-TLV	This document
3	IOAM Directly Export Option Sub-TLV	This document
4	IOAM Edge-to-Edge Option Sub-TLV	This document
5	Enhanced Alternate Marking Sub-TLV	This document

8. Security Considerations

The security mechanisms of the base BGP security model apply to the extensions described in this document as well. See the Security Considerations section of [I-D.ietf-idr-segment-routing-te-policy].

SR operates within a trusted SR domain RFC 8402 [RFC8402] and its security considerations also apply to BGP sessions when carrying SR Policy information. The isolation of BGP SR Policy SAFI peering sessions may be used to ensure that the SR Policy information is not advertised outside the SR domain. Additionally, only trusted nodes (that include both routers and controller applications) within the SR domain must be configured to receive such information.

Implementation of IFIT methods (IOAM and Alternate Marking) are mindful of security and privacy concerns, as explained in [I-D.ietf-ippm-ioam-data] and RFC 8321 [RFC8321]. Anyway incorrect IFIT parameters in the BGP extension SHOULD NOT have an adverse effect on the SR Policy as well as on the network, since it affects only the operation of the telemetry methodology.

IFIT data MUST be propagated in a limited domain in order to avoid malicious attacks and solutions to ensure this requirement are

respectively discussed in [I-D.ietf-ippm-ioam-data] and [I-D.ietf-6man-ipv6-alt-mark].

IFIT methods (IOAM and Alternate Marking) are applied within a controlled domain where the network nodes are locally administered. A limited administrative domain provides the network administrator with the means to select, monitor and control the access to the network, making it a trusted domain also for the BGP extensions defined in this document.

9. Acknowledgements

The authors of this document would like to thank Ketan Talaulikar, Joel Halpern, Jie Dong for their comments and review of this document.

10. References

10.1. Normative References

[I-D.ietf-6man-ipv6-alt-mark]

Fioccola, G., Zhou, T., Cociglio, M., Qin, F., and R. Pang, "IPv6 Application of the Alternate Marking Method", draft-ietf-6man-ipv6-alt-mark-12 (work in progress), October 2021.

[I-D.ietf-idr-segment-routing-te-policy]

Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-14 (work in progress), November 2021.

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

Patel, K., Velde, G. V. D., Sangli, S. R., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-22 (work in progress), January 2021.

[I-D.ietf-ippm-ioam-data]

Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In-situ OAM", draft-ietf-ippm-ioam-data-17 (work in progress), December 2021.

[I-D.ietf-ippm-ioam-direct-export]

Song, H., Gafni, B., Zhou, T., Li, Z., Brockners, F., Bhandari, S., Sivakolundu, R., and T. Mizrahi, "In-situ OAM Direct Exporting", draft-ietf-ippm-ioam-direct-export-07 (work in progress), October 2021.

- [I-D.ietf-ippm-ioam-flags]
Mizrahi, T., Brockners, F., Bhandari, S., Sivakolundu, R., Pignataro, C., Kfir, A., Gafni, B., Spiegel, M., and J. Lemon, "In-situ OAM Loopback and Active Flags", draft-ietf-ippm-ioam-flags-07 (work in progress), October 2021.
- [I-D.ietf-ippm-ioam-ipv6-options]
Bhandari, S. and F. Brockners, "In-situ OAM IPv6 Options", draft-ietf-ippm-ioam-ipv6-options-06 (work in progress), July 2021.
- [I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", draft-ietf-spring-segment-routing-policy-14 (work in progress), October 2021.
- [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>>.
- [RFC7799] Morton, A., "Active and Passive Metrics and Methods (with Hybrid Types In-Between)", RFC 7799, DOI 10.17487/RFC7799, May 2016, <<https://www.rfc-editor.org/info/rfc7799>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.
- [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>>.
- [RFC8321] Fioccola, G., Ed., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., and T. Mizrahi, "Alternate-Marking Method for Passive and Hybrid Performance Monitoring", RFC 8321, DOI 10.17487/RFC8321, January 2018, <<https://www.rfc-editor.org/info/rfc8321>>.
- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.

[RFC8799] Carpenter, B. and B. Liu, "Limited Domains and Internet Protocols", RFC 8799, DOI 10.17487/RFC8799, July 2020, <<https://www.rfc-editor.org/info/rfc8799>>.

10.2. Informative References

- [I-D.chen-pce-pcep-ifit]
Yuan, H., Zhou, T., Li, W., Fioccola, G., and Y. Wang, "Path Computation Element Communication Protocol (PCEP) Extensions to Enable IFIT", draft-chen-pce-pcep-ifit-04 (work in progress), July 2021.
- [I-D.gandhi-mpls-ioam-sr]
Gandhi, R., Ali, Z., Filsfils, C., Brockners, F., Wen, B., and V. Kozak, "MPLS Data Plane Encapsulation for In-situ OAM Data", draft-gandhi-mpls-ioam-sr-06 (work in progress), February 2021.
- [I-D.gandhi-mpls-rfc6374-sr]
Gandhi, R., Filsfils, C., Voyer, D., Salsano, S., and M. Chen, "Performance Measurement Using RFC 6374 for Segment Routing Networks with MPLS Data Plane", draft-gandhi-mpls-rfc6374-sr-05 (work in progress), June 2020.
- [I-D.ietf-mpls-rfc6374-sfl]
Bryant, S., Swallow, G., Chen, M., Fioccola, G., and G. Mirsky, "RFC6374 Synonymous Flow Labels", draft-ietf-mpls-rfc6374-sfl-10 (work in progress), March 2021.

Appendix A.

Authors' Addresses

Fengwei Qin
China Mobile
No. 32 Xuanwumenxi Ave., Xicheng District
Beijing
China

Email: qinfengwei@chinamobile.com

Hang Yuan
UnionPay
1899 Gu-Tang Rd., Pudong
Shanghai
China

Email: yuanhang@unionpay.com

Tianran Zhou
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: zhoutianran@huawei.com

Giuseppe Fioccola
Huawei
Riesstrasse, 25
Munich
Germany

Email: giuseppe.fioccola@huawei.com

Yali Wang
Huawei
156 Beiqing Rd., Haidian District
Beijing
China

Email: wangyalil1@huawei.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: 22 September 2022

A. Wang
China Telecom
B. Khasanov
Yandex LLC
S. Fang
R. Tan
Huawei Technologies, Co., Ltd
C. Zhu
ZTE Corporation
21 March 2022

PCEP Extension for Native IP Network
draft-ietf-pce-pcep-extension-native-ip-18

Abstract

This document defines the Path Computation Element Communication Protocol (PCEP) extension for Central Control Dynamic Routing (CCDR) based application in Native IP network. The scenario and framework of CCDR in native IP is described in [RFC8735] and [RFC8821]. This draft describes the key information that is transferred between Path Computation Element (PCE) and Path Computation Clients (PCC) to accomplish the End to End (E2E) traffic assurance in Native IP network under central control mode.

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 22 September 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
2. Conventions used in this document	3
3. Terminology	3
4. Capability Advertisemnt	4
4.1. Open message	4
5. PCEP messages	4
5.1. The PCInitiate message	5
5.2. The PCRpt message	6
6. PCECC Native IP TE Procedures	7
6.1. BGP Session Establishment Procedures	7
6.2. Explicit Route Establish Procedures	9
6.3. BGP Prefix Advertisement Procedures	12
7. New PCEP Objects	14
7.1. CCI Object	14
7.2. BGP Peer Info Object	15
7.3. Explicit Peer Route Object	17
7.4. Peer Prefix Advertisement Object	20
8. End to End Path Protection	21
9. Re-Delegation and Clean up	21
10. BGP Considerations	22
11. New Error-Types and Error-Values Defined	22
12. Deployment Considerations	23
13. Implementation Status	24
13.1. Proof of Concept based on ODL	24
14. Security Considerations	25
15. IANA Considerations	25
15.1. Path Setup Type Registry	25
15.2. PCECC-CAPABILITY sub-TLV's Flag field	25
15.3. PCEP Object Types	25
15.4. PCEP-Error Object	26
16. Contributor	27
17. Acknowledgement	27
18. Normative References	27
Authors' Addresses	29

1. Introduction

Generally, Multiprotocol Label Switching Traffic Engineering (MPLS-TE) requires the corresponding network devices support Multiprotocol Label Switching (MPLS) or Resource ReSerVation Protocol (RSVP)/Label Distribution Protocol (LDP) technologies to assure the End-to-End (E2E) traffic performance. In Segment Routing either IGP extensions or BGP are used to steer a packet through an SR Policy instantiated as an ordered list of instructions called "segments". But in native IP network, there will be no such signaling protocol to synchronize the action among different network devices. It is necessary to use the central control mode that described in [RFC8283] to correlate the forwarding behavior among different network devices. [RFC8821] describes the architecture and solution philosophy for the E2E traffic assurance in Native IP network via Multi Border Gateway Protocol (BGP) solution. This draft describes the corresponding Path Computation Element Communication Protocol (PCEP) extensions to transfer the key information about BGP peer info, peer prefix advertisement and the explicit peer route on on-path routers.

2. Conventions used in this document

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.

3. Terminology

This document uses the following terms defined in [RFC5440]: PCE, PCEP

The following terms are defined in this document:

- * CCDR: Central Control Dynamic Routing
- * E2E: End to End
- * BPI: BGP Peer Info
- * EPR: Explicit Peer Route
- * PPA: Peer Prefix Advertisement
- * QoS: Quality of Service

4. Capability Advertisement

4.1. Open message

During the PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of Native IP extensions.

This document defines a new Path Setup Type (PST) [RFC8408] for Native-IP, as follows:

- * PST = TBD1: Path is a Native IP path as per [RFC8821].

A PCEP speaker MUST indicate its support of the function described in this document by sending a PATH-SETUP-TYPE-CAPABILITY TLV in the OPEN object with this new PST included in the PST list.

[RFC9050] defined the PCECC-CAPABILITY sub-TLV to exchange information about their PCECC capability. A new flag is defined in PCECC-CAPABILITY sub-TLV for Native IP:

N (NATIVE-IP-TE-CAPABILITY - 1 bit - TBD2): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker is capable for TE in Native IP network as specified in this document. The flag MUST be set by both the PCC and PCE in order to support this extension.

If a PCEP speaker receives the PATH-SETUP-TYPE-CAPABILITY TLV with the newly defined path setup type, but without the N bit set in PCECC-CAPABILITY sub-TLV, it MUST:

- * Send a PCErr message with Error-Type=10(Reception of an invalid object) and Error-Value TBD3(PCECC NATIVE-IP-TE-CAPABILITY bit is not set).
- * Terminate the PCEP session

5. PCEP messages

PCECC Native IP TE solution utilizing the existing PCE LSP Initiate Request message(PCInitiate) [RFC8281], and PCE Report message(PCRppt) [RFC8281] to accomplish the multi BGP sessions establishment, E2E TE path deployment, and route prefixes advertisement among different BGP sessions. A new PST for Native-IP is used to indicate the path setup based on TE in Native IP networks.

The extended PCInitiate message described in [RFC9050] is used to download or cleanup central controller's instructions (CCIs). [RFC9050] specifies an object called CCI for the encoding of central controller's instructions. This document specifies a new CCI object-

type for Native IP. The PCEP messages are extended in this document to handle the PCECC operations for Native IP. Three new PCEP Objects (BGP Peer Info (BPI) Object, Explicit Peer Route (EPR) Object and Peer Prefix Advertisement (PPA) Object) are defined in this document. Refer to Section 7 for detail object definitions.

5.1. The PCInitiate message

The PCInitiate Message defined in [RFC8281] and extended in [RFC9050] is further extended to support Native-IP CCI.

The format of the extended PCInitiate message is as follows:

```

<PCInitiate Message> ::= <Common Header>
                           <PCE-initiated-lsp-list>
Where:
  <Common Header> is defined in [RFC5440]

  <PCE-initiated-lsp-list> ::= <PCE-initiated-lsp-request>
                               [<PCE-initiated-lsp-list>]

  <PCE-initiated-lsp-request> ::=
    (<PCE-initiated-lsp-instantiation>|
     <PCE-initiated-lsp-deletion>|
     <PCE-initiated-lsp-central-control>)

  <PCE-initiated-lsp-central-control> ::= <SRP>
                                           <LSP>
                                           (<cci-list>|
                                           ((<BPI>|<EPR>|<PPA>)
                                           <CCI>))

  <cci-list> ::= <CCI>
                [<cci-list>]

```

Where:

```

  <cci-list> is as per
  [I-D.ietf-pce-pcep-extension-for-pce-controller].
  <PCE-initiated-lsp-instantiation> and
  <PCE-initiated-lsp-deletion> are as per
  [RFC8281].

```

The LSP and SRP objects are defined in [RFC8231].

When PCInitiate message is used create Native IP instructions, the SRP, LSP and CCI objects MUST be present. The error handling for missing SRP, LSP or CCI object is as per [RFC9050]. Further only one of BPI, EPR, or PPA object MUST be present. The PLSP-ID within the

LSP object should be set by PCC uniquely according to the Symbolic Path Name TLV that included in the CCI object. The Symbolic Path Name is used by the PCE/PCC to identify uniquely the E2E native IP TE path.

If none of them are present, the receiving PCC MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=TBD4 (Native IP object missing). If there are more than one of BPI, EPR or PPA object are presented, the receiving PCC MUST send a PCErr message with Error-type=19 (Invalid Operation) and Error-value=TBD5 (Only one of the BPI, EPR or PPA object can be included in this message).

To cleanup the SRP object must set the R (remove) bit.

5.2. The PCRpt message

The PCRpt message is used to acknowledge the Native-IP instructions received from the central controller (PCE).

The format of the PCRpt message is as follows:

```
<PCRpt Message> ::= <Common Header>
                        <state-report-list>
```

Where:

```
<state-report-list> ::= <state-report> [<state-report-list>]
```

```
<state-report> ::= (<lsp-state-report> |
                    <central-control-report>)
```

```
<lsp-state-report> ::= [<SRP>]
                        <LSP>
                        <path>
```

```
<central-control-report> ::= [<SRP>]
                              <LSP>
                              (<cci-list> |
                               ((<BPI> | <EPR> | <PPA>)
                                <CCI>))
```

Where:

<path> is as per [RFC8231] and the LSP and SRP object are also defined in [RFC8231].

The error handling for missing CCI object is as per [RFC9050]. Further only one of BPI, EPR, or PPA object MUST be present.

If none of them are present, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=TBD4 (Native IP object missing). If there are more than one of BPI, EPR or PPA object are presented, the receiving PCE MUST send a PCErr message with Error-type=19(Invalid Operation) and Error-value=TBD5(Only one of the BPI, EPR or PPA object can be included in this message).

6. PCECC Native IP TE Procedures

The detail procedures for the TE in native IP environment are described in the following sections.

6.1. BGP Session Establishment Procedures

The PCInitiate message can be used to configure the parameters for a BGP peer session using the PCInitiate and PCRpt message pair. This pair of PCE messages is exchanged with a PCE function attached to each BGP peer which needs to be configured. After the BGP peer session has been configured via this pair of PCE messages the BGP session establishment process operates in a normal fashion. All BGP peers are configured for peer to peer communication whether the peers are E-BGP peers or I-BGP peers. One of the IBGP topologies requires that multiple I-BGPs peers operate in a route-reflector I-BGP peer topology. The example below shows two I-BGP route reflector clients interacting with one Route Reflector (RR), but Route Reflector topologies may have up to 100s of clients. Centralized configuration via PCE provides mechanisms to scale auto-configuration of small and large topologies.

The PCInitiate message should be sent to PCC which acts as BGP router and/or route reflector(RR).

The route reflector topology for a single AS is shown in Figure 1. The BGP routers R1, R3, and R7 are within a single AS. R1 and R7 are BGP router-reflector clients, and R3 is a Route Reflector. The PCInitiate message should be sent all of the BGP routers that need to be configured R1 (M3), R3 (M2 & M3), and R7 (M4).

PCInitiate message creates an auto-configuration function for these BGP peers providing the indicated Peer AS and the Local/Peer IP Address.

When PCC receives the BPI and CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC should try to establish the BGP session with the indicated Peer AS and Local/Peer IP address.

When PCC clears successfully the specified BGP session, it should report the result via the PCRpt message, with the BPI object included, and the corresponding SRP and CCI object.

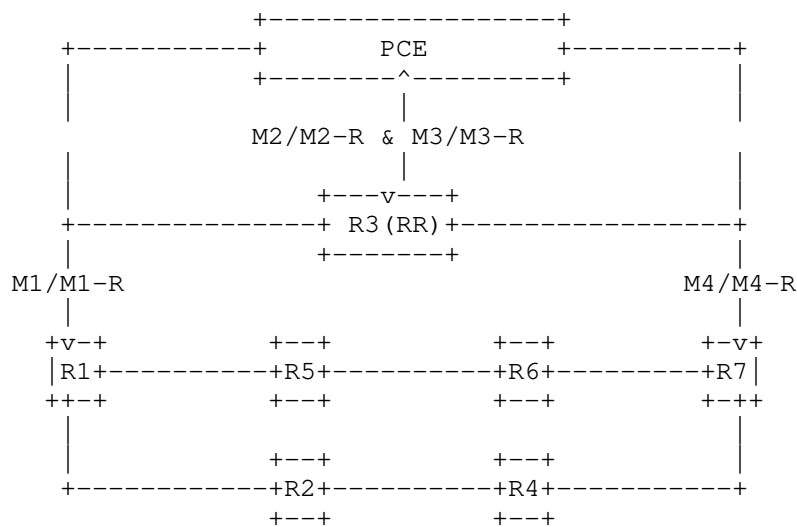


Table 1: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) BPI Object (Local_IP=R1_A, Peer_IP=R3_A)
M2 M2-R	PCE/R3	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) BPI Object (Local_IP=R3_A, Peer_IP=R1_A)
M3 M3-R	PCE/R3	PCInitiate PCRpt	CC-ID=X3 (Symbolic Path Name=Class A) BPI Object (Local_IP=R3_A, Peer_IP=R7_A)
M4 M4-R	PCE/R7	PCInitiate PCRpt	CC-ID=X4 (Symbolic Path Name=Class A) BPI Object (Local_IP=R7_A, Peer_IP=R3_A)

If the PCC cannot establish the BGP session that required by this object, it should report the error values via PCErr message with the newly defined error type (Error-type=TBD6) and error value (Error-value=TBD7, Peer AS not match; or Error-Value=TBD8, Peer IP can't be reached), which is indicated in Section 11

If the Local IP Address or Peer IP Address within BPI object is used in other existing BGP sessions, the PCC should report such error situation via PCErr message with Err-type=TBD6 and error value (Error-value=TBD9, Local IP is in use; Error-value=TBD10, Remote IP is in use).

6.2. Explicit Route Establish Procedures

The explicit route establishment procedures can be used to install a route via PCE in the PCC/BGP Peer, using PCInitiate and PCRpt message pair. Although the BGP policy might redistribute the routes installed by explicit route, the PCE-BGP implementation needs to prohibit the redistribution of the explicit route. PCE explicit routes operate similar to static routes installed by network management protocols (netconf/restconf) but the routes are associated with the PCE routing module. Explicit route installations (like NM static routes) must carefully install and uninstall static routes in an specific order so that the pathways are established without loops.

The PCInitiate message should be sent to the on-path routers respectively. In the example, for explicit route from R1 to R7, the PCInitiate message should be sent to R1 (M1), R2 (M2) and R4 (M3), as shown in Figure 2. For explicit route from R7 to R1, the PCInitiate message should be sent to R7 (M1), R4 (M2) and R2 (M3), as shown in Figure 3.

When PCC clear successfully the explicit route that indicated by this object, it should report the result via the PCRpt message, with the EPR object included, and the corresponding SRP and CCI object.

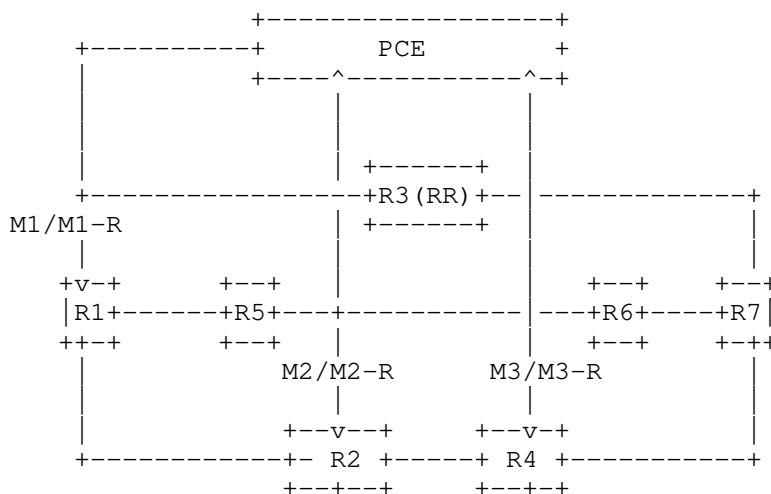


Table 2: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) EPR Object (Peer Address=R7_A, Next Hop=R2_A)
M2 M2-R	PCE/R2	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) EPR Object (Peer Address=R7_A, Next Hop=R4_A)
M3 M3-R	PCE/R4	PCInitiate PCRpt	CC-ID=X3 (Symbolic Path Name=Class A) EPR Object (Peer Address=R7_A, Next Hop=R7_A)

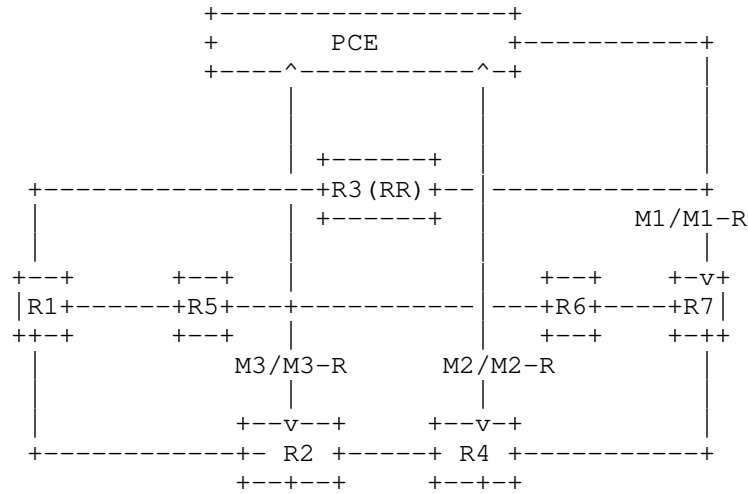


Figure 3: Explicit Route Establish Procedures (From R7 to R1)

The message number, message peers, message type and message key parameters in the above figures are shown in below table:

Table 3: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R7	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) EPR Object (Peer Address=R1_A, Next Hop=R4_A)
M2 M2-R	PCE/R4	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) EPR Object (Peer Address=R1_A, Next Hop=R2_A)
M3 M3-R	PCE/R2	PCInitiate PCRpt	CC-ID=X3 (Symbolic Path Name=Class A) EPR Object (Peer Address=R1_A, Next Hop=R1_A)

In order to avoid the transient loop during the deploy of explicit peer route, the EPR object should be sent to the PCCs in the reverse order of the E2E path. To remove the explicit peer route, the EPR object should be sent to the PCCs in the same order of E2E path.

To accomplish ECMP effects, the PCE can send multiple EPR objects to the same node, with the same route priority and peer address value but different next hop addresses.

The PCC should verify that the next hop address is reachable. Upon the error occurs, the PCC SHOULD send the corresponding error via PCErr message, with an error information (Error-type=TBD6, Error-value=TBD12, Explicit Peer Route Error) that defined in Section 11.

When the peer info is not the same as the peer info that indicated in BPI object in PCC for the same path that is identified by Symbolic Path Name TLV, an error (Error-type=TBD6, Error-value=17, EPR/BPI Peer Info mismatch) should be reported via the PCErr message.

6.3. BGP Prefix Advertisement Procedures

The detail procedures for BGP prefix advertisement are shown below, using PCInitiate and PCRpt message pair.

The PCInitiate message should be sent to PCC that acts as BGP peer router only. In the example, it should be sent to R1(M1) or R7(M2) respectively.

When PCC receives the PPA and the CCI object (with the R bit set to 0 in SRP object) in PCInitiate message, the PCC should send the prefixes indicated in this object to the appointed BGP peer.

When PCC sends successfully the prefixes to the appointed BGP peer, it should report the result via the PCRpt messages, with PPA object and the corresponding SRP and CCI object included.

When PCC receives the PPA and the CCI object with the R bit set to 1 in SRP object in PCInitiate message, the PCC should withdraw the prefixes advertisement to the peer that indicated by this object.

When PCC withdraws successfully the prefixes that indicated by this object, it should report the result via the PCRpt message, with the PPA object included, and the corresponding SRP and CCI object.

The allowed AFI/SAFI for the IPv4 BGP session should be 1/1(IPv4 prefix) and the allowed AFI/SAFI for the IPv6 BGP session should be 2/1(IPv6 prefix). If mismatch occur, an error(Error-type=TBD6, Error-value=TBD18, BPI/PPR address family mismatch) should be reported via PCErr message.

When the peer info is not the same as the peer info that indicated in BPI object in PCC for the same path that is identified by Symbolic Path Name TLV, an error (Error-type=TBD6, Error-value=TBD19, PPA/BPI peer info mismatch) should be reported via the PCErr message.

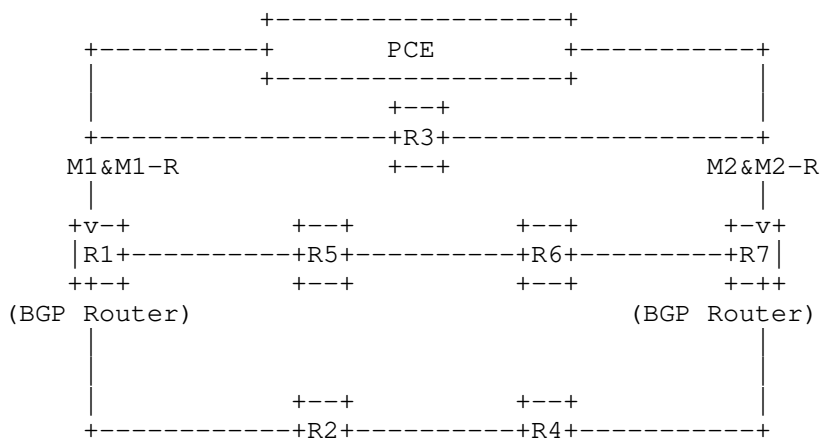


Figure 4: BGP Prefix Advertisement Procedures

Table 4: Message Information

No.	Peers	Type	Message Key Parameters
M1 M1-R	PCE/R1	PCInitiate PCRpt	CC-ID=X1 (Symbolic Path Name=Class A) PPA Object (Peer IP=R7_A, Prefix=1_A)
M2 M2-R	PCE/R7	PCInitiate PCRpt	CC-ID=X2 (Symbolic Path Name=Class A) PPA Object (Peer IP=R1_A, Prefix=7_A)

7. New PCEP Objects

One new CCI Object and three new PCEP objects are defined in this draft. All new PCEP objects are as per [RFC5440]

7.1. CCI Object

The Central Control Instructions (CCI) Object is used by the PCE to specify the forwarding instructions is defined in [RFC9050]. This document defines another object-type for Native-IP.

CCI Object-Type is TBD13 for Native-IP as below

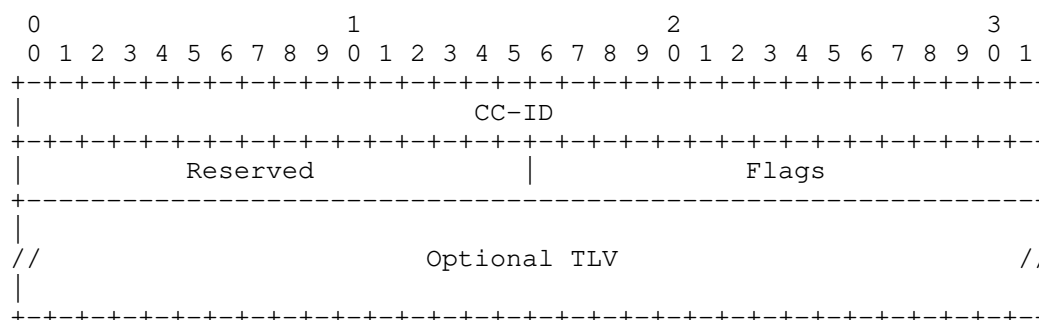


Figure 5: CCI Object for Native IP

Figure 1

The field CC-ID is as described in [RFC9050]. Following fields are defined for CCI Object-Type TBD13

Reserved: is set to zero while sending, ignored on receipt.

Flags: is used to carry any additional information pertaining to the CCI. Currently no flag bits are defined.

The Symbolic Path Name TLV [RFC8231] MUST be included in the CCI Object-Type TBD13 to identify the E2E TE path in Native IP environment and MUST be unique.

7.2. BGP Peer Info Object

The BGP Peer Info object is used to specify the information about the peer that the PCC should establish the BGP relationship with. This object should only be included and sent to the head and end router of the E2E path in case there is no Route Reflection (RR) involved. If the RR is used between the head and end routers, then such information should be sent to head router, RR and end router respectively.

By default, there MUST be no prefix be distributed via such BGP session that established by this object.

By default, the Local/Peer IP address SHOULD be dedicated to the usage of native IP TE solution, and SHOULD NOT be used by other BGP sessions that established by manual or non PCE initiated configuration.

BGP Peer Info Object-Class is TBD14

BGP Peer Info Object-Type is 1 for IPv4 and 2 for IPv6

The format of the BGP Peer Info object body for IPv4 (Object-Type=1) is as follows:

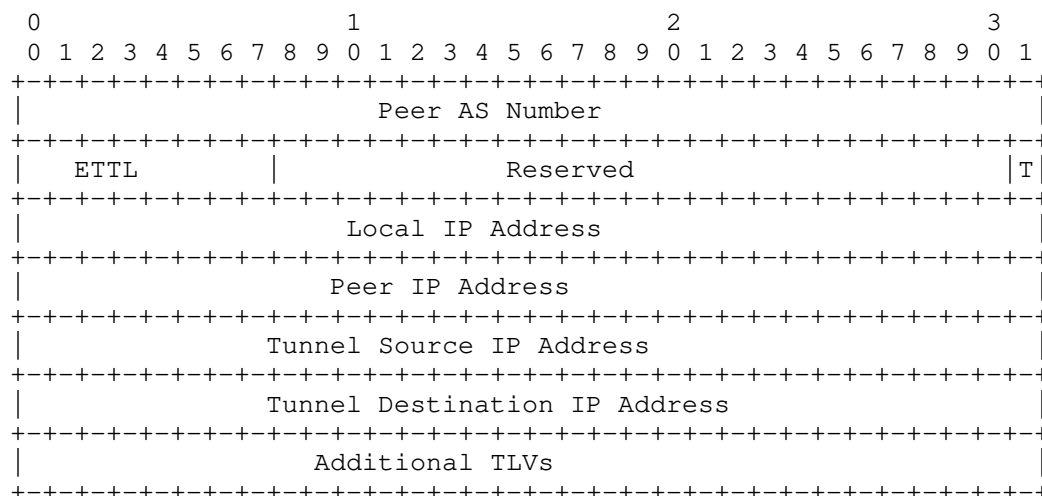


Figure 6: BGP Peer Info Object Body Format for IPv4

The format of the BGP Peer Info object body for IPv6(Object-Type=2) is as follows:

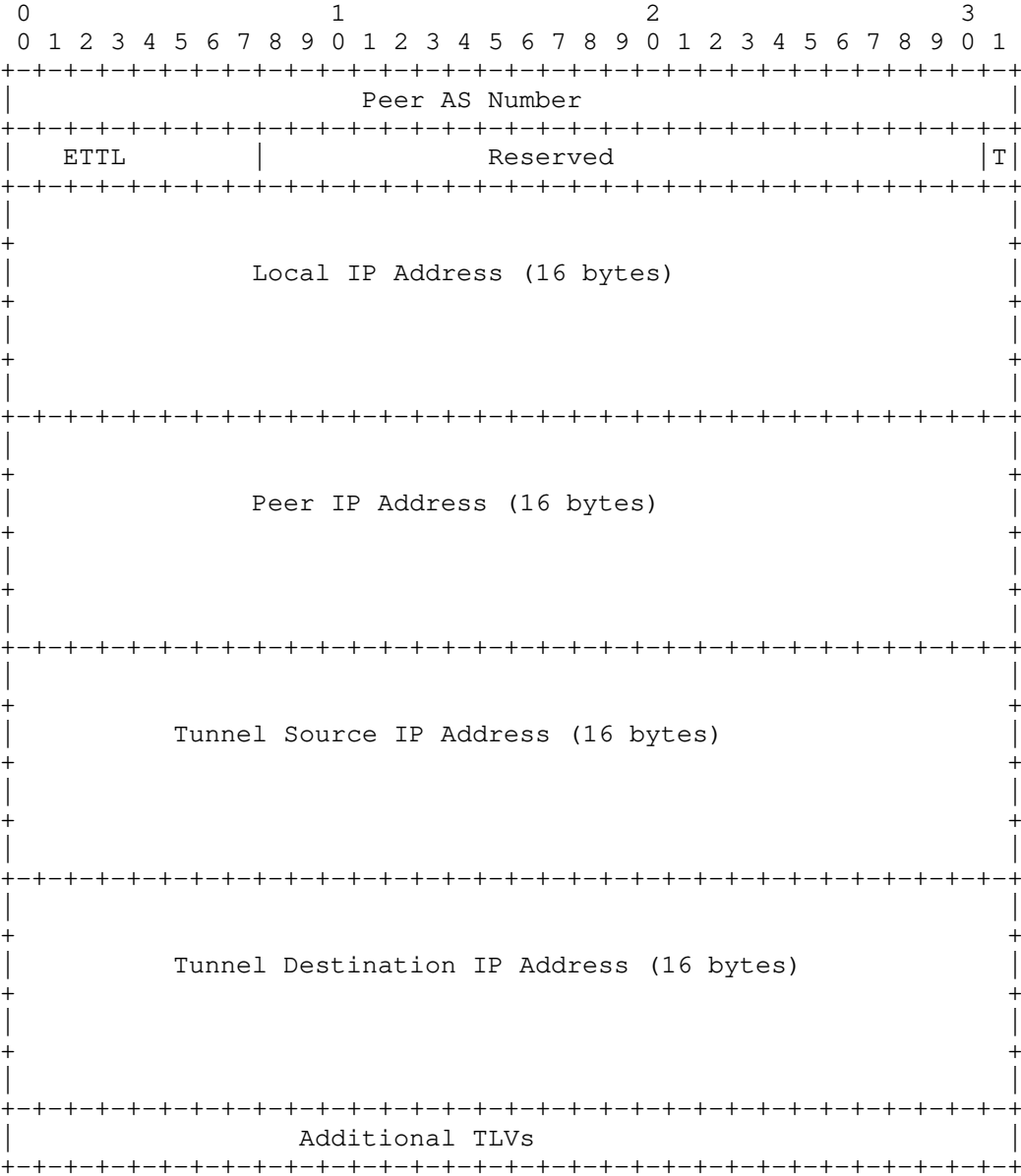


Figure 7: BGP Peer Info Object Body Format for IPv6

Peer AS Number: 4 Bytes, to indicate the AS number of Remote Peer.

ETTL: 1 Byte, to indicate the multihop count for EBGp session. It should be 0 and ignored when Local AS and Peer AS is same.

Reserved: is set to zero while sending, ignored on receipt.

T bit: Indicates whether the traffic that associated with the prefixes advertised via this BGP session is transported via IPinIP tunnel (when T bit is set) or not (when T bit is clear).

Local IP Address(4/16 Bytes): IP address of the local router, used to peer with other end router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes.

Peer IP Address(4/16 Bytes): IP address of the peer router, used to peer with the local router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes;

Tunnel Source IP Address(4/16 Bytes): IP address of the tunnel source, should be owned by the local router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes.

Tunnel Destination IP Address(4/16 Bytes): IP address of the tunnel destination, should be owned by the peer router. When Object-Type is 1, length is 4 bytes; when Object-Type is 2, length is 16 bytes. Should be different from the Peer IP Address.

Additional TLVs: TLVs that associated with this object, can be used to convey other necessary information for dynamic BGP session establishment. Their definition are out of the current document.

When PCC receives BPI object, with Object-Type=1, it should try to establish BGP session with the peer in AFI/SAFI=1/1; when PCC receives BPI object with Object-Type=2, it should try to establish the BGP session with the peer in AFI/SAFI=2/1. Other BGP capabilities, for example, Graceful Restart (GR) that enhance the BGP performance should also be negotiated and used by default.

7.3. Explicit Peer Route Object

The Explicit Peer Route object is defined to specify the explicit peer route to the corresponding peer address on each device that is on the E2E assurance path. This Object should be sent to all the devices that locates on the E2E assurance path that calculated by PCE.

The path established by this object should have higher priority than other path calculated by dynamic IGP protocol, but should be lower priority than the static route configured by manual or NETCONF or by other means.

Explicit Peer Route Object-Class is TBD15.

Explicit Peer Route Object-Type is 1 for IPv4 and 2 for IPv6

The format of Explicit Peer Route object body for IPv4 (Object-Type=1) is as follows:

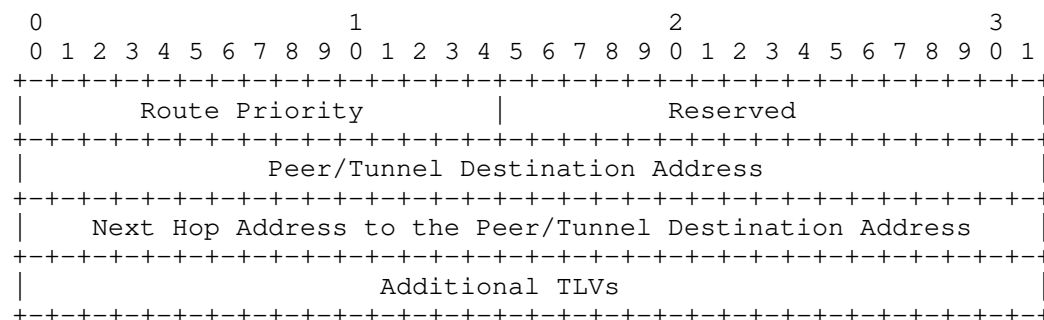


Figure 8: Explicit Peer Route Object Body Format for IPv4

The format of Explicit Peer Route object body for IPv6 (Object-Type=2) is as follows:

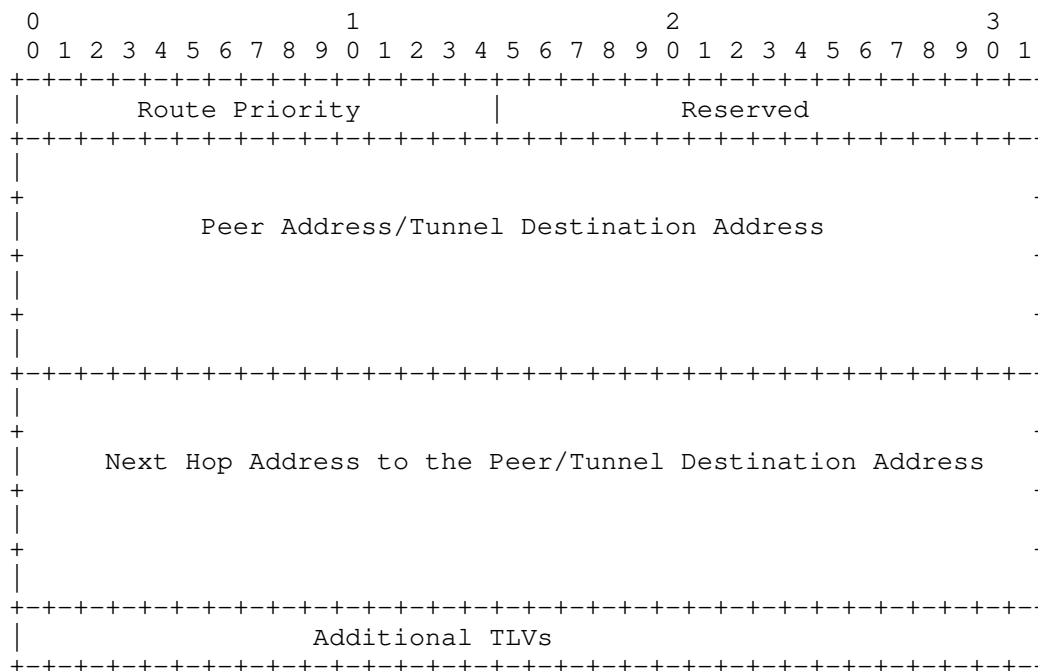


Figure 9: Explicit Peer Route Object Body Format for IPv6

Route Priority: 2 Bytes, The priority of this explicit route. The higher priority should be preferred by the device. This field is used to indicate the backup path at each hop.

Reserved: is set to zero while sending, ignored on receipt.

Peer/Tunnel Destination Address: To indicate the peer address(4/16 Bytes). When T bit is set in the associated BPI object, use the tunnel destination address in BPI object; when T bit is clear, use the peer address in BPI object.

Next Hop Address to the Peer/Tunnel Destination Address: To indicate the next hop address(4/16 Bytes) to the corresponding peer/tunnel destination address.

Additional TLVs: TLVs that associated with this object, can be used to convey other necessary information for explicit peer path establishment. Their definitions are out of the current document.

7.4. Peer Prefix Advertisement Object

The Peer Prefix Advertisement object is defined to specify the IP prefixes that should be advertised to the corresponding peer. This object should only be included and sent to the head/end router of the end2end path.

The prefixes information included in this object MUST only be advertised to the indicated peer, MUST NOT be advertised to other BGP peers.

Peer Prefix Advertisement Object-Class is TBD16

Peer Prefix Advertisement Object-Type is 1 for IPv4 and 2 for IPv6

The format of the Peer Prefix Advertisement object body is as follows:

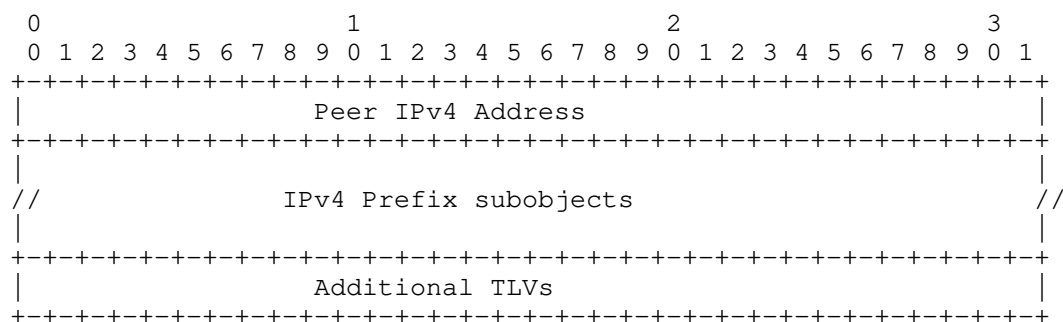


Figure 10: Peer Prefix Advertisement Object Body Format for IPv4

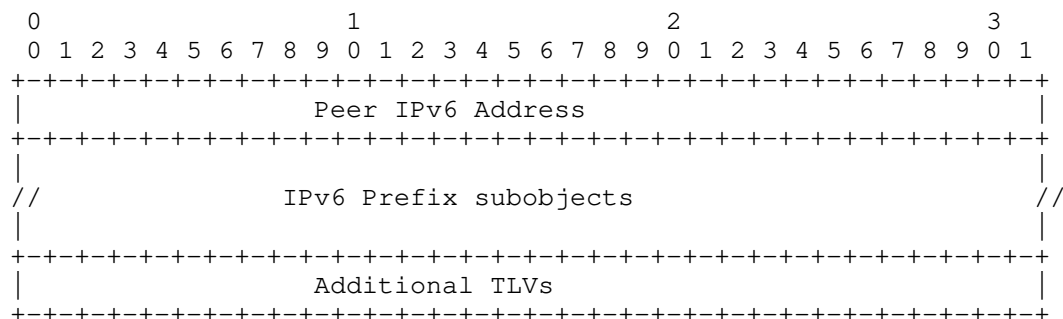


Figure 11: Peer Prefix Advertisement Object Body Format for IPv6

Peer IPv4 Address: 4 Bytes. Identifies the peer IPv4 address that the associated prefixes will be sent to.

IPv4 Prefix subobjects: List of IPv4 Prefix subobjects that defined in [RFC3209], identify the prefixes that will be sent to the peer that identified by Peer IPv4 Address List.

Peer IPv6 Address: 16 Bytes. Identifies the peer IPv6 address that the associated prefixes will be sent to.

IPv6 Prefix subobjects: List of IPv6 Prefix subobjects that defined in [RFC3209], identify the prefixes that will be sent to the peer that identified by Peer IPv6 Address List.

Additional TLVs: TLVs that associated with this object, can be used to convey other necessary information for prefixes advertisement. Their definitions are out of the current document.

8. End to End Path Protection

[RFC8697] defines the path associations procedures between sets of Label Switched Path (LSP). Such procedures can also be used for the E2E path protection. To accomplish this, the PCE should attach the ASSOCIATION object with the EPR object in the PCInitiate message, with the association type set to 1 (Path Protection Association). The Extended Association ID that included within the Extended Association ID TLV, which is included in the ASSOCIATION object, should be set to the Symbolic Path Name of different E2E path. This PCInitiate should be sent to the head-end of the E2E path.

The head-end of the path can use the existing path detection mechanism(for example, Bidirectional Forwarding Detection [RFC5880]), to monitor the status of the active path. Once it detects the failure, it can switch the backup protection path immediately.

9. Re-Delegation and Clean up

In case of a PCE failure, a new PCE can gain control over the central controller instructions. As per the PCEP procedures in [RFC8281], the State Timeout Interval timer is used to ensure that a PCE failure does not result in automatic and immediate disruption for the services. Similarly, as per [RFC9050], the central controller instructions are not removed immediately upon PCE failure. Instead, they could be re-delegated to the new PCE before the expiration of this timer, or be cleaned up on the expiration of this timer. The allows for network clean up without manual intervention. The PCC MUST support the removal of CCI as one of the behaviors applied on expiration of the State Timeout Interval timer.

10. BGP Considerations

This draft defines the procedures and objects to create the BGP sessions and advertises the associated prefixes dynamically. Only the key information, for example peer IP addresses, peer AS number are exchanged via the PCEP protocol. Other parameters that are needed for the BGP session setup should be derived from their default values, as described in Section 7.2. Upon receives such key information, the BGP module on the PCC should try to accomplish the task that appointed by the PCEP protocol and report the status to the PCEP modules.

There is no influence to current implementation of BGP Finite State Machine(FSM). The PCEP cares only the success and failure status of BGP session, and act upon such information accordingly.

The error handling procedures related to incorrect BGP parameters are specified in Section 6.1, Section 6.2, and Section 6.3. The handling of the dynamic BGP sessions and associated prefixes on PCE failure is described in Section 9.

11. New Error-Types and Error-Values Defined

A PCEP-ERROR object is used to report a PCEP error and is characterized by an Error-Type that specifies that type of error and an Error-value that provides additional information about the error. An additional Error-Type and several Error-values are defined to represent some the errors related to the newly defined objects, which are related to Native IP TE procedures.

Error-Type	Meaning	Error-value
TBD6	Native IP TE failure	
		0: Unassigned
		TBD7: Peer AS not match
		TBD8:Peer IP can't be reached
		TBD9:Local IP is in use
		TBD10:Remote IP is in use
		TBD11:Exist BGP session broken
		TBD12:Explicit Peer Route Error
		TBD17:EPR/BPI Peer Info mismatch
		TBD18:BPI/PPA Address Family mismatch
		TBD19:PPA/BPI Peer Info mismatch

Figure 12: Newly defined Error-Type and Error-Value

12. Deployment Considerations

The information transferred in this draft is mainly used for the light weight BGP session setup, explicit route deployment and the prefix distribution. The planning, allocation and distribution of the peer addresses within IGP should be accomplished in advanced and they are out of the scope of this draft.

[RFC8232] describes the state synchronization procedure between stateful PCE and PCC. The communication of PCE and PCC described in this draft should also follow this procedures, treat the three newly defined objects that associated with the same symbolic path name as the attribute of the same path in the LSP-DB.

When PCE detects one or some of the PCCs are out of control, it should recompute and redeploy the traffic engineering path for native IP on the active PCCs. When PCC detects that it is out of control of the PCE, it should clear the information that initiated by the PCE. The PCE should assure the avoidance of possible transient loop in such node failure when it deploy the explicit peer route on the PCCs.

If the established BGP session is broken after some time, the PCC should also report such error via PCErr message with Err-type=TBD6 and error value(Error-value=TBD11, Existing BGP session is broken). Upon receiving such PCErr message, the PCE should clear the prefixes advertisement on the previous BGP session, clear the explicit peer route to the previous peer address; select other Local_IP/Peer_IP pair to establish the new BGP session, deploy the explicit peer route to the new peer address, and advertises the prefixes on the new BGP session.

13. Implementation Status

[Note to the RFC Editor - remove this section before publication, as well as remove 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".

13.1. Proof of Concept based on ODL

.At the time of posting the -18 version of this document, there are no known implementations of this mechanism. A proof of concept for the overall design has been verified using another SBI protocol on the Open DayLight (ODL) controller.

14. Security Considerations

The setup of BGP sessions, prefix advertisement, and explicit peer route establishment are all controlled by the PCE. See [RFC4271] and [RFC4272] for BGP security considerations. Security consideration part in [RFC5440] and [RFC8231] should be considered. To prevent a bogus PCE sending harmful messages to the network nodes, the network devices should authenticate the validity of the PCE and ensure a secure communication channel between them. Mechanisms described in [RFC8253] should be used.

15. IANA Considerations

15.1. Path Setup Type Registry

[RFC8408] created a sub-registry within the "Path Computation Element Protocol (PCEP) Numbers" registry called "PCEP Path Setup Types". IANA is requested to allocate a new code point within this registry, as follows:

Value	Description	Reference
TBD1	Native IP TE Path	This document

15.2. PCECC-CAPABILITY sub-TLV's Flag field

[RFC9050] created a sub-registry within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the value of the PCECC-CAPABILITY sub-TLV's 32-bits Flag field. IANA is requested to allocate a new bit position within this registry, as follows:

Value	Description	Reference
TBD2(N)	NATIVE-IP-TE-CAPABILITY	This document

15.3. PCEP Object Types

IANA is requested to allocate new registry for the PCEP Object Type:

Object-Class Value	Name	Reference
44	CCI Object Object-Type TBD13: Native IP	This document
TBD14	BGP Peer Info Object-Type 1: IPv4 address 2: IPv6 address	This document
TBD15	Explicit Peer Route Object-Type 1: IPv4 address 2: IPv6 address	This document
TBD16	Peer Prefix Advertisement Object-Type 1: IPv4 address 2: IPv6 address	This document

15.4. PCEP-Error Object

IANA is requested to allocate new error types and error values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry for the following errors::

Error-Type	Meaning	Error-value
		Reference
6	Mandatory Object missing	TBD4:Native IP object missing This document
10	Reception of an invalid object	TBD3:PCECC NATIVE-IP-TE-CAPABILITY bit is not set This document
19	Invalid Operation	TBD5:Only one of the BPI,EPR or PPA object can be included in this message This document
TBD6	Native IP TE failure	This document TBD7:Peer AS not match TBD8:Peer IP can't be reached TBD9:Local IP is in use TBD10:Remote IP is in use TBD11:Exist BGP session broken TBD12:Explicit Peer Route Error TBD17:EPR/BPI Peer Info mismatch TBD18:BPI/PPA Address Family mismatch TBD19:PPA/BPI Peer Info mismatch

16. Contributor

Dhruv Dhody has contributed the contents of this draft.

17. Acknowledgement

Thanks Mike Koldychev, Susan Hares, Siva Sivabalan, Adam Simpson for his valuable suggestions and comments.

18. 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>>.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<https://www.rfc-editor.org/info/rfc4272>>.
- [RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <<https://www.rfc-editor.org/info/rfc5440>>.
- [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>>.
- [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>>.
- [RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <<https://www.rfc-editor.org/info/rfc8231>>.

- [RFC8232] Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", RFC 8232, DOI 10.17487/RFC8232, September 2017, <<https://www.rfc-editor.org/info/rfc8232>>.
- [RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <<https://www.rfc-editor.org/info/rfc8253>>.
- [RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <<https://www.rfc-editor.org/info/rfc8281>>.
- [RFC8283] Farrel, A., Ed., Zhao, Q., Ed., Li, Z., and C. Zhou, "An Architecture for Use of PCE and the PCE Communication Protocol (PCEP) in a Network with Central Control", RFC 8283, DOI 10.17487/RFC8283, December 2017, <<https://www.rfc-editor.org/info/rfc8283>>.
- [RFC8408] Sivabalan, S., Tantsura, J., Minei, I., Varga, R., and J. Hardwick, "Conveying Path Setup Type in PCE Communication Protocol (PCEP) Messages", RFC 8408, DOI 10.17487/RFC8408, July 2018, <<https://www.rfc-editor.org/info/rfc8408>>.
- [RFC8697] Minei, I., Crabbe, E., Sivabalan, S., Ananthakrishnan, H., Dhody, D., and Y. Tanaka, "Path Computation Element Communication Protocol (PCEP) Extensions for Establishing Relationships between Sets of Label Switched Paths (LSPs)", RFC 8697, DOI 10.17487/RFC8697, January 2020, <<https://www.rfc-editor.org/info/rfc8697>>.
- [RFC8735] Wang, A., Huang, X., Kou, C., Li, Z., and P. Mi, "Scenarios and Simulation Results of PCE in a Native IP Network", RFC 8735, DOI 10.17487/RFC8735, February 2020, <<https://www.rfc-editor.org/info/rfc8735>>.
- [RFC8821] Wang, A., Khasanov, B., Zhao, Q., and H. Chen, "PCE-Based Traffic Engineering (TE) in Native IP Networks", RFC 8821, DOI 10.17487/RFC8821, April 2021, <<https://www.rfc-editor.org/info/rfc8821>>.

[RFC9050] Li, Z., Peng, S., Negi, M., Zhao, Q., and C. Zhou, "Path Computation Element Communication Protocol (PCEP) Procedures and Extensions for Using the PCE as a Central Controller (PCECC) of LSPs", RFC 9050, DOI 10.17487/RFC9050, July 2021, <<https://www.rfc-editor.org/info/rfc9050>>.

Authors' Addresses

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing
Beijing, 102209
China
Email: wangaj3@chinatelecom.cn

Boris Khasanov
Yandex LLC
Ulitsa Lva Tolstogo 16
Moscow
Email: bhassanov@yahoo.com

Sheng Fang
Huawei Technologies, Co., Ltd
Huawei Bld., No.156 Beiqing Rd.
Beijing
China
Email: fsheng@huawei.com

Ren Tan
Huawei Technologies, Co., Ltd
Huawei Bld., No.156 Beiqing Rd.
Beijing
China
Email: tanren@huawei.com

Chun Zhu
ZTE Corporation
50 Software Avenue, Yuhua District
Nanjing
Jiangsu, 210012
China
Email: zhu.chun1@zte.com.cn

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 6 September 2022

W. Li
H. Wang
J. Dong
Huawei Technologies
5 March 2022

Extension of Link Bandwidth Extended Community
draft-li-idr-link-bandwidth-ext-01

Abstract

[I-D.ietf-idr-link-bandwidth] defines a BGP link bandwidth extended community attribute, which can enable devices to implement unequal-cost load-balancing. However, the bandwidth value encapsulated by the extended community attribute is of the floating-point type, which is inconvenient to use. In this document, a set of new types of link bandwidth extended community are introduced to facilitate the configuration and calculation of link bandwidth.

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 6 September 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	2
2. Link Bandwidth Extended Community	3
3. Deployment Considerations	3
4. IANA Considerations	4
5. Security Considerations	4
6. Acknowledgements	4
7. References	4
7.1. Normative References	4
7.2. References	4
Authors' Addresses	4

1. Introduction

In [I-D.ietf-idr-link-bandwidth], the link bandwidth extended community attribute is added to implement unequal-cost load balancing based on the bandwidth on a path. As defined in the draft, the bandwidth of a link is expressed in 4-octets in IEEE floating-point format.

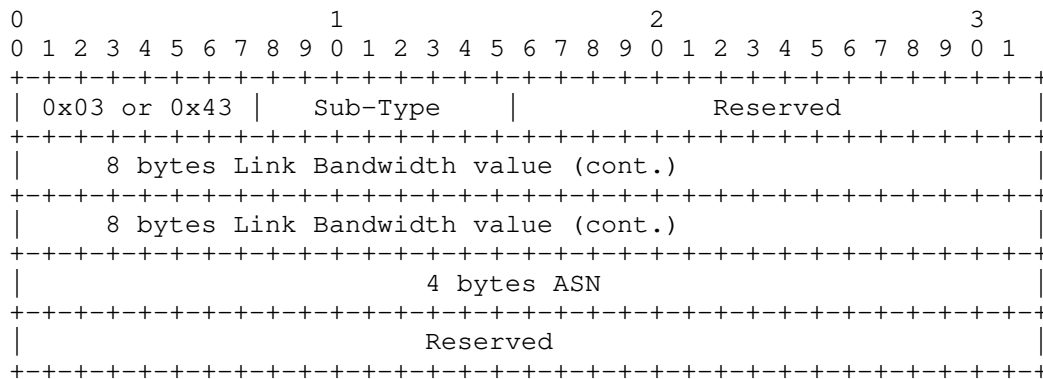
In practice, the use of this floating-point format may result errors in configuration and computation. When an operator needs to manually specify the bandwidth, you also need to consider the conversion from the bandwidth value to the floating-point number. This mode is not user-friendly, especially when the routing policy is used for bandwidth matching.

This document introduce a more intuitive expression of link bandwidth in BGP. It uses an unsigned long integer value to describe the link bandwidth value. This is easier for operators to use and understand, and can avoid configuration and computation errors.

2. Link Bandwidth Extended Community

The type of Link Bandwidth Extended Community is 0x40, and the subtype is 0x04. In the attribute value, the global administrator subfield is set to the AS number of the route to which the Link Bandwidth attribute is added. In the local administrator subfield, the link bandwidth value [I-D.ietf-idr-link-bandwidth] is set to the IEEE floating-point type.

A new type of IPv6 Address Specific Extended Community[RFC5701] is added in this document. The ASN field of this attribute is set to the AS number of the route to which the link bandwidth attribute is added. The Link Bandwidth value field (8 bytes) is set to the link bandwidth. The following extended contents are added:



- * The value of the high-order octet of the extended Type, refer to [RFC4360], It is recommended that 0x03 and 0x43 be used.
- * New Link Bandwidth, subtype is TBD. The value of the Link Bandwidth subfield is an unsigned long integer, in bytes per second.

The subtypes defined here can be used for both optional transitive and non-transitive extended community attributes.

3. Deployment Considerations

The extended link bandwidth extended community attribute in this document should not be used together with the standard link bandwidth extended community attribute. If a route carries both the standard link bandwidth extended community attribute and the unit link bandwidth extended community attribute, the standard link bandwidth extended community attribute is ignored.

In actual deployment, if a routing policy is used to match link bandwidth attributes, you can directly perform exact value matching.

4. IANA Considerations

This document defines a specific application of the two-octet AS specific extended community. IANA is requested to assign new sub-types for both non-transitive and transitive extended communities.

SubType	Description
TBD	Link Bandwidth EC in bytes per second

5. Security Considerations

There are no additional security risks introduced by this design.

6. Acknowledgements

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, <<https://www.rfc-editor.org/info/rfc2119>>.

7.2. References

[I-D.ietf-idr-link-bandwidth] Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", Work in Progress, Internet-Draft, draft-ietf-idr-link-bandwidth-07, 5 March 2018, <<https://www.ietf.org/archive/id/draft-ietf-idr-link-bandwidth-07.txt>>.

[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>>.

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, DOI 10.17487/RFC5701, November 2009, <<https://www.rfc-editor.org/info/rfc5701>>.

Authors' Addresses

Wenyan Li
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing
100095
China
Email: liwenyan@huawei.com

Haibo Wang
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing
100095
China
Email: rainsword.wang@huawei.com

Jie Dong
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing
100095
China
Email: jie.dong@huawei.com

IDR WG
Internet-Draft
Intended status: Standards Track
Expires: 22 June 2022

Y. Liu
S. Peng
ZTE
19 December 2021

BGP Extensions of SR Policy for Path Protection
draft-lp-idr-sr-path-protection-02

Abstract

This document proposes extensions of BGP to provide protection information of segment lists within a candidate path when delivering SR policy. And it also extends BGP-LS to provide some extra information of the segment list in the advertisement.

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 22 June 2022.

Copyright Notice

Copyright (c) 2021 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	2
1.1. Requirements Language	3
2. BGP Extensions for Advertising Segment List	3
2.1. Extensions of Segment List sub-TLV	3
2.2. List Identifier Sub-TLV	4
2.2.1. List Protection Sub-TLV	4
3. BGP-LS Extensions for Distributing Segment List States	7
4. IANA Considerations	7
4.1. New Registry: Flag Field of Segment List sub-TLV	7
4.2. Existing Registry: BGP Tunnel Encapsulation Attribute sub-TLVs	7
4.3. New Registry: List Identifier Sub-TLVs	8
4.4. Existing Registry: Flag Field of SR Segment List TLV . .	8
5. Security Considerations	8
6. References	8
6.1. Normative References	8
6.2. Informative References	9
Authors' Addresses	9

1. Introduction

Segment Routing [RFC8402] allows a headend node to steer a packet flow along any path. [I-D.ietf-spring-segment-routing-policy] details the concept of SR Policy and steering into an SR Policy. An SR Policy is a set of candidate paths, each consisting of one or more segment lists. The headend of an SR Policy may learn multiple candidate paths for an SR Policy.

Candidate path can be used for path protection, that is, the lower preference candidate path may be designated as the backup for a specific or all (active) candidate path(s). Backup candidate path provide protection only when all the segment lists in the active CP are invalid.

If a candidate path is associated with a set of Segment-Lists, each Segment-List is associated with weight for weighted load balancing.

The protection mechanism for SR Policy is not flexible enough. For example, there're three segment lists(SL1, SL2, SL3) in candidate path 1, it may be desired that SL1 and SL2 are the primary path, SL3 are the backup path for SL1 and will be active only when SL1 fails.

[I-D.ietf-pce-multipath] proposes extensions to PCEP to specify the protection relationship between segment lists in the candidate path.

[I-D.ietf-idr-segment-routing-te-policy] specifies BGP extensions for the advertisement of SR Policies and each candidate path is carried in an NLRI. This document proposes extensions of BGP in order to provide protection information of segment lists when delivering SR policy.

[I-D.ietf-idr-te-lsp-distribution] describes a mechanism to collect the SR policy information that is locally available in a node and advertise it into BGP Link State (BGP-LS) updates. This document also extends it to provide some extra information of the segment list in a candidate path in the BGP-LS advertisement.

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 [RFC2119].

2. BGP Extensions for Advertising Segment List

2.1. Extensions of Segment List sub-TLV

Segment List sub-TLV is introduced in [I-D.ietf-idr-segment-routing-te-policy] and it includes the elements of the paths (i.e., segments).

This document introduces a one-bit flag in the RESERVED field.

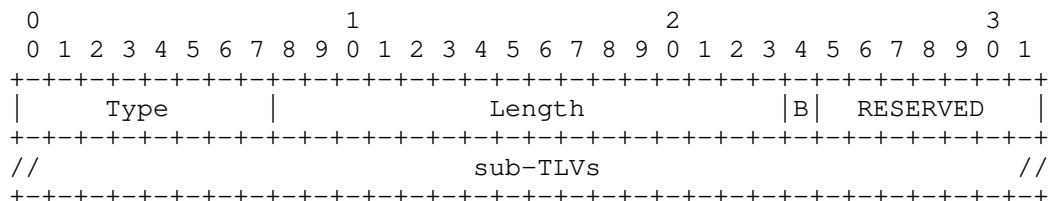


Figure 1: Segment List sub-TLV

B-Flag(Backup Flag): one bit. When set to 0, it indicates that the segment list acts as the active member in the candidate path. When set to 1, it indicates that the segment list acts as the backup path in the candidate path.

Using segment lists for path protection can be compatible with using candidate paths. When a path fails, the backup segment list within the same candidate path is used preferentially for path protection. If the backup list is also invalid, then other candidate path can be enabled for protection.

2.2. List Identifier Sub-TLV

This document introduces a new sub-sub-tlv of Segment List sub-TLV, where,

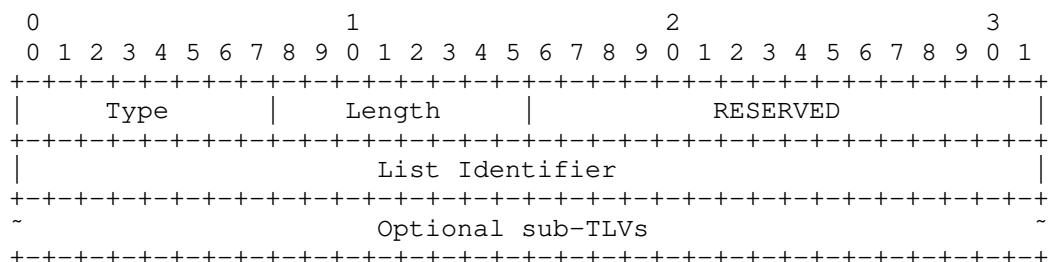


Figure 2: List Identifier Sub-TLV

- * Type: 1 octet. TBD.
- * Length: 1 octet, specifies the length of the value field not including Type and Length fields.
- * RESERVED: 2 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- * List Identifier: 4 octets. It is the identifier of the corresponding segment list, so that the segment list can be operated according to the specified Segment List identifier.
- * This sub-TLV is optional and it MUST NOT appear more than once inside the Segment List sub-TLV.

2.2.1. List Protection Sub-TLV

The List Protection Info sub-TLV is an optional sub-TLV of List Identifier sub-TLV, where:

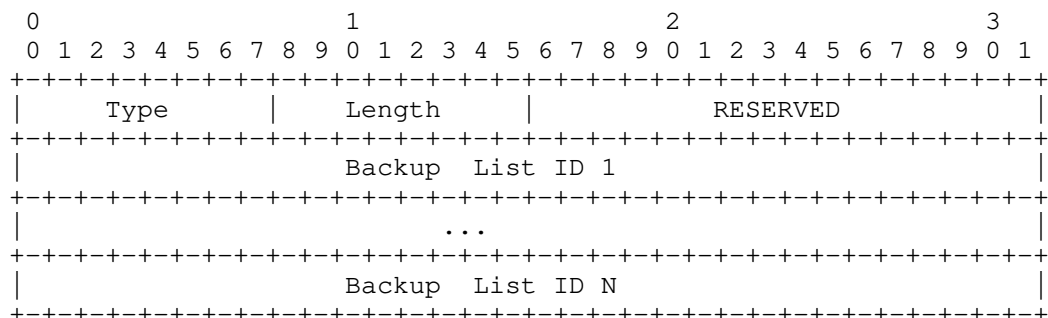


Figure 3: List Protection Info Sub-TLV

- * Type: 1 octet. TBD.
- * Length: 1 octet, specifies the length of the value field not including Type and Length fields.
- * RESERVED: 2 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- * Backup List ID: 4 octets. It is the List Identifier of the backup segment list that protects this segment list. If there're multiple backup paths, the list ID of each path should be included in the TLV.

As defined in [I-D.ietf-idr-segment-routing-te-policy], 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
 - Priority
 - Policy Name
 - Explicit NULL Label Policy (ENLP)
 - Segment List
 - Weight
 - Segment
 - Segment
 - ...
 - Segment List
 - ...
 - ...

The new SR Policy encoding structure with List Identifier sub-TLV is shown as below:

SR Policy SAFI NLRI: <Distinguisher, Policy-Color, Endpoint>
Attributes:
Tunnel Encaps Attribute (23)

- Tunnel Type: SR Policy
- Binding SID
- SRv6 Binding SID
- Preference
- Priority
- Policy Name
- Policy Candidate Path Name
- Explicit NULL Label Policy (ENLP)
- Segment List
 - List Identifier
 - List Protection Info
 - Weight
 - Segment
 - Segment
 - ...
- Segment List
- ...
- ...

3. BGP-LS Extensions for Distributing Segment List States

[I-D.ietf-idr-te-lsp-distribution] describes a mechanism to collect the SR Policy information that is locally available in a node and advertise it into BGP Link State (BGP-LS) updates. The SR Policy information includes status of the candidate path, e.g, whether the candidate path is administrative shut or not.

SR Segment List TLV is defined in [I-D.ietf-idr-te-lsp-distribution] to report the SID-List(s) of a candidate path. Figure 4 shows the flags in SR Segment List TLV.

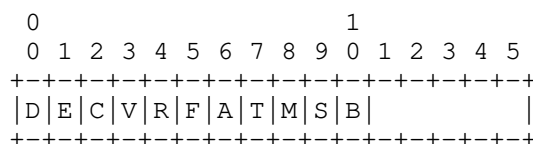


Figure 4: Flag Field of SR Segment List TLV

The D,E,C,V,R,F,A,M flags are defined in [I-D.ietf-idr-te-lsp-distribution].

This document introduces two new flags, where,

- * S-Flag : Indicates the segment list is in administrative shut state when set.
- * B-Flag : Indicates the segment list is the backup path within the candidate path when set, otherwise it is the active path.

4. IANA Considerations

4.1. New Registry: Flag Field of Segment List sub-TLV

This document introduces a one-bit flag field in the Segment List sub-TLV [I-D.ietf-idr-segment-routing-te-policy] for the Backup Flag (B-Flag).

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

This document defines a new sub-TLV in the registry "SR Policy List Sub-TLVs" [I-D.ietf-idr-segment-routing-te-policy] to be assigned by IANA:

Codepoint	Description	Reference
TBD	List Identifier Sub-TLV	This document

4.3. New Registry: List Identifier Sub-TLVs

This document requests the creation of a new registry called "List Identifier Sub-TLVs" under the "BGP Tunnel Encapsulation" registry. Following initial Sub-TLV codepoint are assigned by this document.

Codepoint	Description	Reference
TBD	List Protection Sub-TLV	This document

4.4. Existing Registry: Flag Field of SR Segment List TLV

This document requests bit 9 and bit 10 in the flag field of "SR Segment List TLV" [I-D.ietf-idr-te-lsp-distribution] under the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry.

Bit	Description	Reference
9	Administrative Shut State Flag(S-Flag)	This document
10	Backup Path State Flag(B-Flag)	This document

5. Security Considerations

Procedures and protocol extensions defined in this document do not affect the security considerations discussed in [I-D.ietf-idr-segment-routing-te-policy] and [I-D.ietf-idr-te-lsp-distribution].

6. References

6.1. Normative References

[I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filisfilis, C., Talaulikar, K., Mattes, P., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-segment-routing-te-policy-14, 10 November 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-segment-routing-te-policy-14>>.

[I-D.ietf-idr-te-lsp-distribution]
Previdi, S., Talaulikar, K., Dong, J., Chen, M., Gredler, H., and J. Tantsura, "Distribution of Traffic Engineering (TE) Policies and State using BGP-LS", Work in Progress, Internet-Draft, draft-ietf-idr-te-lsp-distribution-16, 22 October 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-te-lsp-distribution-16>>.

[I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Talaulikar, K., Voyer, D., Bogdanov, A., and
P. Mattes, "Segment Routing Policy Architecture", Work in
Progress, Internet-Draft, draft-ietf-spring-segment-
routing-policy-14, 25 October 2021,
<[https://datatracker.ietf.org/doc/html/draft-ietf-spring-
segment-routing-policy-14](https://datatracker.ietf.org/doc/html/draft-ietf-spring-segment-routing-policy-14)>.

[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>>.

6.2. Informative References

[I-D.ietf-pce-multipath]
Koldychev, M., Sivabalan, S., Saad, T., Beeram, V. P.,
Bidgoli, H., Yadav, B., and S. Peng, "PCEP Extensions for
Signaling Multipath Information", Work in Progress,
Internet-Draft, draft-ietf-pce-multipath-03, 25 October
2021, <[https://datatracker.ietf.org/doc/html/draft-ietf-
pce-multipath-03](https://datatracker.ietf.org/doc/html/draft-ietf-pce-multipath-03)>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L.,
Decraene, B., Litkowski, S., and R. Shakir, "Segment
Routing Architecture", RFC 8402, DOI 10.17487/RFC8402,
July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.

Authors' Addresses

Yao Liu
ZTE
Nanjing
China

Email: liu.yao71@zte.com.cn

Shaofu Peng
ZTE
Nanjing
China

Email: peng.shaofu@zte.com.cn

IDR Workgroup
Internet-Draft
Intended status: Standards Track
Expires: 12 November 2022

A. Retana
Y. Qu
Futurewei Technologies, Inc.
J. Tantsura
Microsoft
11 May 2022

Use of Streams in BGP over QUIC
draft-retana-idr-bgp-quic-stream-02

Abstract

This document specifies the use of QUIC Streams to support multiple BGP sessions over one connection in order to achieve high resiliency.

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 12 November 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	2
1.1. Requirements Language	3
2. Multiple BGP Sessions	3
2.1. Multiple QUIC Streams	3
2.2. Multiple BGP Sessions Using QUIC Streams	4
3. MultiStream Capability	4
4. Error Handling	5
5. BGP Session Establishment and Collision Avoidance	6
6. Modifications to FSM	7
7. Operational Considerations	7
7.1. Backward Compatibility	7
7.2. Session Prioritization	7
7.3. Other Considerations	8
8. Security Considerations	8
9. IANA Considerations	8
10. Acknowledgement	9
11. References	9
11.1. Normative References	9
11.2. Informative References	10
Authors' Addresses	11

1. Introduction

The Border Gateway Protocol (BGP) [RFC4271] uses TCP as its transport protocol. BGP establishes peer relationships between routers using a TCP session on port 179. TCP also provides reliable packet communication.

Multiprotocol Extensions for BGP-4 (MP-BGP) [RFC4760] allow BGP to carry information for multiple Network Layer protocols. However, only a single TCP connection can reach the Established state between a pair of peers [RFC4271].

As pointed out by [I-D.ietf-idr-bgp-multisession], there are some disadvantages of using a single BGP session:

A common criticism of BGP is the fact that most malformed messages cause the session to be terminated. While this behavior is necessary for protocol correctness, one may observe that the protocol machinery of a given implementation may only be defective with respect to a given AFI/SAFI. Thus, it would be desirable to allow the session related to that family to be terminated while leaving other AFI/SAFI unaffected. As BGP is commonly deployed, this is not possible.

A second criticism of BGP is that it is difficult or in some cases impossible to manage control plane resource contention when BGP is used to support diverse services over a single session. In contrast, if a single BGP session carries only information for a single service (or related set of services) it may be easier to manage such contention.

QUIC [RFC9000] is a UDP-based multiplexed and secure transport protocol. QUIC can provide low latency and encrypted transport with resilient connections. [I-D.chen-idr-bgp-over-quic] specifies the procedure to use BGP over QUIC. Complementary to it, this document specifies a mechanism to support multiple BGP sessions using QUIC streams.

Each BGP session operates independently. Thus, an error on one session has no impact on any other session. The Network Layer protocol(s) negotiated in the BGP OPEN message distinguish the sessions.

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. Multiple BGP Sessions

2.1. Multiple QUIC Streams

QUIC [RFC9000] is a UDP-based secure transport protocol that provides connection-oriented and stateful interaction between a client and server. It integrates TLS and allows the exchange of application data as soon as possible.

In QUIC, application protocols exchange information via streams, and multiple streams can be multiplexed onto an underlying connection. Each stream is a separate unidirectional or bidirectional channel of "order stream of bytes." Moreover, each stream has flow control which limits bytes sent on a stream, together with flow control of the connection.

2.2. Multiple BGP Sessions Using QUIC Streams

BGP over QUIC [I-D.chen-idr-bgp-over-quic] proposes different options to map streams. This document specifies a complementary and backward compatible mechanism to establish multiple BGP sessions using QUIC streams. An implementation can assign one or more Network Layer protocols to a BGP session.

A QUIC stream is created by sending a BGP OPEN message, and each stream MUST be bidirectional as described in Section 2.1 of [RFC9000]. In addition, the corresponding stream MUST end (clean termination) as described in Section 2.4 of [RFC9000] when a BGP session is terminated.

Section 5 describes the Connection Collision Detection procedure to be used with streams. Each BGP session operates independently, which means critical conditions (such as a malformed message) in one session won't affect others.

3. MultiStream Capability

The MultiStream Capability (MSC) is defined to indicate that a BGP speaker supports multiple sessions as specified in this document. The capability [RFC5492] is defined as follows:

Capability code (1 octet): TBD1

Capability length (1 octet): 1

Capability value (1 octet): flag field reserved.

```

 0 1 2 3 4 5 6 7
+---+---+---+---+
|   Reserved   |
+---+---+---+---+
```

Flags: bitfield - MUST be set to zero and ignored by the receiver.

The MSC only applies when using BGP over QUIC [I-D.chen-idr-bgp-over-quic]. It MUST be included in all OPEN messages. It MUST be ignored otherwise.

This specification applies only if both peers advertise the MSC during the establishment of the "initial session." Otherwise, the processes specified in [I-D.chen-idr-bgp-over-quic] MUST be followed. In particular, if a peer that advertises the MSC doesn't receive an OPEN message with the MSC from its peer, it SHOULD NOT terminate the session.

Using the MSC allows peers to establish multiple BGP sessions, one per QUIC stream. Each new BGP session is established using a separate OPEN message [RFC4271] and MUST include the MSC. If both peers exchange the MSC in the "initial session," they MUST include it when establishing other sessions. Otherwise, the new session MUST be terminated, and the Error Subcode MUST be set to MultiStream Conflict (TBD2), defined in Section 4.

Once a BGP session is established, it follows the procedures specified in [RFC4271].

4. Error Handling

OPEN message error handling is defined in section 6.2 of [RFC4271]. This document introduces the following OPEN Message Error subcodes:

TBD2 - MultiSession Conflict - Used if the MSC is exchanged by both peers in the "initial session" but is not present when establishing a new session.

TBD3 - Session Capability Mismatch - Used if a BGP speaker terminates a session in the case where it sends an OPEN message with the MSC but receives an OPEN message without it.

TBD4 - Network Layer Protocol Mismatch - Used if a BGP session has already been established for a signaled Network Layer Protocol, either individually or as part of a set.

Section 3 recommends not terminating a session when only one peer supports the MSC. If such a BGP speaker does terminate the session, the Error Subcode MUST be set to Session Capability Mismatch (TBD3).

Any individual BGP session can be terminated as specified in [RFC4486]. If multiple sessions are to be terminated, then the procedure MUST be followed for each one.

5. BGP Session Establishment and Collision Avoidance

Before creating a new session, a BGP speaker should check that no session exists for the same Network Layer protocol(s). If a session already exists, the BGP speaker SHOULD NOT attempt to create a new one.

If a pair of BGP speakers try to establish a BGP session with each other simultaneously, then two parallel sessions will be formed. In the case of BGP over QUIC, the IP addresses of the connection cannot be used to resolve collisions when using multiple streams.

To avoid connection collisions, a session is identified by the My Autonomous System and BGP Identifier fields pair in the OPEN message. In this context, a connection collision is the attempt to open a BGP session for which the set of Network Layer protocols is the same. One of the connections MUST be closed.

The connection collision is resolved using the extension specified in [RFC6286]. In other words, the session with the higher-valued BGP Identifier is preserved [RFC4271]. If the BGP Identifiers are identical, then the session with the larger ASN is preserved [RFC6286].

Upon receiving an OPEN message, the local system MUST examine all of its sessions in the OpenConfirm state. A BGP speaker MAY also examine sessions in an OpenSent state if it knows the BGP Identifier of the peer by means outside of the protocol. If among these sessions, there is one to a remote BGP speaker whose BGP Identifier and ASN pair equals the one in the OPEN message, and this session collides with the connection over which the OPEN message is received, then the local system performs the following collision resolution procedure:

- 1) The BGP Identifier of the local system is compared to the BGP Identifier of the remote system (as specified in the OPEN message). Comparing BGP Identifiers is done by converting them to host byte order and treating them as 4-octet unsigned integers.
- 2) If the value of the local BGP Identifier is less than the remote one, the local system closes the BGP connection that already exists (the one that is already in the OpenConfirm state) and accepts the BGP connection initiated by the remote system.
- 2a) Otherwise, the local system closes the newly created BGP connection (the one associated with the recently received OPEN message) and continues to use the existing one (the one that is already in the OpenConfirm state).

3) If the BGP Identifiers of the peers involved in the connection collision are identical, then the session initiated by the BGP speaker with the larger AS number is preserved.

Unless allowed via configuration, a connection collision with an existing BGP session in the Established state causes the closing of the newly created session.

Closing the BGP session (that results from the collision resolution procedure) is accomplished by sending the NOTIFICATION message with the Error Code Cease, Subcode Connection Collision Resolution (7) [RFC4486].

The remainder of the process is as specified in [RFC4271].

6. Modifications to FSM

The modifications to BGP FSM is described in section 4.4 of [I-D.chen-idr-bgp-over-quic]. For simplicity and security reason, it is suggested that 1-RTT is used.

This specification does not modify BGP FSM, but the collision handling procedure should be replaced with the procedure described in this document.

7. Operational Considerations

7.1. Backward Compatibility

A BGP speaker that doesn't understand the MSC will ignore it [RFC5492]. Section 3 recommends not terminating a session when only one peer supports the MSC. Instead, the operation will continue as specified in [I-D.chen-idr-bgp-over-quic].

7.2. Session Prioritization

One of the drawbacks of a single BGP session is that control plane messages for all supported Network Layer protocols use the same connection, which may cause resource contention.

QUIC [RFC9000] does not provide a mechanism for exchanging prioritization information. Instead, it recommends that implementations provide ways for an application to indicate the relative priority of streams, in this case, mapped to BGP sessions. An operator should prioritize BGP sessions (streams) that carry critical control plane information if the functionality is available. The definition of this functionality and the determination of the importance of a BGP session are both outside the scope of this document.

An example implementation is to have four priority (0-3) defined, and smaller number means higher priority. Each AFI/SAFI should be assigned a default priority and optional configuration to modify the default value. For example, IPv4 and IPv6 unicast AFI/SAFI (1/1 and 2/1) may have priority of 1, while BGP-LS (16388/71 and 16388/72) may have a priority of 3, and BGP FlowSpec (1/133 and 1/134) may have a priority of 4.

7.3. Other Considerations

A configuration command SHOULD be implemented to allow grouping of some AFI/SAFIs into one session.

8. Security Considerations

This document specifies how to establish multiple BGP sessions over a single QUIC connection. The general operation of BGP is not changed, nor is its security model. The security considerations of [I-D.chen-idr-bgp-over-quic] apply. Also, the non-TCP-related considerations of [RFC4271], [RFC4272], and [RFC7454] apply to the specification in this document.

By separating the control plane traffic over multiple sessions, the effect of a session-based vulnerability is reduced; only a single session is affected and not the whole connection. The result is increased resiliency.

On the other hand, a high number of BGP sessions may result in higher resource utilization and the risk of depletion. Also, more sessions may imply additional configuration and operational complexity. However, this risk is mitigated by the fact that BGP sessions typically require explicit configuration by the operator.

9. IANA Considerations

IANA is asked to assign a new Capability Code for the MultiStream Capability (Section 3) as follows:

Value	Description	Reference	Change Controller
TBD1	MultiStream Capability	[This Document]	IETF

Table 1: MultiStream Capability

IANA is asked to assign three values from the OPEN Message Error subcodes registry as follows:

Value	Name	Reference
TBD2	MultiSession Conflicty	[This Document]
TBD3	Session Capability Mismatch	[This Document]
TBD4	Network Layer Protocol Mismatch	[This Document]

Table 2

10. Acknowledgement

This document references the text and procedures defined in [I-D.ietf-idr-bgp-multisession], and we are grateful for their contributions.

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

11. References

11.1. Normative References

- [I-D.chen-idr-bgp-over-quic]
Chen, S., Zhang, Y., Wang, H., and Z. Li, "BGP Over QUIC", Work in Progress, Internet-Draft, draft-chen-idr-bgp-over-quic-00, 3 June 2021, <<https://www.ietf.org/archive/id/draft-chen-idr-bgp-over-quic-00.txt>>.
- [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>>.
- [RFC4486] Chen, E. and V. Gillet, "Subcodes for BGP Cease Notification Message", RFC 4486, DOI 10.17487/RFC4486, April 2006, <<https://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, <<https://www.rfc-editor.org/info/rfc5492>>.
- [RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, DOI 10.17487/RFC6286, June 2011, <<https://www.rfc-editor.org/info/rfc6286>>.
- [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>>.
- [RFC9000] Iyengar, J., Ed. and M. Thomson, Ed., "QUIC: A UDP-Based Multiplexed and Secure Transport", RFC 9000, DOI 10.17487/RFC9000, May 2021, <<https://www.rfc-editor.org/info/rfc9000>>.

11.2. Informative References

- [I-D.ietf-idr-bgp-multisession] Scudder, J., Appanna, C., and I. Varlashkin, "Multisession BGP", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-multisession-07, 13 September 2012, <<http://www.ietf.org/internet-drafts/draft-ietf-idr-bgp-multisession-07.txt>>.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<https://www.rfc-editor.org/info/rfc4272>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.
- [RFC7454] Durand, J., Pepelnjak, I., and G. Doering, "BGP Operations and Security", BCP 194, RFC 7454, DOI 10.17487/RFC7454, February 2015, <<https://www.rfc-editor.org/info/rfc7454>>.

Authors' Addresses

Alvaro Retana
Futurewei Technologies, Inc.
2330 Central Expressway
Santa Clara, CA 95050
United States of America
Email: aretana@futurewei.com

Yingzhen Qu
Futurewei Technologies, Inc.
2330 Central Expressway
Santa Clara, CA 95050
United States of America
Email: yingzhen.qu@futurewei.com

Jeff Tantsura
Microsoft
United States of America
Email: jefftant.ietf@gmail.com

Interdomain Routing Working Group
Internet-Draft
Intended status: Standards Track
Expires: 27 April 2022

H. Wang
M. Shen
J. Dong
Huawei Technologies
24 October 2021

Revised Error Handling for BGP Messages
draft-wang-idr-bgp-error-enhance-00

Abstract

This document supplements and revises RFC7606. According to RFC 7606, when an UPDATE packet received from a neighbor contains an attribute of incorrect format, the BGP session cannot be reset directly. Instead, the BGP session must be reset based on the specific problem. Error packets must minimize the impact on routes and do not affect the correctness of the protocol. Different error handling methods are used. The error handling methods include discarding attributes, withdrawing routes, disabling the address family, and resetting sessions.

RFC 7606 specifies the error handling methods of some existing attributes and provides guidance for error handling of new attributes.

This document supplements the error handling methods for common attributes that are not specified in RFC7606, and provides suggestions for revising the error handling methods for some attributes. The general principle remains unchanged: Maintain established BGP sessions and keep valid routes updated. However, discard or delete incorrect attributes or packets to minimize the impact on the current session.

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 27 April 2022.

Copyright Notice

Copyright (c) 2021 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. Scenarios	3
3. Error-Handling Procedures Update for NLRI	4
3.1. Prefix Length Error	4
3.2. Appears More Than Once	4
4. Error-Handling Procedures Update for Existing Attributes . .	5
4.1. Nexthop	5
4.2. MP_REACH_NLRI	5
4.3. Prefix SID	6
4.4. AGGREGATE and AS4_AGGREGATOR	6
4.5. ORIGINATOR_ID	6
4.6. Cluster-List	6
5. IANA Considerations	6
6. Security Considerations	6
7. References	7
7.1. Normative References	7
7.2. Informative References	7
Authors' Addresses	7

1. Introduction

According to RFC 4271, a BGP session that receives an UPDATE message containing a malformed attribute needs to reset the session that receives the malformed attribute.

According to our experience during maintenance, malformed packets may be incorrectly encapsulated due to software bugs or mis-understanding of standards in software development. Interrupting a neighbor causes neighbor flapping, which does not help solve the problem. The malformed packets may not be recognized by intermediate routers and cannot be incorrectly checked and propagated to other routers that establish sessions. When they reach the router that recognizes and checks the attribute, the neighbor flapping may also occurred. Even because routes are propagated multiple times, a route containing malformed packets may be received from multiple sessions at a checkpoint, causing multiple sessions to be reset, and the harm is multiplied.

For the preceding reasons, RFC 7606 defines a new method for processing incorrect UPDATE packets. If the Update packet received from a neighbor contains incorrect attributes, the BGP session cannot be reset directly. Instead, the BGP session needs to be handled in a specific manner based on the principle that incorrect packets affect routes as little as possible and do not affect protocol correctness. The error handling methods include discarding attributes, withdrawing routes, disabling the address family, and resetting sessions.

However, the error handling methods of some common attributes are not provided in RFC7606 or are different from those of vendors. This document supplements the error handling methods of some common attributes and provides suggestions for modifying the error handling methods of some attributes.

2. Scenarios

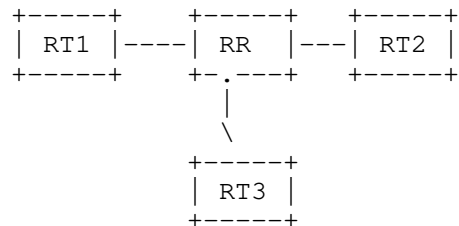


Figure 1 A simple network

Figure 1 shows a simple network. When RT3 has some software bugs or misunderstands the RFC, it may send a malformed packet. The RR receives the packet and considers it a malformed packet according to the error handling rules and resets the session. Later the session between RT3 and RR is re-established. RT3 will resend the packet to RR, and RR continues to repeat the previous action. This will happen continuously until the operator modifies the configuration, such as deleting the configuration about that new feature of the property, and sometimes they don't know how to correct the problem. During this process, RT3 services are unavailable. Frequent neighbor reestablishment and route updating also consumes more RR's system resources.

If the RR does not understand the attribute, the RR sends packets to RT1 and RT2. RT1 and RT2 may perform the same operation as RR. As a result, services between RT1 and RT2 are interrupted.

3. Error-Handling Procedures Update for NLRI

3.1. Prefix Length Error

According to [RFC7606], when a NLRI/UNLRI or MP_REACH_NLRI/MP_REACH_UNLRI with invalid length, eg, IPv4 Prefix length is more than 32, we must drop this Prefix and ignore the following Prefixes. We may keep the prefixes we have parsed correctly before.

Then we may also try to continue parse the next update packet if we can correctly find it.

The NLRI/UNLRI or MP_REACH_NLRI/MP_REACH_UNLRI with invalid length is malformed.

3.2. Appears More Than Once

[RFC7606] described like this:

If the MP_REACH_NLRI attribute or the MP_UNREACH_NLRI [RFC4760] attribute appears more than once in the UPDATE message, then a NOTIFICATION message MUST be sent with the Error Subcode "Malformed Attribute List".

Revised suggestion:

If the MP_REACH_NLRI attribute or the MP_UNREACH_NLRI [RFC4760] attribute appears more than once in the UPDATE message, only the last MP_REACH_NLRI/MP_UNREACH_NLRI SHOULD be processed, the others would be ignore.

4. Error-Handling Procedures Update for Existing Attributes

4.1. Nexthop

[RFC4271] define the IP address in the NEXT_HOP meet the following criteria to be considered semantically incorrect:

- a) It is the IP address of receiving speaker.
- b) The IP address is not EBGp directly neighbor's address or not share a common subnet with the receiving BGP speaker.

An update message with the case a) MAY be install to the RIB but treat as invalid.

Whether an update message with the case b) SHOULD be considered semantically incorrect depends on the user's configuration.

The following criteria also must to be considered semantically incorrect:

- c) The IP address is all zero.
- d) The IP address is all one.
- e) The IP address is multicast address(Class D) or reserved address (Class E).
- f) The IP address is not a invalid ip address.

An update message with the case c) to f) SHOULD be logged, and the route will be treat-as-withdraw.

4.2. MP_REACH_NLRI

[RFC7606] suggest to do "session reset" or "AFI/SAFI disable" approach. But this approach is too strict.

If the Length of Next Hop Network Address field of the MP_REACH attribute is inconsistent with that which was expected, the attribute is considered malformed. The whole MP_REACH attribute will be ignore and try to parse the next update packet. When it cannot correctly locate the next update packet, it will do the procedure suggested according to [RFC7606] . Otherwise, only the error SHOULD be logged and continued to do packet parsing.

An update message may both contained MP_REACH_NLRI and MP_REACH_UNLRI. If there are same Prefixes in both MP_REACH_NLRI and MP_REACH_UNLRI, the message SHOULD NOT be consider malformed. In this case, it should be firstly process the Prefixes in the MP_REACH_NLRI then process the Prefixes in the MP_REACH_UNLRI.

4.3. Prefix SID

According to [RFC8669], an update message containing a malformed or invalid BGP Prefix-SID attribute will be ignore and not advertise it to other BGP peers. But this procedure may lead to unexpected results.

The error handling is revised to be treat-as-withdraw.

4.4. AGGREGATE and AS4_AGGREGATOR

When the router-id in AGGREGATE or AS4_AGGREGATE attribute is zero, the attribute SHOULD be consider semantically incorrect, and the attribute SHOULD be logged and discard.

4.5. ORIGINATOR_ID

The error handling of [RFC4456] and [RFC7606] is revised as follows.

When the BGP Identifier in ORIGINATOR_ID attribute is zero, the attribute SHOULD be consider semantically incorrect, and the attribute SHOULD be logged and the UPDATE message SHALL be handled using the approach of "treat-as- withdraw".

4.6. Cluster-List

The error handling of [RFC4456] and [RFC7606] is revised as follows.

When the CLUSTER_ID value in ORIGINATOR_ID attribute is zero, the attribute SHOULD be consider semantically incorrect, and the attribute SHOULD be logged and the UPDATE message SHALL be handled using the approach of "treat-as- withdraw".

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

This document helps reduce the impact of malformed packets on the network and devices.

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, <<https://www.rfc-editor.org/info/rfc2119>>.

7.2. Informative References

- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [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>>.
- [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>>.
- [RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <<https://www.rfc-editor.org/info/rfc8669>>.

Authors' Addresses

Haibo Wang
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing
100095
China

Email: rainsword.wang@huawei.com

Ming Shen
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing

100095
China

Email: shenming2@huawei.com

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

Email: jie.dong@huawei.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: 30 October 2022

K. Zhang
Z. Hu
J. Dong
Huawei
28 April 2022

BGP SR Policy Extensions for template
draft-zhang-idr-sr-policy-template-01

Abstract

Segment Routing(SR) Policies can be advertised using BGP. An SR Policy may has lots of constraints, and as the application and features evolve, the SR Policy may need have more and more attribute constraints. To avoid modifying BGP when constraints are added to an SR Policy, we can define a template. The identifier and content of the template are defined by the receiver of the SR Policy. The advertiser of an SR policy only needs to know the ID of the template. When advertising SR policy, the advertiser carries the template ID in the tunnel encapsulation information of the SR policy. After receiving the SR Policy information, the receiver obtains the corresponding template and content according to the template ID, thereby obtaining abundant constraint configuration 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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

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

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

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

This Internet-Draft will expire on 30 October 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	2
2. Terminology	3
3. Template ID defination	3
4. SR Policy and Tunnel Encapsulation Attribute Update	3
4.1. Template ID sub-TLV	4
5. SR Policy Operations	5
5.1. Advertisement of SR Policies	5
5.2. Reception of an SR Policy	5
6. Acknowledgements	5
7. IANA Considerations	5
8. Security Considerations	5
9. References	5
Authors' Addresses	6

1. Introduction

[I-D.ietf-idr-segment-routing-te-policy] defines some attributes encoding of the SR Policy path. However, in actual applications, there are many other constraints of SR Policy path. These constraints are valid only on the device where the SR Policy path is installed. Such constraints may include backup protection, Bidirectional Forwarding Detection information, traffic statistics collection, or in-situ Flow Information Telemetry detection information, etc. If these constraints are directly delivered through BGP, the BGP SR Policy protocol may change frequently. This document defines a general method to carry the path constraints of SR Policies.

2. Terminology

SR Policy: An ordered list of segments.

Candidate Path: the unit for signaling of an SR Policy to a headend via protocol extensions like Path Computation Element (PCE) Communication Protocol (PCEP) [RFC8664] [I-D.ietf-pce-segment-routing-policy-cp] or BGP SR Policy [I-D.ietf-idr-segment-routing-te-policy].

SRPM: SR Policy Module.

Template: A collection of constraints sets.

Template ID: The identifier of a template.

3. Template ID defination

To support the constraints extension of SR Policies, this document defines a constraint template identifier. The constraint template ID is valid only for the recipient. The SR policy publisher only needs to carry the template ID when publishing the SR policy. The receiver of the SR Policy may create a template corresponding to the template identifier in advance before receiving the SR Policy, or may define a corresponding template after receiving the template definition of the SR Policy. The template can contain any constraints on the SR Policy path, including but not limited to backup protection, Bidirectional Forwarding Detection information, traffic statistics collection, or in-situ Flow Information Telemetry detection information, etc. After receiving the SR Policy information, the receiver matches the template information based on the template ID and adds constraints to the SR Policy based on the constraints defined in the template.

4. SR Policy and Tunnel Encapsulation Attribute Update

As the template ID is defined, the tunnel attribute encapsulation of the BGP SR Policy needs to be updated.

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
    Priority
    Policy Name
    Policy Candidate Path Name
    Explicit NULL Label Policy (ENLP)
    Template ID
    Segment List
      Weight
      Segment
      Segment
      ....
    ....

```

Where Template ID indicates the template ID for the SR Policy candidate path.

4.1. Template ID sub-TLV

A new sub-TLV called Template ID sub-TLV is defined. Template ID sub-TLV specifies the template ID of an SR policy candidate path. Each sub-TLV is encoded as shown in Figure 1.

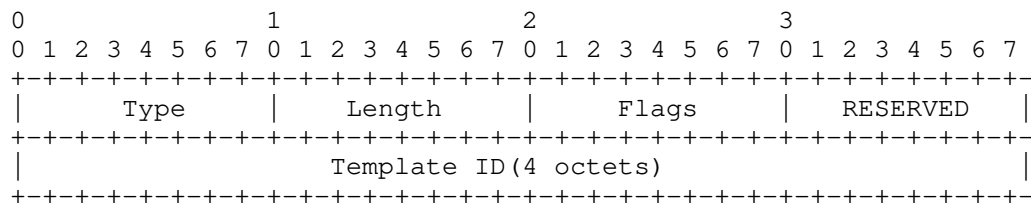


Figure 1: Figure 1: Template ID Sub-TLV

Type: Template ID, 1 octet, TBD.

Length: 6.

Flags: 1 octet of flags. None are defined at this stage. Flags SHOULD be set to zero on transmission and MUST be ignored on receipt.

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

Template ID: a 4-octet value.

5. SR Policy Operations

5.1. Advertisement of SR Policies

When BGP advertises an SR Policy, different candidate paths of the same SR Policy may have different template IDs or the same template ID, depending on the constraints required by the candidate paths of the SR Policy.

5.2. Reception of an SR Policy

When a BGP speaker receives an SR Policy NLRI from a neighbor, BGP Speaker determines if it's acceptable as described in [I-D.ietf-idr-segment-routing-te-policy]. Once BGP on the receiving node has determined that the SR Policy NLRI is usable, it passes the SR Policy candidate path to the SRPM. The SRPM then determine how to use the template ID in SR Policy.

The SRPM should find the template by template ID, and determines the constraints to use when install the candidate path. If there is no template find, the SRPM should ignore the template ID and use the candidate path as there is no template ID.

6. Acknowledgements

TBD.

7. IANA Considerations

This document requests that IANA allocates a new sub-TLV type as defined in Section 4.1 from the "Sub-TLVs for SR Policy" registry as specified.

Value	Description	Reference
TBD	SR Policy Template ID	This document

Figure 2: Figure 2: Template ID sub-TLV

8. Security Considerations

These extensions to BGP SR Policy do not add any new security issues to the existing protocol.

9. References

[I-D.ietf-idr-segment-routing-te-policy]

Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-segment-routing-te-policy-17, 14 April 2022, <<https://www.ietf.org/archive/id/draft-ietf-idr-segment-routing-te-policy-17.txt>>.

[I-D.ietf-pce-segment-routing-policy-cp]

Koldychev, M., Sivabalan, S., Barth, C., Peng, S., and H. Bidgoli, "PCEP extension to support Segment Routing Policy Candidate Paths", Work in Progress, Internet-Draft, draft-ietf-pce-segment-routing-policy-cp-07, 21 April 2022, <<https://www.ietf.org/archive/id/draft-ietf-pce-segment-routing-policy-cp-07.txt>>.

[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>>.

[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>>.

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

Authors' Addresses

Ka Zhang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China
Email: zhangka@huawei.com

Zhibo Hu
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China

Email: huzhibo@huawei.com

Jie Dong
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China
Email: jie.dong@huawei.com

IDR
Internet-Draft
Intended status: Standards Track
Expires: 2 September 2022

Y. Liu
S. Peng
ZTE
1 March 2022

BGP Extension for SR-MPLS Entropy Label Position
draft-zhou-idr-bgp-srmppls-elp-04

Abstract

This document proposes extensions for BGP to indicate the entropy label position in the SR-MPLS label stack when delivering SR Policy via BGP.

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 2 September 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	2
2. Conventions used in this document	3
2.1. Requirements Language	3
2.2. Terminology and Acronyms	3
3. Entropy Label Position in SR-MPLS with the Controller	3
4. BGP Extensions for ELP in SR Policy	5
5. Operations	5
6. IANA Considerations	5
7. Security Considerations	6
8. References	6
8.1. Normative References	6
8.2. Informative References	6
Authors' Addresses	7

1. Introduction

Segment Routing (SR) leverages the source routing paradigm. Segment Routing can be instantiated on MPLS data plane which is referred to as SR-MPLS [RFC8660]. SR-MPLS leverages the MPLS label stack to construct the SR path.

Entropy labels (ELs) [RFC6790] are used in the MPLS data plane to provide entropy for load-balancing. The idea behind the entropy label is that the ingress router computes a hash based on several fields from a given packet and places the result in an additional label named "entropy label". Then, this entropy label can be used as part of the hash keys used by an LSR. Using the entropy label as part of the hash keys reduces the need for deep packet inspection in the LSR while keeping a good level of entropy in the load-balancing.

[RFC8662] proposes to use entropy labels for SR-MPLS networks and multiple <ELI, EL> pairs may be inserted in the SR-MPLS label stack. The ingress node may decide the number and position of the ELI/ELs which need to be inserted into the label stack, that is termed as ELP (Entropy Label Position) in this document. But in some cases, the controller (e.g. PCE) can be used to perform the TE path computation as well as the Entropy Label Position which is useful for inter-domain scenarios.

[I-D.ietf-idr-segment-routing-te-policy] specifies the way to use BGP to distribute one or more of the candidate paths of an SR Policy to the headend of that policy.

This document proposes extensions for BGP to indicate the ELP in the segment list when delivering SR Policy via BGP in SR-MPLS networks.

2. Conventions used in this document

2.1. 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 [RFC2119].

2.2. Terminology and Acronyms

EL: Entropy Label

ELI: Entropy Label Indicator

ELC: Entropy Label Capability

ERLD: Entropy Readable Label Depth

ELP: Entropy Label Position

MSD: Maximum SID Depth

3. Entropy Label Position in SR-MPLS with the Controller

As described in [RFC8662] section 7, ELI/EL placement is not an easy decision, multiple criteria may be taken into account.

First is the Maximum SID Depth (MSD), it defines the maximum number of labels that a particular node can impose on a packet, and it is a limit when the ingress node imposing ELI/EL pairs on the SR label stack.

The Entropy Readable Label Depth(ERLD) value is another important parameter to consider when inserting an ELI/EL. The ERLD is defined as the number of labels a router can both read in an MPLS packet received on its incoming interface(s) and use in its load-balancing function. An ELI/EL pair must be within the ERLD of the LSR in order for the LSR to use the EL during load-balancing. It's necessary to get the ERLD of the nodes along the SR path to achieve efficient load-balancing.

An implementation MAY try to evaluate if load-balancing is really expected at a particular node based on the segment type of its label, which also influences the ELP of a segment list.

Other criteria includes maximizing number of LSRs that will load-balance, preference for a part of the path, and etc. Using which criteria and how to decide the ELP based on the criteria is a matter of implementation.

As shown in Figure 1, in the inter-domain scenario, a path from A to Z is required, a centralized controller performs the computation of the end-to-end path, along which traffic load-balancing is required.

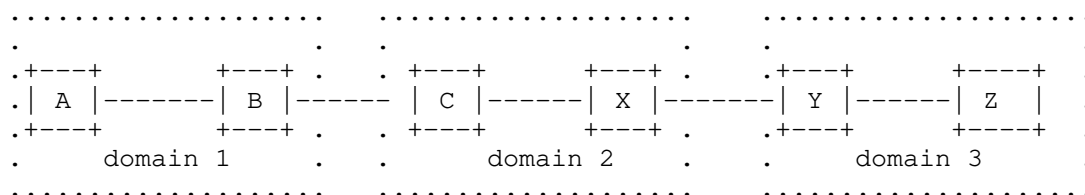


Figure 1: Entropy Labels in SR-MPLS Inter-Domain Scenario

When the headend node in the first domain can't get the information of the nodes/SIDs in other domains, e.g, the ERLD of each node or the type of the SID bounded to a node/link, it's difficult for the headend node to decide the ELP of the segment list for the path.

Performing the computation of the ELP by the controller is an alternate, since it's easier for the controller to get the required information along the segment list prescribed by itself.

For example, the ERLD value can be advertised via IS-IS[I-D.ietf-isis-mpls-elc] and OSPF[I-D.ietf-ospf-mpls-elc] within the domain, in each domain, one or more nodes are configured with BGP-LS so the controller can get the ERLD value of all the nodes through BGP-LS[RFC9085]. The controller can acquire the MSD of the headend node or the Binding SID anchor node via BGP-LS[RFC8814] or PCEP[RFC8664].

Another benefit of utilizing the controller to calculate ELP is that if the criteria or calculation algorithm is changed, the corresponding modification only needs to be made on the controller instead of each headend node in the network.

When the controller performs the computation of the the ELP for a segment list, the considerations for the placement of ELI/ELs introduced in [RFC8662] are still applicable. How the controller computes the ELP is out of scope of the document.

After the ELP of an SR path is decided, the controller SHOULD inform the result to the headend node of the path, so the node knows where to insert the ELI/ELs when needed. Section 4 proposes the detailed extensions for BGP to carry this information.

4. BGP Extensions for ELP in SR Policy

The Segment Flags for Segment Sub-TLVs are defined in Section 2.4.4.2.12 of [I-D.ietf-idr-segment-routing-te-policy]. In this document, the ELP information is transmitted by extending the flags of Segment Sub-TLVs.

```

      0 1 2 3 4 5 6 7
    +--+--+--+--+--+--+
    |V|A|S|B|E|   |
    +--+--+--+--+--+--+

```

E-Flag: This flag, when set, indicates that presence of < ELI, EL> label pairs which are inserted after this segment. E-Flag is applicable to Segment Types A, C, D, E, F, G and H. If E-Flag appears with Segment Types B, I, J and K, it MUST be ignored.

5. Operations

Node A receives an SR Policy NLRI with an Segment List sub-TLV from the controller. The Segment List sub-TLV contains multiple Segment sub-TLVs, e.g, <S1, S2, S3, S4, S5, S6>, the E-Flags of S3 and S6 are set, it indicates that if load-balancing is required, two <ELI, EL> pairs SHOULD be inserted into the label stack of the SR-TE forwarding entry, respectively after the Label for S3 and Label for S6.

The value of EL is supplemented by the ingress node according to load-balancing function of the appropriate keys extracted from a given packet. After inserting ELI/ELs, the label stack on the ingress node would be <S1, S2, S3, ELI, EL, S4, S5, S6, ELI, EL>.

6. IANA Considerations

This document requests bit 4 for Entropy Label Flag in "SR Policy Segment Flags" under the "BGP Tunnel Encapsulation" registry.

Bit	Description	Reference
4	Entropy Label Position Flag(E-Flag)	This document

7. Security Considerations

Procedures and protocol extensions defined in this document do not introduce any new security considerations beyond those already listed in [RFC8662] and [I-D.ietf-idr-segment-routing-te-policy].

8. References

8.1. Normative References

- [I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", Work in Progress, Internet-Draft, draft-ietf-idr-segment-routing-te-policy-14, 10 November 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-segment-routing-te-policy-14>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <<https://www.rfc-editor.org/info/rfc4360>>.
- [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, <<https://www.rfc-editor.org/info/rfc6790>>.
- [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>>.
- [RFC8662] Kini, S., Kompella, K., Sivabalan, S., Litkowski, S., Shakir, R., and J. Tantsura, "Entropy Label for Source Packet Routing in Networking (SPRING) Tunnels", RFC 8662, DOI 10.17487/RFC8662, December 2019, <<https://www.rfc-editor.org/info/rfc8662>>.

8.2. Informative References

- [I-D.ietf-isis-mpls-elc]
Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using IS-IS", Work in Progress, Internet-Draft, draft-ietf-isis-mpls-elc-01, 10 November 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-isis-mpls-elc-01>>.

Progress, Internet-Draft, draft-ietf-isis-mpls-elc-13, 28 May 2020, <<https://datatracker.ietf.org/doc/html/draft-ietf-isis-mpls-elc-13>>.

[I-D.ietf-ospf-mpls-elc]

Xu, X., Kini, S., Psenak, P., Filsfils, C., Litkowski, S., and M. Bocci, "Signaling Entropy Label Capability and Entropy Readable Label Depth Using OSPF", Work in Progress, Internet-Draft, draft-ietf-ospf-mpls-elc-15, 1 June 2020, <<https://datatracker.ietf.org/doc/html/draft-ietf-ospf-mpls-elc-15>>.

[RFC8476] Tantsura, J., Chunduri, U., Aldrin, S., and P. Psenak, "Signaling Maximum SID Depth (MSD) Using OSPF", RFC 8476, DOI 10.17487/RFC8476, December 2018, <<https://www.rfc-editor.org/info/rfc8476>>.

[RFC8491] Tantsura, J., Chunduri, U., Aldrin, S., and L. Ginsberg, "Signaling Maximum SID Depth (MSD) Using IS-IS", RFC 8491, DOI 10.17487/RFC8491, November 2018, <<https://www.rfc-editor.org/info/rfc8491>>.

[RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <<https://www.rfc-editor.org/info/rfc8660>>.

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

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

[RFC9085] Previdi, S., Talaulikar, K., Ed., Filsfils, C., Gredler, H., and M. Chen, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing", RFC 9085, DOI 10.17487/RFC9085, August 2021, <<https://www.rfc-editor.org/info/rfc9085>>.

Authors' Addresses

Yao Liu
ZTE
Nanjing
China
Email: liu.yao71@zte.com.cn

Shaofu Peng
ZTE
Nanjing
China
Email: peng.shaofu@zte.com.cn