Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: 10 September 2023

K. Talaulikar
A. MahendraBabu
Cisco Systems

9 March 2023

# Advertising Flexible Algorithm Extensions in BGP Link-State
## draft-aravindbabu-idr-bgp-ls-flex-algo-ext-00

Abstract

Flexible Algorithm is a solution that allows some routing protocols (e.g., OSPF and IS-IS) to compute paths over a network based on user-defined (and hence, flexible) constraints and metrics. The computation is performed by routers participating in the specific network in a distributed manner using a Flexible Algorithm Definition. This Definition is provisioned on one or more routers and propagated through the network by OSPF and IS-IS flooding.

BGP Link-State (BGP-LS) enables the collection of various topology information from the network. RFC9351 introduced BGP-LS support for the advertisement of Flexible Algorithm Definition as a part of the topology information from the network. This document specifies the advertisement of further Flexible Algorithm related extensions in BGP-LS.

Table of Contents



## 1. Introduction

Flexible Algorithm is a solution that allows some routing protocols (e.g., OSPF and IS-IS) to compute paths over a network based on user-defined (and hence, flexible) constraints and metrics. The computation is performed by routers participating in the specific network in a distributed manner using a Flexible Algorithm Definition. This Definition is provisioned on one or more routers and propagated through the network by OSPF and IS-IS flooding. [RFC9350] defines the base Flexible Algorithm solution that allows IGPs themselves to compute constraint-based paths over the network.

The extensions to BGP-LS [RFC7752] for the advertisement of the Flexible Algorithm Definition (FAD) information to enable learning of the mapping of the flexible algorithm number to its Definition in each area/domain of the underlying IGPs are defined in [RFC9351].

This document defines further extensions to BGP-LS for Flexible Algorithm as below:

* The extensions to the Flexible Algorithm so that it can be used with the regular IPv4 and IPv6 forwarding as defined for IGPs in [I-D.ietf-lsr-ip-flexalgo].

* The Flexible Algorithm Definition that is used to exclude based on bandwidth and metric constraints and to automatically calculate metrics for use in SPF calculation as defined for IGPs in [I-D.ietf-lsr-flex-algo-bw-con].

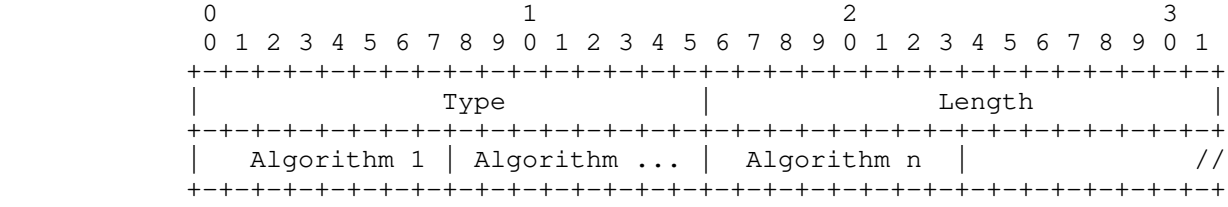
* The Flexible Algorithm Definition that is used to include/exclude links in the reverse direction of the traffic flow for SPF calculation as defined for IGPs in [I-D.ppsenak-lsr-igp-flex-algo-reverse-affinity].

## 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. Advertising IP Algorithm Participation

The IP Algorithm TLV is a BGP-LS Attribute TLV associated with the Node NLRI that is used for the algorithms associated with a given node. The format of this TLV is as follows:




where:

Figure 1: IP Algorithm TLV

* Type: TBA

* Length: Variable

* Algorithm: Single octet algorithm value between 128 and 255 inclusive.

The IP Algorithm TLV is derived from the following IGP protocol-specific advertisements:

* In the case of IS-IS, from the IS-IS IP Algorithm sub-TLV defined in [I-D.ietf-lsr-ip-flexalgo].

* In the case of OSPFv2/OSPFv3, from the OSPF IP Algorithm sub-TLV defined in [I-D.ietf-lsr-ip-flexalgo].

The IP Algorithm TLV is optional and it MUST NOT be advertised more than once in the BGP-LS Attribute. If multiple instances are present, then the first one MUST be considered valid, and the rest MUST be ignored.

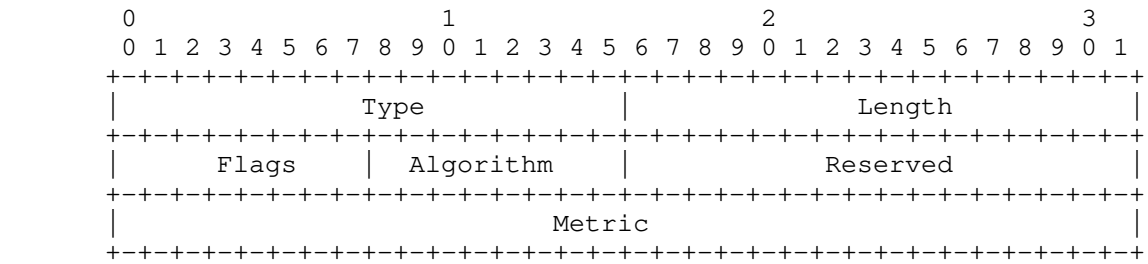
## 3. Advertising IP Algorithm Reachability

The normal or base (i.e., algorithm 0) prefix reachabilities are done using the BGP-LS IPv4/IPv6 Topology Prefix NLRIs defined in [RFC7752] along with its associated IGP metric carried within the IGP Metric TLV (TLV 1095) in the BGP-LS Attributed associated with the NLRI. The presence of IGP Metric TLV is what identifies the base/normal prefix reachability.

The IP algorithm-specific reachability advertisements are also done using the BGP-LS IPv4/IPv6 Topology Prefix NLRIs. However, these algorithm-specific advertisements MUST NOT carry an IGP Metric TLV along with them. Instead, the metric associated with the IP algorithm-specific prefix reachability is carried within the TLVs introduced in the following subsections.

A BGP-LS Consumer receiving an IPv4/IPv6 Topology Prefix NLRI advertisement that carries both an IGP Metric TLV along with any of the TLVs introduced in the following subsections, MUST consider it as a normal (i.e., algorithm 0) prefix reachability advertisement and MUST ignore all TLVs corresponding to algorithm-specific prefix reachability advertisements.

The IP Algorithm Prefix Reachability TLV is a BGP-LS Attribute TLV associated with the IPv4/IPv6 Topology Prefix NLRI that is used for the advertisement of the algorithm-specific prefix reachability from a given node. The format of this TLV is as follows:




where:

Figure 2: IP Algorithm Prefix Reachability TLV

* Type: TBA

* Length: 8

* Flags: 1 octet of flags. The flags are copied from the IS-IS IPv4/IPv6 Algorithm Prefix Reachability TLV [I-D.ietf-lsr-ip-flexalgo] or the OSPFv2/OSPFv3 IP Algorithm Prefix Reachability sub-TLV [I-D.ietf-lsr-ip-flexalgo] in the case of IS-IS or OSPFv2/OSPFv3 respectively.

* Algorithm: 1 octet value providing the Associated Algorithm from 128 to 255.

* Reserved: 2 octet value that MUST be set to 0 by the originator and ignored by the receiver.

* Metric: The algorithm-specific metric value.

The IP Algorithm Prefix Reachability TLV is derived from the following IGP protocol-specific advertisements:

* In the case of IS-IS, from the IS-IS IPv4/IPv6 Algorithm Prefix Reachability TLV defined in [I-D.ietf-lsr-ip-flexalgo]. The sub-TLVs are encoded using the BGP-LS Attribute TLVs defined for the IPv4/IPv6 Topology Prefix NLRI.

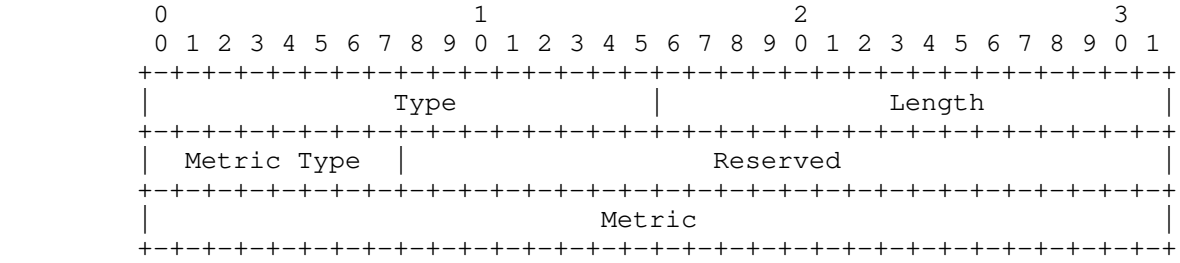
* In the case of OSPF, from the OSPFv2/OSPFv3 IP Algorithm Prefix Reachability sub-TLV defined in [I-D.ietf-lsr-ip-flexalgo].

The Multi-topology ID (MTID) associated with the underlying IGP advertisements is encoded using the Multi-Topology Identifier TLV (TLV 263) [RFC7752] as a Prefix Descriptor TLV when the advertisement is associated with a non-default topology. The IP Prefix value itself is encoded using the IP Reachability Information TLV (TLV 265) [RFC7752] as a Prefix Descriptor TLV.

The IP Algorithm Prefix Reachability TLV MUST NOT be advertised more than once in the BGP-LS Attribute. If multiple instances are present, then the first one MUST be considered valid, and the rest MUST be ignored.

## 4. Advertising Generic Metric for Links

The Generic Metric TLV is a BGP-LS Attribute TLV associated with the Link NLRI that is used for the advertisement of the generic metric(s) associated with a link. The format of this TLV is as follows:




where:

Figure 3: Generic Metric TLV

* Type: TBA

* Length: 8.

* Metric Type: 1 octet that carries a metric type from the IGP Metric Type registry

* Reserved: 3 octet value that MUST be set to 0 by the originator and ignored by the receiver.

* Metric: 4-octet field carrying the metric value. In the case of IS-IS, the value MUST be in the range of 0 - 16,777,215.

The Generic Metric TLV is derived from the following IGP protocol-specific advertisements:

* In the case of IS-IS, from the IS-IS Generic Metric sub-TLV defined in [I-D.ietf-lsr-flex-algo-bw-con].

* In the case of OSPF, from the OSPF Generic Metric sub-TLV defined in [I-D.ietf-lsr-flex-algo-bw-con].

The Generic Metric TLV MAY be advertised more than once in the BGP-LS Attribute, one for each metric type. If multiple instances are present for the same metric type, then the first one MUST be considered valid, and the rest MUST be ignored.

## 5. Advertising Flexible Algorithm Definition Extensions

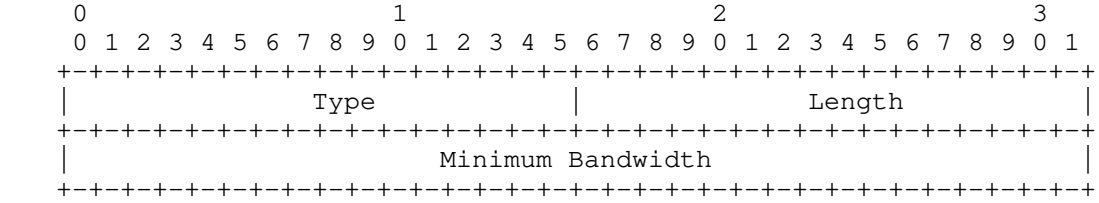
[RFC9351] introduced the Flexible Algorithm Definition (FAD) TLV that is advertised in the BGP-LS Attribute along with the Node NLRI for the advertisement of the Flexible Algorithm definition advertised by a given node in IGPs.

The following subsections define new sub-TLVs of the FAD TLV to cover further extensions of the IGP Flexible Algorithm solution.

### 5.1. FAD Exclude Minimum Bandwidth Sub-TLV

The FAD Exclude Minimum Bandwidth sub-TLV is an optional sub-TLV that is used to carry the minimum bandwidth associated with the FAD that are used in the computation of the specific algorithm as described in [I-D.ietf-lsr-flex-algo-bw-con].

The sub-TLV has the following format:




where:

Figure 4: FAD Exclude Minimum Bandwidth sub-TLV

* Type: TBA

* Length: 4 octets.

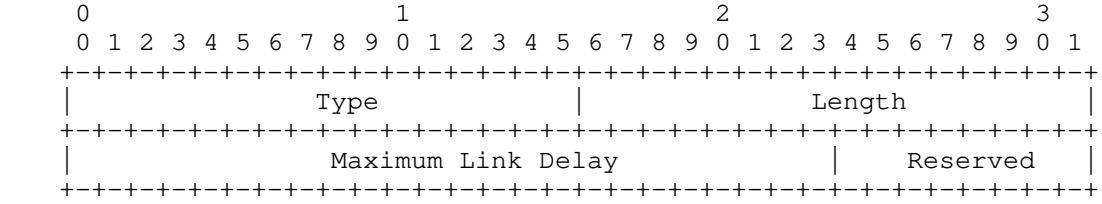
* Min Bandwidth: The minimum link bandwidth is encoded in 32 bits in IEEE floating point format. The units are bytes per second.

The information in the FAD Exclude Minimum Bandwidth sub-TLV is derived from the IS-IS and OSPF protocol-specific FAD Exclude Minimum Bandwidth sub-TLVs as defined in [I-D.ietf-lsr-flex-algo-bw-con].

## 5.2. FAD Exclude Maximum Link Delay Sub-TLV

The FAD Exclude Maximum Link Delay sub-TLV is an optional sub-TLV that is used to carry the maximum link delay information associated with the FAD that is used in the computation of the specific algorithm as described in [I-D.ietf-lsr-flex-algo-bw-con].

The sub-TLV has the following format:




where:

Figure 5: FAD Exclude Maximum Link Delay sub-TLV

* Type: TBA

* Length: 4 octets.

* Maximum Link Delay: The maximum link delay is encoded in microseconds.

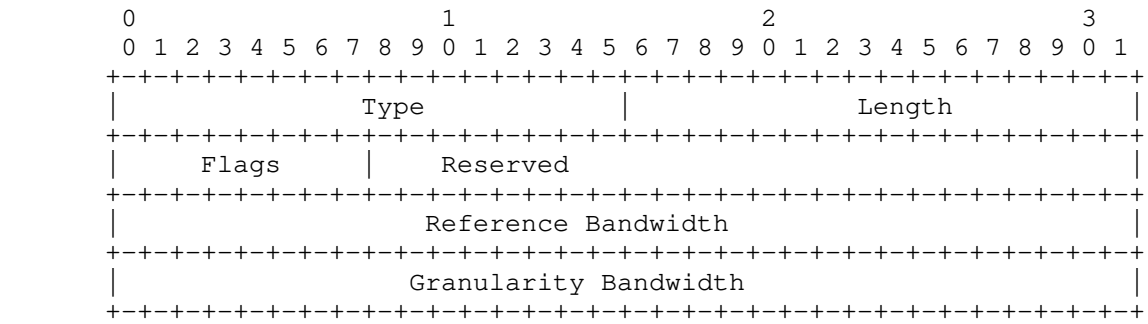
* Reserved: 1 octet field that MUST be set to 0 by the originator and ignored by the receiver.

The information in the FAD Exclude Maximum Link Delay sub-TLV is derived from the IS-IS and OSPF protocol-specific FAD Exclude Maximum Link Delay sub-TLVs as defined in [I-D.ietf-lsr-flex-algo-bw-con].

## 5.3. FAD Reference Bandwidth Sub-TLV

The FAD Reference Bandwidth sub-TLV is an optional sub-TLV that is used to carry the information needed for the reference bandwidth method of metric calculation associated with the FAD that is used in the computation of the specific algorithm as described in [I-D.ietf-lsr-flex-algo-bw-con].

The sub-TLV has the following format:




where:

Figure 6: FAD Reference Bandwidth sub-TLV

* Type: TBA

* Length: 12 octets.

* Flags: 1 octet of flags. The flags are copied from the IS-IS FAD Reference Bandwidth sub-TLV [I-D.ietf-lsr-flex-algo-bw-con] or the OSPF FAD Reference Bandwidth sub-TLV [I-D.ietf-lsr-flex-algo-bw-con] in the case of IS-IS or OSPF respectively.

* Reserved: 3 octet field that MUST be set to 0 by the originator and ignored by the receiver.

* Reference Bandwidth: The reference bandwidth is encoded in 32 bits in IEEE floating point format. The units are bytes per second.

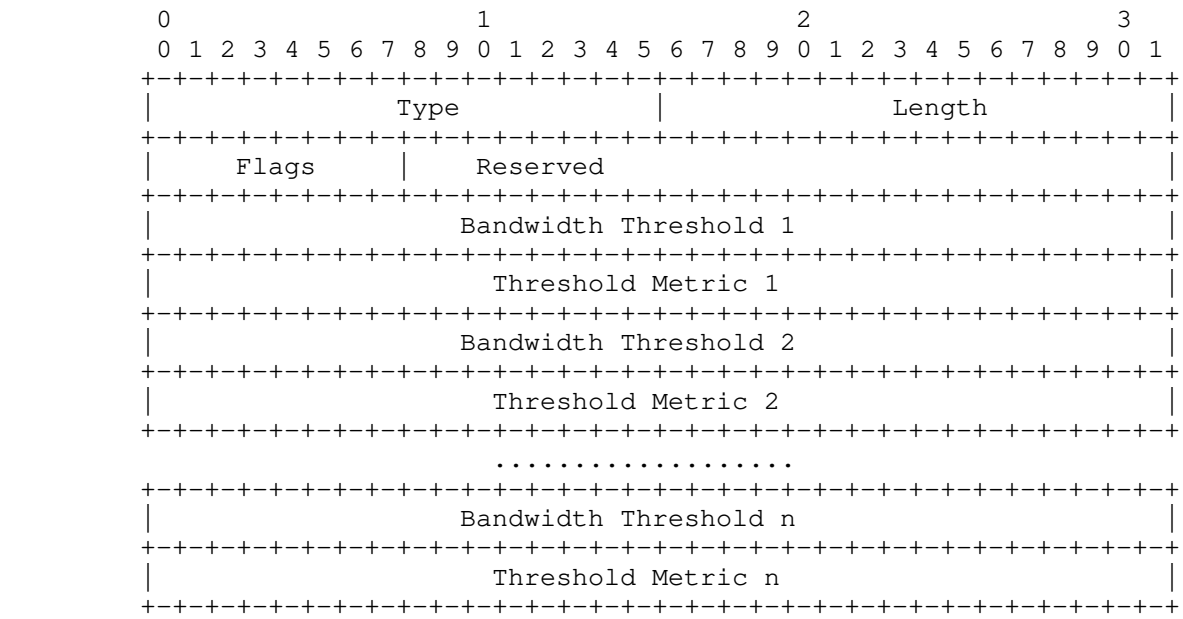
* Granularity Bandwidth: The granularity bandwidth is encoded in 32 bits in IEEE floating point format. The units are bytes per second.

The information in the FAD Reference Bandwidth sub-TLV is derived from the IS-IS and OSPF protocol-specific FAD Reference Bandwidth sub-TLV as defined in [I-D.ietf-lsr-flex-algo-bw-con].

## 5.4. FAD Bandwidth Thresholds Sub-TLV

The FAD Bandwidth Thresholds sub-TLV is an optional sub-TLV that is used to carry the information needed for bandwidth threshold method of metric calculation associated with the FAD that are used in the computation of the specific algorithm as described in [I-D.ietf-lsr-flex-algo-bw-con].

The sub-TLV has the following format:




where:

Figure 7: FAD Bandwidth Thresholds sub-TLV

* Type: TBA

* Length: 4 + (n * 8) octets. Here n is equal to the number of Threshold Metrics specified. n MUST be greater than or equal to 1.

* Flags: 1 octet of flags. The flags are copied from the IS-IS FAD Bandwidth Thresholds sub-TLV [I-D.ietf-lsr-flex-algo-bw-con] or the OSPF FAD Bandwidth Thresholds sub-TLV [I-D.ietf-lsr-flex-algo-bw-con] in the case of IS-IS or OSPF respectively.

* Reserved: 3 octet field that MUST be set to 0 by the originator and ignored by the receiver.

* Bandwidth Threshold (1 ... n): The bandwidth threshold is encoded in 32 bits in IEEE floating point format. The units are bytes per second.

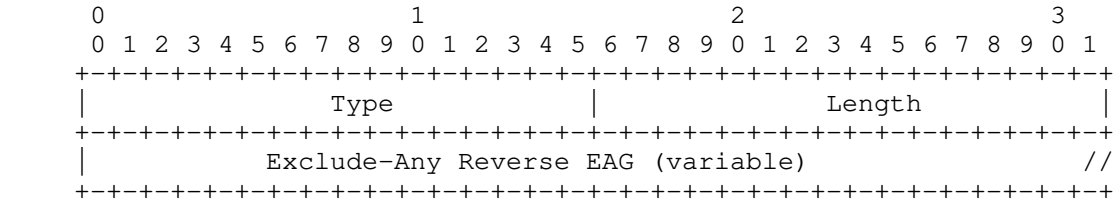
* Threshold Metric (1 ... n): 4 octet field carrying the threshold metric value. In the case of IS-IS, the value MUST be in the range of 0 - 16,777,215.

The information in the FAD Bandwidth Thresholds sub-TLV is derived from the IS-IS and OSPF protocol-specific FAD Bandwidth Thresholds sub-TLV as defined in [I-D.ietf-lsr-flex-algo-bw-con].

## 5.5. Flexible Algorithm Exclude-Any Reverse Affinity Sub-TLV

The Flexible Algorithm Exclude-Any Reverse Affinity sub-TLV is an optional sub-TLV that is used to carry the reverse affinity constraints associated with the FAD and enable the exclusion of links carrying any of the specified affinities from the computation of the specific algorithm as described in [I-D.ppsenak-lsr-igp-flex-algo-reverse-affinity]. The affinity is expressed in terms of Extended Admin Group (EAG) as defined in [RFC7308].

The sub-TLV has the following format:




where:

Figure 8: Flexible Algorithm Exclude-Any Reverse Affinity sub-TLV

* Type: TBA

* Length: The total length of the value field in octets dependent on the size of the EAG. It MUST be a non-zero value and a multiple of 4.

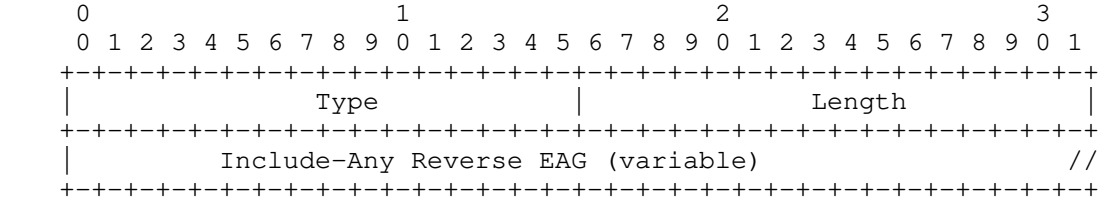
* Exclude-Any Reverse EAG: the EAG value.

The information in the Flexible Algorithm Exclude Any Reverse Affinity sub-TLV is derived from the IS-IS and OSPF protocol-specific Flexible Algorithm Exclude Admin Group sub-TLV as defined in [I-D.ppsenak-lsr-igp-flex-algo-reverse-affinity].

### 5.6. Flexible Algorithm Include-Any Reverse Affinity Sub-TLV

The Flexible Algorithm Include-Any Reverse Affinity sub-TLV is an optional sub-TLV that is used to carry the affinity constraints associated with the FAD and enable the inclusion of links carrying any of the specified affinities in the computation of the specific algorithm as described in [I-D.ppsenak-lsr-igp-flex-algo-reverse-affinity]. The affinity is expressed in terms of Extended Admin Group (EAG) as defined in [RFC7308].

The sub-TLV has the following format:




where:

Figure 9: Flexible Algorithm Include-Any Reverse Affinity sub-TLV

* Type: TBA

* Length: The total length of the value field in octets dependent on the size of the EAG. It MUST be a non-zero value and a multiple of 4.

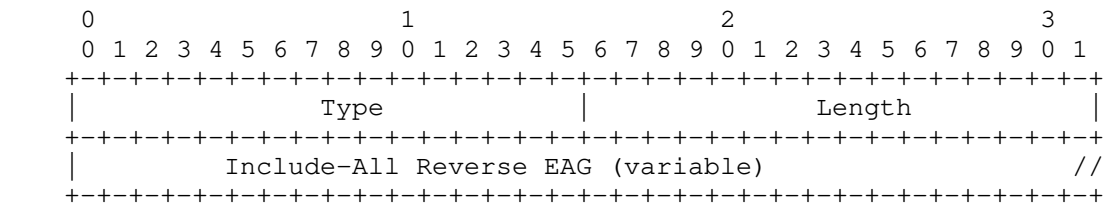
* Include-Any EAG: the EAG value.

The information in the Flexible Algorithm Include-Any Reverse Affinity sub-TLV is derived from the IS-IS and OSPF protocol-specific Flexible Algorithm Include-Any Reverse Admin Group sub-TLV as defined in [I-D.ppsenak-lsr-igp-flex-algo-reverse-affinity].

### 5.7. Flexible Algorithm Include-All Reverse Affinity Sub-TLV

The Flexible Algorithm Include-All Reverse Affinity sub-TLV is an optional sub-TLV that is used to carry the affinity constraints associated with the FAD and enable the inclusion of links carrying all of the specified affinities in the computation of the specific algorithm as described in [I-D.ppsenak-lsr-igp-flex-algo-reverse-affinity]. The affinity is expressed in terms of Extended Admin Group (EAG) as defined in [RFC7308].

The sub-TLV has the following format:




where:

Figure 10: Flexible Algorithm Include-All Reverse Affinity sub-TLV

* Type: TBA

* Length: The total length of the value field in octets dependent on the size of the EAG. It MUST be a non-zero value and a multiple of 4.

* Include-All EAG: the EAG value.

The information in the Flexible Algorithm Include-All Reverse Affinity sub-TLV is derived from the IS-IS and OSPF protocol-specific Flexible Algorithm Include-All Reverse Admin Group sub-TLV as defined in [I-D.ppsenak-lsr-igp-flex-algo-reverse-affinity].

## 6. IANA Considerations

This document requests IANA to allocate code points from the "BGP-LS NLRI and Attribute TLVs" sub-registry of the "Border Gateway Protocol - Link-State (BGP-LS) Parameters" registry group.

| Code Point | Description | Reference |
|---|---|---|
| TBA | IP Algorithm | this document |
| TBA | IP Algorithm Prefix Reachability | this document |
| TBA | Generic Metric | this document |
| TBA | Flexible Algorithm Exclude Minimum Bandwidth | this document |
| TBA | Flexible Algorithm Exclude Maximum Link Delay | this document |
| TBA | Flexible Algorithm Reference Bandwidth | this document |
| TBA | Flexible Algorithm Bandwidth Thresholds | this document |
| TBA | Flexible Algorithm Exclude Any Reverse Affinity | this document |
| TBA | Flexible Algorithm Include Any Reverse Affinity | this document |
| TBA | Flexible Algorithm Include All Reverse Affinity | this document |

Figure 11: BGP-LS Flexible Algorithm Extensions Code Points

## 7. Manageability Considerations

This document does not introduce any new manageability considerations beyond those covered by [RFC9351].

## 8. Security Considerations

This document does not introduce any new security considerations beyond those covered by [RFC9351].

## 9. Acknowledgements

## 10. References

### 10.1. Normative References

[I-D.ietf-lsr-flex-algo-bw-con]
Hegde, S., Britto, W., Shetty, R., Decraene, B., Psenak, P., and T. Li, "Flexible Algorithms: Bandwidth, Delay, Metrics and Constraints", Work in Progress, Internet-Draft, draft-ietf-lsr-flex-algo-bw-con-05, 16 January 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-lsr-flex-algo-bw-con-05>.

[I-D.ietf-lsr-ip-flexalgo]
Britto, W., Hegde, S., Kaneriya, P., Shetty, R., Bonica, R., and P. Psenak, "IGP Flexible Algorithms (Flex-Algorithm) In IP Networks", Work in Progress, Internet-Draft, draft-ietf-lsr-ip-flexalgo-08, 19 December 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-lsr-ip-flexalgo-08>.

[I-D.ppsenak-lsr-igp-flex-algo-reverse-affinity]
Psenak, P., Horn, J., and A. Dhamija, "IGP Flexible Algorithms Reverse Affinity Constraint", Work in Progress, Internet-Draft, draft-ppsenak-lsr-igp-flex-algo-reverse-affinity-01, 4 January 2023, <https://datatracker.ietf.org/doc/html/draft-ppsenak-lsr-igp-flex-algo-reverse-affinity-01>.

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

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC9350] Psenak, P., Ed., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", RFC 9350, DOI 10.17487/RFC9350, February 2023, <https://www.rfc-editor.org/info/rfc9350>.

[RFC9351] Talaulikar, K., Ed., Psenak, P., Zandi, S., and G. Dawra, "Border Gateway Protocol - Link State (BGP-LS) Extensions for Flexible Algorithm Advertisement", RFC 9351, DOI 10.17487/RFC9351, February 2023, <https://www.rfc-editor.org/info/rfc9351>.

## 10.2. Informative References

[RFC7308] Osborne, E., "Extended Administrative Groups in MPLS Traffic Engineering (MPLS-TE)", RFC 7308, DOI 10.17487/RFC7308, July 2014, <https://www.rfc-editor.org/info/rfc7308>.

Authors' Addresses

Ketan Talaulikar
Cisco Systems
India
Email: ketant.ietf@gmail.com

Aravind Babu MahendraBabu
Cisco Systems
India
Email: aravindbabu.mahendrababu@gmail.com

IDR
Internet-Draft
Intended status: Standards Track
Expires: 22 April 2023

R. Chen
D. Zhao
ZTE Corporation
L. Gong
China mobile

19 October 2022

# SR Policies Extensions for NRP in BGP-LS
draft-chen-idr-bgp-ls-sr-policy-nrp-01

## Abstract

This document defines a new TLV which enable the headed to report the configuration and the states of SR policies carrying NRP information by using BPG-LS.

## 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 April 2023.

## Copyright Notice

Table of Contents



## 1. Introduction

SR Policy is an ordered list of segments (i.e. instructions) that represent a source-routed policy. Packet flows are steered into a SR Policy on a node where it is instantiated called a headend node. The packets steered into an SR Policy carry an ordered list of segments associated with that SR Policy.[I-D.ietf-idr-te-lsp-distribution] describes a mechanism to distribute traffic engineering policy information (SR Policies , TE-LSPs, etc) to external components using BGP-LS.

[I-D.ietf-teas-ns-ip-mpls] introduces a Slice-Flow Aggregate as the collection of packets (from one or more IETF network slice traffic streams) that match an NRP Policy selection criteria and are offered the same forwarding treatment. The NRP Policy is used to realize an NRP by instantiating specific control and data plane resources on select topological elements in an IP/MPLS network. The NRP Identifier (NRP-ID) is globally unique within an NRP domain and that can be used in the control or management plane to identify the resources associated with the NRP.

Based on the mechanism defined in [I-D.ietf-idr-te-lsp-distribution], this document defines a new TLV which enable the headed to report the configuration and the states of SR policies carrying NRP information by using BPG-LS.

## 2. Requirements Language

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

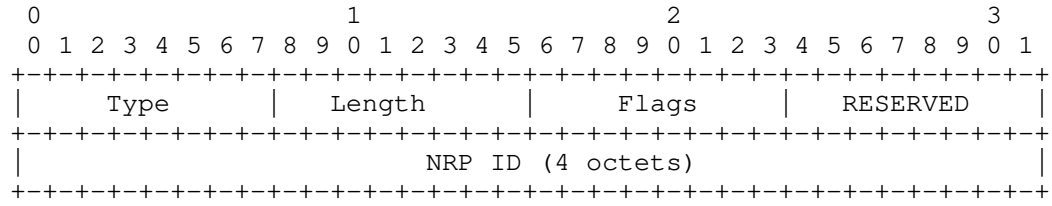
cloud transport network: It is usually a national or province backbone network to achieve interconnection between multiple regional clouds/core clouds deployed in the country/province.

## 3. Carrying NRP Sub-TLV in BGP-LS

[I-D.liu-idr-bgp-network-slicing] and [I-D.dong-idr-sr-policy-nrp] define extensions to BGP in order to advertise NRP in SR policies.

In order to collect configuration and states of the NRP SR policies, this document defines a new SR Policy state TLV.

The TLV has the following format:




where:

Type: TBD1.

Length: 6 octets.

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.

NRP: 4 octet global identifier of Network Resource Partition.

## 4. Acknowledgements

TBD.

## 5. IANA Considerations

TBD.

## 6. Security Considerations

TBD.

## 7. Normative References

[I-D.dong-idr-sr-policy-nrp]
Dong, J., Hu, Z., and R. Pang, "BGP SR Policy Extensions for Network Resource Partition", Work in Progress, Internet-Draft, draft-dong-idr-sr-policy-nrp-01, 11 July 2022, <https://www.ietf.org/archive/id/draft-dong-idr-sr-policy-nrp-01.txt>.

[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-18, 22 August 2022, <https://www.ietf.org/archive/id/draft-ietf-idr-te-lsp-distribution-18.txt>.

[I-D.ietf-teas-ns-ip-mpls]
Saad, T., Beeram, V. P., Dong, J., Wen, B., Ceccarelli, D., Halpern, J., Peng, S., Chen, R., Liu, X., Luis Contreras, M., Rokui, R., and L. Jalil, "Realizing Network Slices in IP/MPLS Networks", Work in Progress, Internet-Draft, draft-ietf-teas-ns-ip-mpls-00, 16 June 2022, <https://www.ietf.org/archive/id/draft-ietf-teas-ns-ip-mpls-00.txt>.

[I-D.liu-idr-bgp-network-slicing]
Yao, L. and P. Shaofu, "BGP Extensions to Support Packet Network Slicing in SR Policy", Work in Progress, Internet-Draft, draft-liu-idr-bgp-network-slicing-01, 15 October 2020, <https://www.ietf.org/archive/id/draft-liu-idr-bgp-network-slicing-01.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>.

## Authors' Addresses

Ran Chen
ZTE Corporation
Nanjing
China
Email: chen.ran@zte.com.cn

Detao Zhao
ZTE Corporation
Nanjing
China
Email: zhao.detao@zte.com.cn

Liyan Gong
China mobile
Beijing
China
Email: gongliyan@chinamobile.com

IDR Working Group
Internet-Draft
Intended status: Informational
Expires: 8 September 2023

E. Chen
Palo Alto Networks
R. Raszuk
Arrcus

7 March 2023

# Applying TCP User Timeout Parameter to BGP Sessions
## draft-chen-idr-tcp-user-timeout-01

## Abstract

In this document we discuss the TCP "User Timeout" parameter and recommend using it to handle stuck BGP sessions.

## 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 8 September 2023.

## Copyright Notice

Table of Contents



## 1. Introduction

A BGP session [RFC4271] is said, informally, to be "stuck" when BGP messages are not transmitted over the session for an extended period of time. Certainly the stuck BGP session should have been terminated by the BGP holdtimer. Such a case could occur, though, due to software defects or under certain unusual circumstances. Currently it's difficult to know for sure due to lacking of automated, real-time detection mechanisms in BGP implementations.

It has been speculated that some BGP sessions may have stuck from time to time, and that may have contributed to stale routes (e.g., missing route withdraws) in the routing system.

Here is a specific scenario of a stuck BGP session between two BGP speakers A and B:

* Due to certain software defect, B stops reading data from TCP [RFC0793] for an extended period of time, resulting in B advertises a zero value for its TCP window size after its TCP buffer fills up.

* B fails to generate BGP HOLDTIME expiration, although it has not read from TCP and thus has not received any BGP KEEPALIVE from A during that time.

* B, however, continues to send BGP KEEPALIVE to A on time.

In this scenario A would not be able to send routing updates to B during that period of time. The routing system may become stale, not only on B, but on its BGP neighbors and beyond.

It's desirable for a BGP speaker (e.g., A in the example) to be able to detect and then terminate such a stuck session so that the stale routes are purged from the routing system.

The availability of such a mechanism may also help accelerate the resolution of the software defect involved.

In this document we discuss the TCP "User Timeout" parameter [RFC0793] and recommend using it to handle stuck BGP sessions.

## 2. Discussion on TCP User Timeout

The TCP "User Timeout" parameter is designed to terminate a connection in a variety of cases where a TCP session does not progress within certain time period. It is specified in [RFC0793] as follows:

> USER TIMEOUT
>
> For any state if the user timeout expires, flush all queues, signal the user "error: connection aborted due to user timeout" in general and for any outstanding calls, delete the TCB, enter the CLOSED state and return.

Clearly the TCP "User Timeout" applies when the application data is not delivered on time, including the cases that transmitted data may remain unacknowledged, or buffered data may remain untransmitted (due to zero window size).

The TCP "User Timeout" parameter is well summarized in [RFC5482], although the zero-window case is not explicitly called out:

> The Transmission Control Protocol (TCP) specification RFC0793 defines a local, per-connection "user timeout" parameter that specifies the maximum amount of time that transmitted data may remain unacknowledged before TCP will forcefully close the corresponding connection. Applications can set and change this parameter with OPEN and SEND calls.

Regarding the implementation of the TCP "User Timeout" parameter, one example is Linux's "TCP_USER_TIMEOUT" socket option documented in [LINUX-TCP].

## 3. Recommendations

As discussed in the introduction, a BGP session is considered "stuck" when BGP messages are not delivered for an extended period of time.

Given that BGP messages are TCP data, and TCP is responsible for delivering the data, thus it would be more natural and more complete to address the issue at the TCP layer rather than in BGP itself (particularly in the case of persistent TCP zero-window).

As the TCP "User Timeout" parameter is specifically defined to terminate the TCP connection when something in TCP is "stuck", we thus recommend using it to detect and terminate these stuck BGP sessions.

We RECOMMEND that the TCP "User Timeout" parameter be set for all BGP sessions, and the default timeout value be five times the configured BGP holdtime value but no less than ten minutes in order to tolerate certain short-lived, transient conditions. The TCP "User Timeout" value for a BGP session SHOULD be configurable.

We also RECOMMEND that the TCP "User Timeout" parameter be set only after the End-of-RIB marker [RFC4724], if expected, is received from each of the (AFI, SAFI) being exchanged over the BGP session, or otherwise thirty minutes after the BGP session is established. The delay for setting the parameter SHOULD be configurable.

When the TCP "User Timeout" for a BGP session expires, the BGP speaker SHOULD log the event locally. In addition, the administrator of the remote BGP speaker SHOULD be informed (by means outside the scope of this document) so that the issue can be investigated.

The procedures for BGP Graceful Restart [RFC4724] SHOULD be followed when the TCP session is terminated due to TCP "User Timeout" expiration.

## 4. IANA Considerations

This document has no request for IANA.

## 5. Security Considerations

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

## 6. Acknowledgments

TBD

## 7. References

### 7.1. Normative References

[RFC0793] Postel, J., "Transmission Control Protocol", RFC 793, DOI 10.17487/RFC0793, September 1981, <https://www.rfc-editor.org/info/rfc793>.

[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>.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <https://www.rfc-editor.org/info/rfc4724>.

[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>.

### 7.2. Informative References

[LINUX-TCP] TCP(7), "Linux Man Pages", March 2021, <https://man7.org/linux/man-pages/man7/tcp.7.html>.

[RFC5482] Eggert, L. and F. Gont, "TCP User Timeout Option", RFC 5482, DOI 10.17487/RFC5482, March 2009, <https://www.rfc-editor.org/info/rfc5482>.

## Authors' Addresses

Enke Chen
Palo Alto Networks
Email: enchen@paloaltonetworks.com

Robert Raszuk
Arrcus
Email: robert@raszuk.net

IDR Working Group
Internet-Draft
Intended status: Informational
Expires: 12 September 2023

X. Ding
Z. Tan
L. Wang
Huawei Technologies

11 March 2023

# Route Target Constraint for BGP Flow Spec(BGP Flow) and BGP Segment Routing Policies(BGP SR-Policy)
## draft-ding-idr-rtc-for-bgp-flow-sr-00

## Abstract

This document introduces an extension to the application scenarios of Route Target Constraints (RTC). By using the global administrator field of the IPv4 Address Specific Extended Community to represent a network node and exchanging BGP Route-Target routes, a BGP speaker could generate an egress policy for filtering one or a group of network nodes, which could implement precise control and distribution of services such as BGP Flow Spec and BGP Segment Routing Policies.

## 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 12 September 2023.

Table of Contents



## 1. Introduction

BGP [RFC4271] has been used to distribute different types of routing and policy information. In some scenarios, the distributed routing information is specific for certain services, such as BGP/MPLS IP VPNs.

Route Target Constraints (RTC) [RFC4684], extends Outbound Route Filtering (ORF), describes how route targets are exchanged through the BGP RTC address family on a BGP/MPLS IP VPN network to generate egress policies. This feature enables the BGP/MPLS IP VPN network to control the advertisement of VPN routing information in a more refined manner.

This document introduces an extension to the application scenarios of Route Target Constraints (RTC) [RFC4684] to control the distribution of routing information to one or a group of network nodes, which could implement precise control of services such as BGP Flow Spec [RFC8955] and BGP Segment Routing Policies [I-D.ietf-idr-segment-routing-te-policy].

### 1.1. Terminology

This document introduces the following terms:

RTC Route Target Constraints [RFC 4684]

ORF Outbound Route Filtering

Flowspec BGP Flow Specification

SR-Policy BGP Segment Routing Policy

NLRI Network Layer Reachability Information

## 2. Route Target Membership NLRI Advertisements

The encapsulation of Route Target membership NLRI is defined in Route Target Constraints (RTC) [RFC4684], the NLRI is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes. The (AFI, SAFI) value pair used to identify this NLRI is (AFI=1, SAFI=132).

The route-target field in the NLRI indicates a network node and is encoded as a IPv4 Address Specific Extended Community [RFC4360], as shown blow:

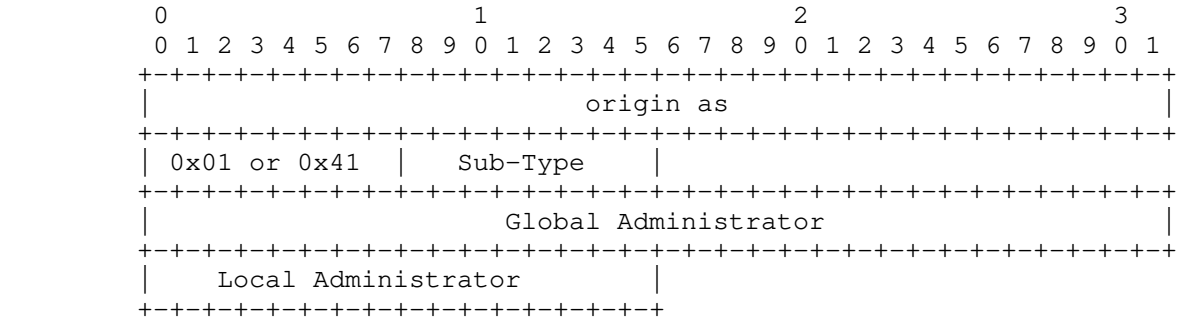



Figure 1: Route Target membership NLRI Format

While encoding these fields:

* Global Administrator: 4 octets, indicates the router identifier of the node. If the Global Administrator is set to 0.0.0.0, it means that the peer node accepts all policy rules from the RR.

* Local Administrator: 2 octets, reserved for future use, MUST be set to 0 upon the sender and MUST be ignored upon the receiver.

## 3. Use case

This section describes a few use-case scenarios.

### 3.1. BGP Flow Spec ORF

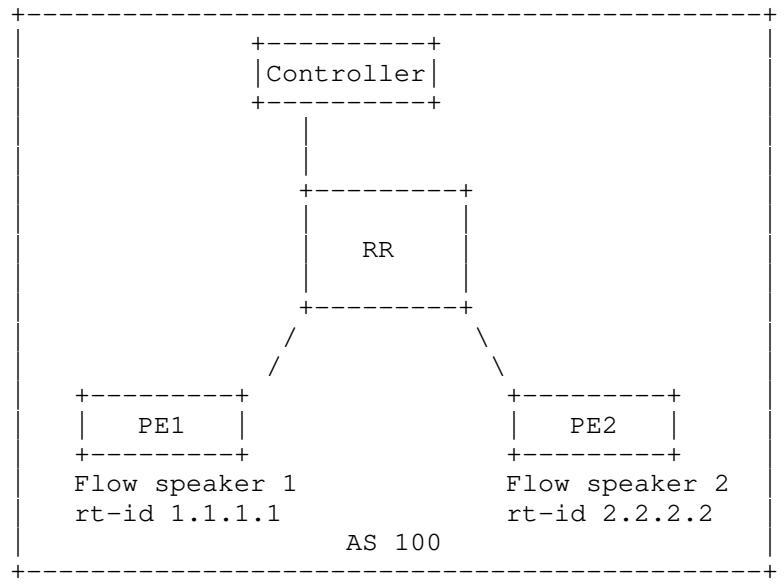



Figure 2: BGP Flow Spec ORF

In the topology above, the Controller, PE1, and PE2 establish IBGP peer relationships with the RR respectively. PE1 and PE2 are clients of the RR. The Controller distributes Flowspec rules through the RR, and the RR reflects the Flowspec rules to PE1 and PE2.

PE1 sends route target membership NLRI{100, 1.1.1.1:0} to the RR, and PE2 sends route target membership NLRI{100, 2.2.2.2:0} to the RR. After receiveing the UPDATE messages with Route Target Membership NLRI, the RR will trigger the RIB-OUTS of the Flowspec route to match the egress policies and update the route to PEs.

If hierarchical RRs are deployed, the RRs need to advertise all received route target membership NLRI routes to the upper-layer RRs.

### 3.2. BGP Segment Routing Policies ORF

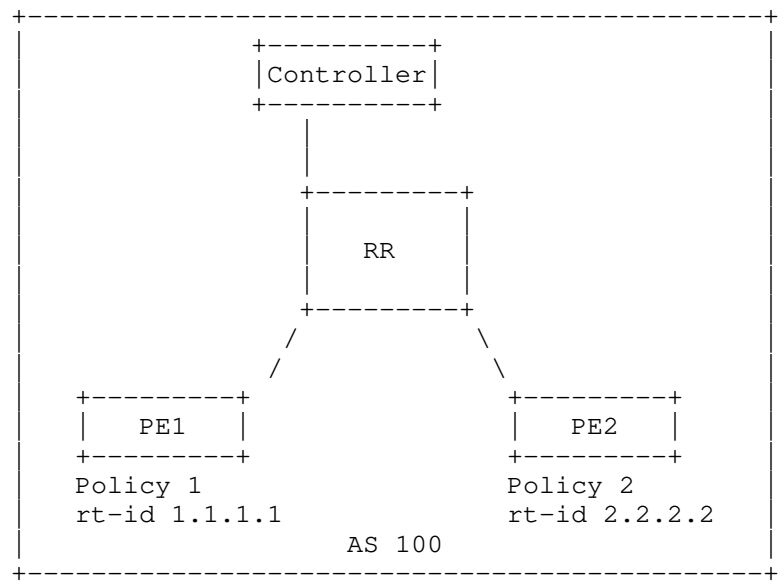



Figure 3: BGP Segment Routing Policies ORF

It is described in BGP Segment Routing Policies [I-D.ietf-idr-segment-routing-te-policy] that one or more route targets SHOULD be attached to the advertisement, where each route target identifies one or more intended headends for the advertised SR Policy update. In the topology above, when the controller needs to deliver SR policies to PE1 and PE2, it will advertises SR policies with route target extended communities, SR Policy1 with {1.1.1.1:0} and SR Policy2 with {2.2.2.2:0}, to RR. The RR will reflect SR Policies to both PE1 and PE2. PEs need to do an ingress filtering, by matching route target extended community with its own router-id. In this case, PE1 will keep SR Policy1 and drop SR Policy2, as well as PE2 will keep SR Policy2 and drop SR Policy1. During this process, even though SR policies are correctly provisioned, the RR advertises all routes to all peers, which may cause network congestion.

The ORF operations described in this document work as an egress filter on RR. PE1 sends route target membership NLRI{100, 1.1.1.1:0} to the RR, and PE2 sends route target membership NLRI{100, 2.2.2.2:0} to the RR. After receiving the Route Target Membership NLRI from the PE, the RR generates a PE-specific egress filter. Before advertising routes to PEs, the RR matches routes with egress policies, and will only deliver SR policy1 to PE1 and SR policy1 to PE2 respectively. In this way, services could be correctly deployed and network bandwidth could be saved.

## 4. IANA Considerations

TBD

## 5. Security Considerations

TBD

## 6. References

### 6.1. 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>.

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

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

### 6.2. 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-20, 27 July 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-segment-routing-te-policy-20>.

[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>.

[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>.

Authors' Addresses

Xiangfeng Ding
Huawei Technologies
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: dingxiangfeng@huawei.com

Zhen Tan
Huawei Technologies
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: tanzhen6@huawei.com

Lili Wang
Huawei Technologies
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: lily.wong@huawei.com

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: 28 August 2023

C. Li, Ed.
H. Shi, Ed.
Huawei Technologies
T. He
R. Pang
China Unicom
G. Qian
Huawei Technologies

24 February 2023

# Distribution of Service Metadata in BGP-LS
draft-ls-idr-bgp-ls-service-metadata-01

## Abstract

In edge computing, a service may be deployed on multiple instances within one or more sites, called edge service. The edge service is associated with an ANYCAST address in IP layer, and the route of it with potential service metatdata will be distributed to the network. The Edge Service Metadata can be used by ingress routers to make path selections not only based on the routing cost but also the running environment of the edge services.

The service route with metadata can be collected by a PCE(Path Compute Element) or an analyzer for calculating the best path to the best site/instance. This draft describes a mechanism to collect the information of the service routes and related service metadata in BGP-LS.

## About This Document

This note is to be removed before publishing as an RFC.

The latest revision of this draft can be found at https://VMatrix1900.github.io/draft-service-metadata-in-BGP-LS/draft-ls-idr-bgp-ls-service-metadata.html. Status information for this document may be found at https://datatracker.ietf.org/doc/draft-ls-idr-bgp-ls-service-metadata/.

Discussion of this document takes place on the Inter-Domain Routing Working Group mailing list (mailto:idr@ietf.org), which is archived at https://mailarchive.ietf.org/arch/browse/idr/. Subscribe at https://www.ietf.org/mailman/listinfo/idr/.

Source for this draft and an issue tracker can be found at https://github.com/VMatrix1900/draft-service-metadata-in-BGP-LS.

Status of This Memo



Copyright Notice



Table of Contents

## 1. Introduction

Many services deploy their service instances in multiple sites to get better response time and resource utilization. These sites are often geographically distributed to serve the user demand. For some services such as VR/AR and intelligent transportation, the QoE will depend on both the network metrics and the compute metrics. For example, if the nearest site is overloaded due to the demand fluctuation, then steer the user traffic to a another light-loaded sites may improve the QoE.

[I-D.ietf-idr-5g-edge-service-metadata] descirbes the BGP extension of distributing service route with network and computing-related metrics. The router connected to the site will received the service routes and service metadata sent from devices inside the egdge site, and then generates the corresponding routes and distributes them to ingress routers. However, the route with service metadata on the router connected to the site can be also collected by a central Controller for calculating the best path to the best site.

This document defines an extension of BGP-LS to carry the service metadata along with the service route. Using the service metadata and the service route, the controller can calculate the best site for the traffic, giving each user the best QoE.

### 1.1. Terminology

### 1.2. Requirements Language

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

## 2. BGP-LS Extension for Service in a Site

The goal of the BGP-LS extension is to collect the information of the service prefix and metadata of the service, such as network metrics and compute metrics. A service is identified by an prefix, and this information is carried by existing prefix NLRI TLV. Other information including service metadata are carried by attributes TLVs.

### 2.1. Prefix NLRI

A service is identified by a prefix, and the Prefix NLRI defined in the [RFC7752] is used to collect the prefix information of the service. The format of the Prefix NLRI is shown in Figure 1 for better understanding.

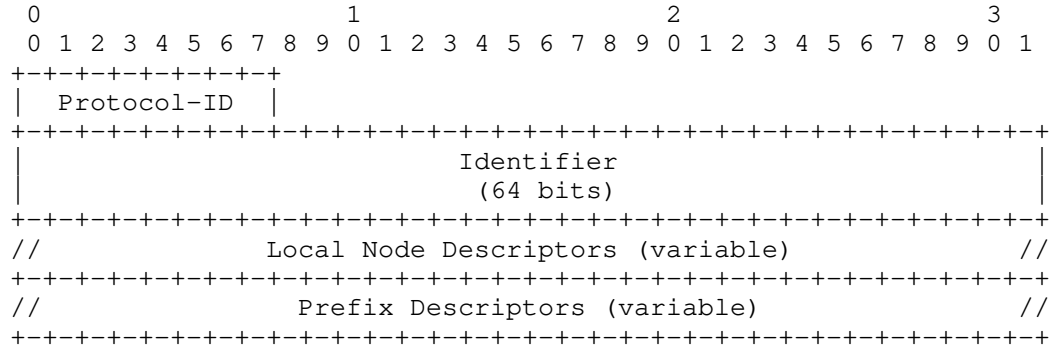



Figure 1: The IPv4/IPv6 Topology Prefix NLRI Format

Specifically, the service prefix is carried by the IP Reachability Information TLV(Figure 2) inside the Prefix Descriptor field. The Prefix Length field contains the length of the prefix in bits. The IP Prefix field contains the most significant octets of the prefix.

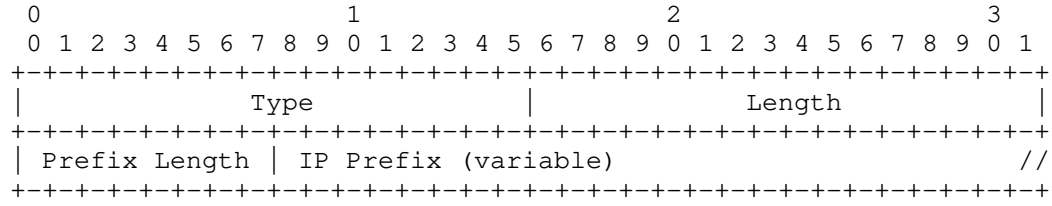



Figure 2: IP Reachability Information TLV Format

### 2.2. Attributes

The following three prefix attribute TLVs are used to carry the metadata of a service instance:

1. Metadata Path Attribute TLV carries the compute metric of the service instance such as site preference, capacity index and load measurement defined in [I-D.ietf-idr-5g-edge-service-metadata].

2. Prefix SID TLV carries a Prefix SID associated to the edge site.

3. Color Attribute TLV carries the service requirement level information of the service

#### 2.2.1. Metadata Path Attribute TLV

The Metadata Path Attribute TLV is an optional attribute to carry the Edge Service Metadata defined in the [I-D.ietf-idr-5g-edge-service-metadata]. It contains multiple sub-TLVs, with each sub-TLV containing a specific metric of the Edge Service Metadata. This document define a new TLV in BGP-LS, which reuse the name and the format of Metadata Path Attribute TLV.

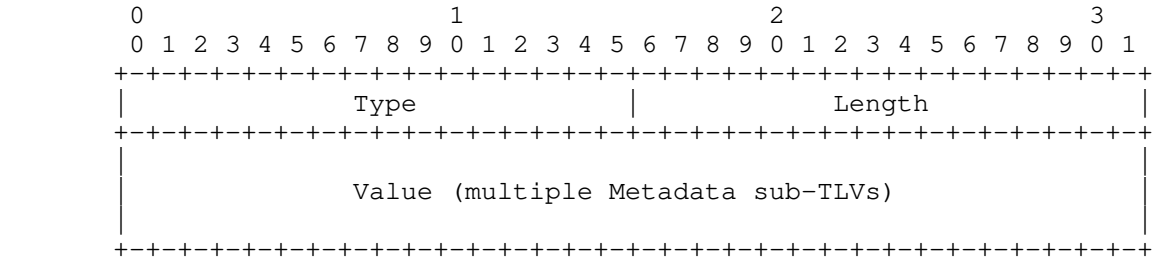



Figure 3: Metadata Path Attribute TLV format

* Type: identify the Metadata Path Attribute, to be assigned by IANA.

* Length: the total number of the octets of the value field.

* Value: contains multiple sub-TLVs.

There are three types of Edge Service Metadata sub-TLVs defined in [I-D.ietf-idr-5g-edge-service-metadata]:

1. Site Preference Index indicates the preference to choose the site.

2. Capacity Index indicates the capability of a site. One Edge Site can be in full capacity, reduced capacity, or completely out of service.

3. Load Measurement indicates the load level of the site.

To collect these information, this document defines TLVs reusing the name and format of the TLVs defined in [I-D.ietf-idr-5g-edge-service-metadata].

### 2.3. Prefix SID Attribute TLV

In some cases, there may be multiple sites connect to one Edge(egress) router through different interfaces. Generally, a overlay path, such a overlay tunnel will be used between the ingress router and the egress for steering the traffic to the best site correctly. In SR-MPLS networks or SRv6 networks, a prefix SID is needed. For example, some SRv6 Endpoint Behaviors such as End.DX6, End.X can be encoded for each site so that the egress router can steer the traffic to the corresponding site. The Prefix SID TLV defined [RFC9085] can be used to collect this information.

The Prefix SID TLV is an optional TLV to carry the Prefix SID associated to the edge site. The TLV format is illustrated in Figure 4.

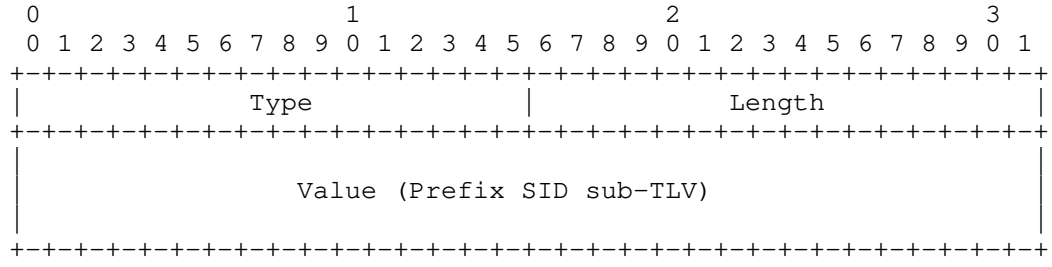



Figure 4: Prefix-SID TLV format

* Type: 1158, identify the Prefix SID Attribute.

* Length: the total number of the octets of the value field.

* Value: contains Prefix SID sub-TLV.

#### 2.3.1. Color Attribute TLV

Color is used to indicate the service level. For example, different site may have different level of service capability which is taken into account of by the controller when calculate the path to the egress router. More details can be added in the future revision.

The TLV format(shown in Figure 5) is similar to the BGP Color Extended Community defined in [RFC9012].

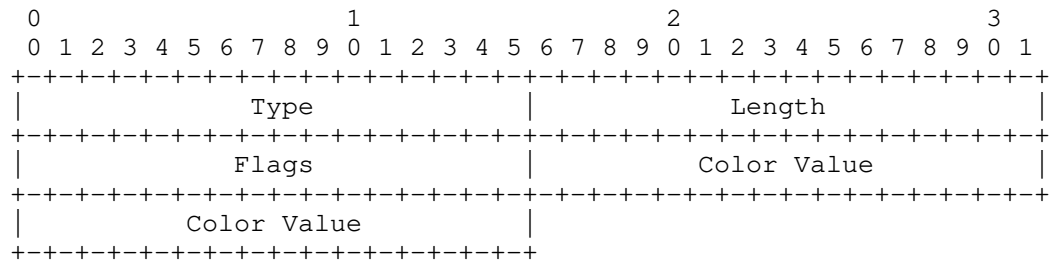



Figure 5: Color Attribute TLV format

* Type: identify the Color Attribute, to be assigned by IANA.

* Length: 6, length of Flags + Color Value.

* Flags and Color is the same as defined in [RFC9012]. Color Value: 32 bit value of color.

## 3. Security Considerations

TBD

## 4. IANA Considerations

This document requires IANA to assign the following code points from the registry called "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs":

| Value | Description | Reference |
|---|---|---|
| TBD1 | Metadata Path Attribute Type | Section 2.2.1 |
| TBD2 | Site Preference Sub-Type | Section 2.2.1 |
| TBD3 | Capacity Sub-Type | Section 2.2.1 |
| TBD4 | Load Measurement Sub-Type1: Aggregated-Cost | Section 2.2.1 |
| TBD5 | Load Measurement Sub-Type2: Raw-Measurements | Section 2.2.1 |
| TBD6 | Color Attribute Type | Section 2.3.1 |

Table 1

## 5. Contributors

Xiangfeng Ding

email: dingxiangfeng@huawei.com

## 6. Normative References

[I-D.ietf-idr-5g-edge-service-metadata]
Dunbar, L., Majumdar, K., Wang, H., and G. S. Mishra, "BGP Extension for 5G Edge Service Metadata", Work in Progress, Internet-Draft, draft-ietf-idr-5g-edge-service-metadata-00, 2 December 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-5g-edge-service-metadata-00>.

[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/rfc/rfc2119>.

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

[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/rfc/rfc9012>.

[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/rfc/rfc9085>.

[RFC9252] Dawra, G., Ed., Talaulikar, K., Ed., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "BGP Overlay Services Based on Segment Routing over IPv6 (SRv6)", RFC 9252, DOI 10.17487/RFC9252, July 2022, <https://www.rfc-editor.org/rfc/rfc9252>.

Acknowledgements

The authors would like to thank Haibo Wang, LiLi Wang, Jianwei Mao for their help.

Authors' Addresses

Cheng Li (editor)
Huawei Technologies
Beijing
China
Email: c.l@huawei.com

Hang Shi (editor)
Huawei Technologies
Beijing
China
Email: shihang9@huawei.com

Tao He
China Unicom
Beijing
China
Email: het21@chinaunicom.cn

Ran Pang
China Unicom
Beijing
China
Email: pangran@chinaunicom.cn

Guofeng Qian
Huawei Technologies
Beijing
China
Email: qianguofeng@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 8 September 2023

S R. Mohanty
J. Alcaide
M. Ghosh
Cisco Systems, Inc.

7 March 2023

# A solution to the Hierarchical Route Reflector issue in RT Constraints
draft-mohanty-idr-rtc-hierarchical-rr-00

## Abstract

Route Target Constraints (RTC) is used to build a VPN route distribution graph such that routers only receive VPN routes corresponding to specified route-targets (RT) that they are interested in. This is done by exchanging the route-targets as routes in the RTC address-family and a corresponding "RT filter" is installed that influences the VPN route advertisement. In networks employing hierarchical Route Reflectors (RR) the use of RTC can lead to incorrect VPN route distribution and loss in connectivity as detailed in an earlier draft . Two solutions were provided to overcome the problem.

This draft presents a method with suggested modifications to the RTC RFC in order to solve the hierarchical RR RTC problem in an efficient manner.

## 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 September 2023.

## Copyright Notice

Table of Contents



## 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. Introduction

Hierarchical RR [RFC4456] deployments with VPN [RFC4364] working in conjunction with RTC [RFC4684] may result in sub-optimal and incorrect VPN route distribution that is nicely described in [I-D.ietf-idr-rtc-hierarchical-rr]. The root reason for this is the way the RR rules for RTC are defined in [RFC4684]. The authors of [I-D.ietf-idr-rtc-hierarchical-rr] furnish two solutions for the problem, one based on add-paths and the other based on diverse-paths constructs. In this memo, we present another another solution to the very same problem.

## 3. RTC and RR Rules

When advertising RT membership NLRI to a route-reflector client, Section 3.2 of [RFC4684] advocates the advertising RR to set the ORIGINATOR_ID attribute [RFC4456] to its own router-id, and the Next-hop attribute to be set to the local address for that session. However, this creates the issue in hierarchical RR setups as explained in [I-D.ietf-idr-rtc-hierarchical-rr]. Fig. 1 represents the same Figure as in [I-D.ietf-idr-rtc-hierarchical-rr]. When RR-2 and RR-3 advertise RT-1 to RR-1, the latter will choose one of the routes to be best and will advertise the same to RR-2 and RR-3 respectively after setting the ORIGINATOR_ID and next-hop to itself. Note that RR-1 will also add its own CLUSTER_ID [RFC4456]to the CLUSTER_LIST but importantly not overwrite the CLUSTER_ID of the sender. This leads to the issue explained in [I-D.ietf-idr-rtc-hierarchical-rr].

## 4. Problem Definition

In the Fig 1, when RR-1 chooses the route from RR-2 as the best route, and formats the next-hop and ORIGINATOR_ID as explained above and then advertises the route to RR-2, RR-2 will drop the route reflected from RR-1 because of the CLUSTER_ID check.

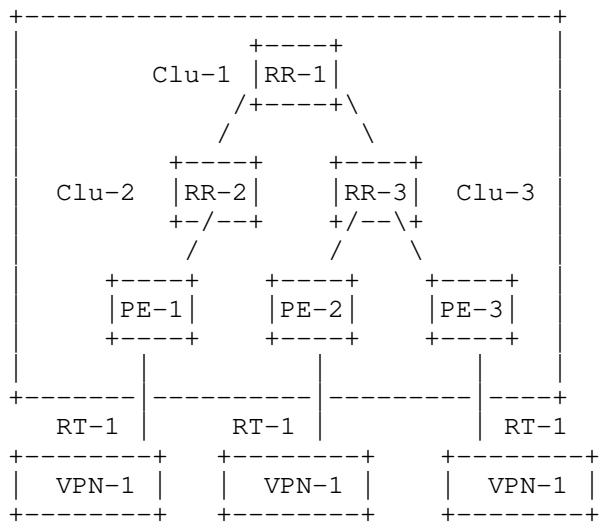



Figure 1 Hierarchical RR Setup with RTC

Figure 1

RR-2 will therefore not form the outbound filter of RT-1 towards RR-1 which means that after convergence RR-2 will not advertise VPN routes to RR-1 anymore. This leads to an incorrect VPN route distribution across the network.

In the scenario of Fig 2. CE-1 is multi-homed to PE-1 and PE-2 and wants to communicate with CE-2 which is behind PE-4. As explained earlier, because RR-1 chooses RR-2 path as best in the RTC family, RR-1 is only receiving the VPN route from RR-3 (and not RR-2) in the steady state.

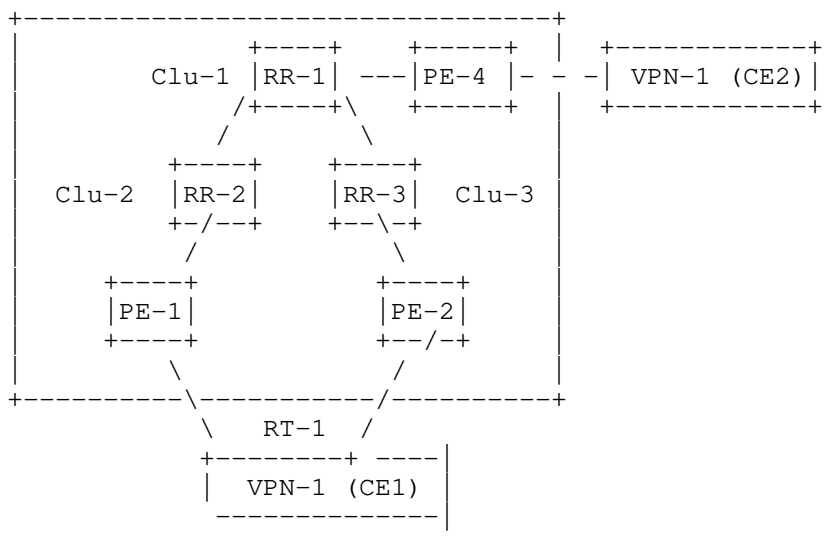



Figure 2 Hierarchical RR Setup with RTC with dual-homed CE

Figure 2

Notice that even though the link between between RR-3 and RR-1 comes down, The RR-2 PATH still remains as best in the RTC address-family at RR-1 and the VPN route advertisements to RR-1 from RR-2 still continue to be blocked. Thus even though there is an alternative connectivity from CE-1 to PE-4 via PE-1, RR-2 and RR-1, the BGP VPN routes cannot be sent. In fact CE-1 is completely cut-off from rest of the network. Generalizing, it means that in a hierarchical RR with only a single first-level RR as its client, the solution is completely broken. Notice that without RTC, RR-1 would have both VPN paths and the loss of connectivity to RR-3 would just result in local convergence at RR-1 subject to the time when the path from RR-2 becomes best.

The solutions presented in [I-D.ietf-idr-rtc-hierarchical-rr] are based on

a. Addpath, RR-1 will advertise both the paths from RR-2 and RR-3 to RR-2 and RR-3 so that each of the first level RRS will accept at least one of the routes and install the filter

b. When RR-1 will advertise the best-path to a client or non-client speaker, and that speaker is the one whose path is the best, the advertising router will use the most "diverse" path (different next-hop and ORIGINATOR_ID than the best-path) to accomplish the same goal, i.e. the path will be accepted at the receiving speaker

In the next section, we provide a solution, that does not require add-path and also improves upon [RFC4684] while solving this hierarchical RR issue in RTC.

## 5. Proposed Solution

A problem that [RFC4684] does not address is limiting the number for VPN routes advertised to AN RR when only one client advertises RTC routes. Consider in Figure 1 that PE-1 is the only one advertising a RTC route to RR-2. RR-2 will advertise back the route towards PE-1 (with next-hop/ORIGINATOR_ID rewriting to avoid PE-1 discarding the route). PE-1 will advertise VPN routes towards RR-1, which is unnecessary and wastes resources on RR-1.

### 5.1. Sender Advertisement Rule

In the description that follows we define Attribute diversity to mean RTC routes with different ORIGINATOR_ID attribute (or router-id of the peer it is received from if ORIGINATOR_ID does not exist), or different first CLUSTER_ID inside the CLUSTER_LIST (an empty CLUSTER_ID is different to any other non-empty CLUSTER_ID). Diversity for ORIGINATOR_ID is typically used for first level RRs, diversity for CLUSTER_LIST is typically used for higher level RRs.

Both diversity attributes can be used in combination when the RR has a mix of clients (that are themselves RRs and non-RRs). The underlying assumptions for looking at CLUSTER_ID for attribute diversity is that any clients that are also route-reflectors and have the same CLUSTER_ID, will themselves have the same set of clients. Imagine a a higher level RR receiving the same route from two lower level RRs with the same CLUSTER_ID. It will not reflect back the RTC if only the same CLUSTER_ID is received from its clients (as first CLUSTER_ID in the CLUSTER_LIST).

The following rule modifies the [RFC4684].

* A RTC route is reflected from a client to a client (including the client the route is received from), only when there is enough attribute diversity amongst the RTC routes received from all the clients.

The rule above will apply as well to a top level RR, guaranteeing that top level RR does not have unnecessary routes.

### 5.2. Receiver Acceptance Rule

We need to take care that when reflecting back a RTC route to the advertising client, this client does not discard the update. RFC 4684 mandates to overwrite next-hop to self and the ORIGINATOR_ID to our local router-id when advertising RTC route a route reflector client (section 3.2, rule (i) ). This rule can be extended to take care of the case where the reflection happens at a higher level RR (See Rule(1) below). Additionally, no attribute overwriting is deemed necessary when reflecting a RTC route from client to non-client. Thus, the following rule is added to RFC 4684.

1. When reflecting a RTC route from RR client to RR client, NEXT_HOP attribute is overwritten to self, ORIGINATOR _ID is set or overwritten to local router-id, and first CLUSTER_ID of CLUSTER_LIST (if not empty) is overwritten to local CLUSTER_ID (this is before regular CLUSTER_ID prepending; thus advertised CLUSTER_LIST may have two repeated CLUSTER_ID at the beginning).

2. when a RR receives an RT-Constraint route that contains its own CLUSTER_ID or ORIGINATOR_ID, it ignores the CLUSTER_ID/ ORIGINATOR_ID check and does not discard the path but keep it as "Received-only". Treating the path as "Received-only" removes it from best-path computation considerations but allows to install the VPN filter.

With the Rule 1. since we are over-writing the cluster-id, the receiver will accept the route. With Rule 2., in the above Fig. 1 and 2 when RR-2 receives the update from RR-1, it will accept the route and will treat it as "Received-only". However, although the route will not be eligible for advertisement, but since this route is accepted, the VPN filter is installed and VPN routes will be advertised from RR-2 to RR-1. Both rules are exclusive of each other.

## 6. Conclusion

With the procedures it is not necessary for the RR to know in which level it is operating. The above rules are compatible. We always advertise best-path for any rule and it is easily seen that RR-2 will accept the RT Constraint path advertised from RR-1 . Since the path is accepted, the RT Filter at RR-2 will pass the VPN routes, and the problem scenarios are resolved accordingly.

With this specification in the RT-Constraint address-family, we solve both the incorrect and sub-optimal issues as mentioned above. There is no need for add-paths. We can also optimize over [RFC4684] on RTC advertisements based on diversity of ORIGINATOR_ID and CLUSTER_ID so that a higher level RR does not have to be populated with VPN routes with a specific RT if that RT is not present in other clusters.

## 7. IANA Considerations

None.

## 8. Operational Considerations

TBD.

## 9. Security Considerations

This document raises no new security issues for RT Constraints.

## 10. Acknowledgements

The authors would like to thank Swadesh Agrawal and M. Mirza for useful discussions related to hierarchical RR RTC.

## 11. Normative References

[I-D.ietf-idr-rtc-hierarchical-rr]
Dong, J., Chen, M., and R. Raszuk, "Extensions to RT-Constrain in Hierarchical Route Reflection Scenarios", Work in Progress, Internet-Draft, draft-ietf-idr-rtc-hierarchical-rr-03, 3 July 2017, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-rtc-hierarchical-rr-03>.

[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>.

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

[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>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

## Authors' Addresses

Satya Ranjan Mohanty
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
United States of America
Email: satyamoh@cisco.com

Juan Alcaide
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
United States of America
Email: jalcaide@cisco.com

Mrinmoy Ghosh
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
United States of America
Email: mrghosh@cisco.com

IDR
Internet-Draft
Updates: 4271 (if approved)
Intended status: Standards Track
Expires: 15 August 2023

J. Snijders
Fastly
B. Cartwright-Cox
Port 179 Ltd

11 February 2023

# Border Gateway Protocol 4 (BGP-4) Send Hold Timer
# draft-spaghetti-idr-bgp-sendholdtimer-09

## Abstract

This document defines the SendHoldTimer session attribute for the Border Gateway Protocol (BGP) Finite State Machine (FSM). Implementation of a SendHoldTimer should help overcome situations where BGP sessions are not terminated after it has become detectable for the local system that the remote system is not processing BGP messages. For robustness, this document specifies that the local system should close BGP connections and not solely rely on the remote system for session closure when BGP timers have expired. This document updates RFC4271.

## Requirements Language

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

## Status of This Memo

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

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

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

This Internet-Draft will expire on 15 August 2023.

Table of Contents



## 1. Introduction

This document defines the SendHoldTimer session attribute for the Border Gateway Protocol (BGP) [RFC4271] Finite State Machine (FSM) defined in section 8.

Failure to terminate a 'stuck' BGP session can result in Denial Of Service, the subsequent failure to generate and deliver BGP WITHDRAW messages to other BGP peers of the local system is detrimental to all participants of the inter-domain routing system. This phenomena is theorised to have contributed to IP traffic backholing events in global Internet routing system [bgpzombies].

This specification intends to improve this situation by requiring sessions to be terminated if the local system has detected that the remote system cannot possibly have received any BGP messages for the duration of the SendHoldTimer. Through codification of the aforementioned requirement, operators will benefit from consistent behavior across different BGP implementations.

BGP speakers following this specification do not exclusively rely on remote systems robustly closing connections, but will also locally close connections.

## 2. Example of a problematic scenario

In implementations lacking the concept of a SendHoldTimer, a malfunctioning or overwhelmed remote peer may cause data on the BGP socket in the local system to accumulate ad infinitum. This could result in forwarding failure and traffic loss, as the overwhelmed peer continues to utilize stale routes.

An example fault state: as BGP runs over TCP [RFC9293] it is possible for hosts in the ESTABLISHED state to encounter a BGP peer that is advertising a TCP Receive Window (RCV.WND) of size zero, this 0 window prevents the local system from sending KEEPALIVE, CEASE, WITHDRAW, UPDATE, or any other critical BGP messages across the network socket to the remote peer. Historically, many BGP implementations were unable to handle this situation in a robust fashion. Previous BGP RFC specifications would not give cause for the session to be torn down in such situations.

Generally BGP implementation have no visibility into lower-layer subsystems such as TCP or the peer's current Receive Window. Therefore, this document relies upon BGP implementations having the ability to detect whether the TCP socket to a BGP peer is progressing (data is being transmitted), or persisting in a stalled state.

## 3. Specification of the Send Hold Timer

BGP speakers are implemented following a conceptual model "BGP Finite State Machine" (FSM), which is outlined in section 8 of [RFC4271]. This specification updates the BGP FSM as following:

### 3.1. Session Attributes

The following mandatory session attributes are added to paragraph 6 of Section 8, before "The state session attribute indicates the current state of the BGP FSM":

9) SendHoldTimer

10) SendHoldTime (an initial value of 8 minutes is recommended)

## 3.2. SendHoldTimer_Expires Event Definition

Section 8.1.3 [RFC4271] is extended as following:

> Event XX1: SendHoldTimer_Expires
> Definition : An event generated when the SendHoldTimer expires.
> Status: Mandatory

If the SendHoldTimer_Expires (Event XX1), the local system:

- logs a message with the BGP Error Notification Code "Send Hold Timer Expired",
- releases all BGP resources,
- sets the ConnectRetryTimer to zero,
- drops the TCP connection,
- increments the ConnectRetryCounter,
- (optionally) performs peer oscillation damping if the DampPeerOscillations attribute is set to TRUE, and
- changes its state to Idle.

If the DelayOpenTimer_Expires event (Event 12) occurs in the Connect state, the local system:

- sends an OPEN message to its peer,
- sets the HoldTimer to a large value, and
- sets the SendHoldTimer to a large value, and
- changes its state to OpenSent.

If the DelayOpen attribute is set to FALSE, the local system:

- stops the ConnectRetryTimer (if running) and sets the ConnectRetryTimer to zero,
- completes BGP initialization
- sends an OPEN message to its peer,

- sets the HoldTimer to a large value, and
- sets the SendHoldTimer to a large value, and
- changes its state to OpenSent.

A HoldTimer value of 4 minutes is suggested.

A SendHoldTimer value of 8 minutes is suggested.

### 3.3. MsgSent Event Definition

Section 8.1.5 [RFC4271] is extended as following:

Event XX2: MsgSent
Definition: An event is generated when a KEEPALIVE or UPDATE message is transmitted.
Status: Mandatory

### 3.4. Restarting the SendHoldTimer

On page 74 [RFC4271] before "If the local system receives an UPDATE message, and the UPDATE message error handling procedure (see Section 6.3) detects an error (Event 28), the local system:", add the following:

If the local system transmits a KEEPALIVE or UPDATE message (MsgSent (Event XX2)), the local system:

- restarts the SendHoldTimer, and
- remains in the Established state.

## 4. Send Hold Timer Expired Error Handling

If a system does not send successive KEEPALIVE, UPDATE, and/or NOTIFICATION messages within the period specified in the Send Hold Time, then the BGP connection is closed and a log message is emitted.

## 5. Operational Considerations

When the local system recognizes a remote peer is not processing any BGP messages for the duration of the Send Hold Timer, the local system will not be able to inform the remote peer through a BGP message as to why the session is being closed (i.e. a NOTIFICATION message with the "Send Hold Timer Expired" error code).

Even so, BGP speakers SHOULD provide this reason as part of their operational state; e.g. bgpPeerLastError in the BGP MIB [RFC4273].

## 6. Security Considerations

This specification addresses the vulnerability of a BGP speaker to a potential attack whereby a BGP peer can pretend to be unable to process BGP messages and in doing so create a scenario where the local system is poisoned with stale routing information.

There are three detrimental aspects to the problem of not robustly handling 'stuck' peers:

* Failure to send BGP messages to a peer implies the peer is operating based on stale routing information.

* Failure to disconnect from a 'stuck' peer hinders the local system's ability to construct a non-stale local Routing Information Base (RIB).

* Failure to disconnect from a 'stuck' peer hinders the local system's ability to inform other BGP peers with current network reachability information.

In other respects, this specification does not change BGP's security characteristics.

## 7. IANA Considerations

This document requests IANA to assign a value named "Send Hold Timer Expired" in the "BGP Error (Notification) Codes" sub-registry under the "Border Gateway Protocol (BGP) Parameters" registry.

## 8. Acknowledgements

The authors would like to thank William McCall, Theo de Raadt, John Heasley, Nick Hilliard, Jeffrey Haas, and Tom Petch for their helpful review of this document.

## 9. References

### 9.1. Normative References

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

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

[RFC4273] Haas, J., Ed. and S. Hares, Ed., "Definitions of Managed Objects for BGP-4", RFC 4273, DOI 10.17487/RFC4273, January 2006, <https://www.rfc-editor.org/info/rfc4273>.

[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>.

[RFC9293] Eddy, W., Ed., "Transmission Control Protocol (TCP)", RFC 9293, DOI 10.17487/RFC9293, August 2022, <https://www.rfc-editor.org/info/rfc9293>.

### 9.2. Informative References

[bgpzombies] Fontugne, R., "BGP Zombies", April 2019, <https://labs.ripe.net/author/romain_fontugne/bgp-zombies/>.

[frr] Lamparter, D., "bgpd: implement SendHoldTimer", May 2022, <https://github.com/FRRouting/frr/pull/11225>.

[neo-bgp] Cartwright-Cox, B., "What does bgp.tools support", August 2022, <https://bgp.tools/kb/bgp-support>.

[openbgpd] Jeker, C., "bgpd send side hold timer", December 2020, <https://marc.info/?l=openbsd-tech&m=160820754925261&w=2>.

## Appendix A. Implementation status - RFC EDITOR: REMOVE BEFORE PUBLICATION

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in RFC 7942. 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 RFC 7942, "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".

* OpenBGPD [openbgpd]

* FRRouting [frr]

* neo-bgp (bgp.tools) [neo-bgp]

## Authors' Addresses

Job Snijders
Fastly
Amsterdam
Netherlands
Email: job@fastly.com

Ben Cartwright-Cox
Port 179 Ltd
London
United Kingdom
Email: ben@benjojo.co.uk

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: 11 September 2023

J. Uttaro
A. Lingala
AT&T
K. Patel
Arrcus, Inc.
D. Rao
Cisco Systems
B. Wen
Comcast
A. Retana
Futurewei Technologies, Inc.
S. Sangli
Juniper Networks
P. Mohapatra
Sproute Networks

10 March 2023

# One Administrative Domain using BGP
draft-uttaro-idr-bgp-oad-00

## Abstract

This document defines a new External BGP (EBGP) peering type known as EBGP-OAD. EBGP-OAD peering is used between two EBGP peers that belong to One Administrative Domain (OAD).

## 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 11 September 2023.

## Copyright Notice

Table of Contents



## 1. Introduction

At each EBGP boundary, BGP path attributes are modified as per standard BGP rules [RFC4271]. This includes prepending the AS_PATH attribute with the autonomous-system number of the BGP speaker and stripping any IBGP-only attributes.

Some networks span more than one autonomous system and require more flexibility in the propagation of path attributes. These networks are said to belong to One Administrative Domain (OAD). It is desirable to carry IBGP-only attributes across EBGP peering when the peers belong to OAD. This document defines a new EBGP peering type known as EBGP-OAD. EBGP-OAD peering is used between two EBGP peers that belong to OAD. This document also defines rules for route announcement and processing for EBGP-OAD peers.

### 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. Discussion

Networks have traditionally been demarcated by an autonomous system/ BGP border which correlates to an administrative boundary. This paradigm no longer serves the needs of network designers or customers due to the decoupling of IGP from BGP, BGP-free core in the underlay (e.g. using BGP labeled unicast [RFC8277]), the use of BGP to facilitate multiple service overlays (e.g., L2VPN, L3VPN, etc.) spanning multiple regions and AS domains, and the instantiation of customer sites on multiple content service providers (CSPs).

For example, sites in a BGP/MPLS VPN [RFC4364] may be distributed across different AS domains. In some cases, the administrator of the VPN may prefer that some attributes are propagated to all their sites to influence the BGP decision process. An example could be LOCAL_PREF which is ignored if received on an EBGP session [RFC4271].

## 3. Operation

[RFC4271] defines two types of BGP peerings used during a BGP protocol session. As part of the extensions defined in this document, the EBGP peering is divided into two types:

1. EBGP as defined in [RFC4271].

2. EBGP-OAD as defined below.

The EBGP-OAD session is a BGP connection between two external peers in different Autonomous Systems that belong to OAD. In general, the EBGP-OAD speakers follow the EBGP route advertisement, route processing, path attribute announcement and processing rules as defined in [RFC4271]. However, EBGP-OAD speakers are also allowed to announce and receive any IBGP-only or non-transitive attributes that were restricted to remain within an Autonomous System [RFC4271].

Unless explicitly specified, all path attributes MAY be advertised over an EBGP-OAD session. The reception of any path attribute over an EBGP-OAD session MUST NOT result in an error, unless it is malformed. Received path attributes SHOULD NOT be ignored by the receiver, unless directed to by local policy.

Unless explicitly specified, the current processes for the advertisement of path attributes remains unchanged when advertised through an EBGP-OAD peering. The process for EBGP advertisement MUST take priority over the process for IBGP advertisement. For example, the AS_PATH attribute is modified as specified in Section 5.1.2 of [RFC4271], bullet b ("BGP speaker advertises the route to an external peer").

An EBGP-OAD speaker MUST support four-octet AS numbers and avertize the "support for four-octet AS number capability" [RFC6793] .

The following sections describe modifications to route advertisements and path attribute announcements that are specific to the EBGP-OAD peering.

### 3.1. Next Hop Handling

It is reasonable for EBGP-OAD peers to share a common Interior Gateway Protocol (IGP). In such a case, NEXT_HOP attribute and the Next Hop in the MP_REACH_NLRI attribute [RFC4760] MAY be left unchanged.

### 3.2. MULTI_EXIT_DISC (MED) Handling

The determination of the neighboring AS for the purpose of BGP Route Selection [RFC4271] MAY also consider the ASN of the EBGP-OAD peer. If so, all the peers in the receiving ASN MUST be configured to use the same criteria.

### 3.3. Route Reflection

BGP Route Reflection [RFC4456] is an alternative to full-mesh IBGP. The ORIGINATOR_ID and CLUSTER_LIST attributes MUST NOT be advertised over an EBGP-OAD session. If received, the procedures in [RFC7606] apply.

## 4. Deployment and Operational Considerations

For the EBGP-OAD session to operate as expected, both BGP speakers MUST be configured with the same session type. If only one BGP speaker is configured that way, and the other uses an EBGP session, the result is that some path attributes may be ignored and others will be discarded, but the BGP session will remain operational.

The default BGP peering type for a session that is across autonomous systems SHOULD be EBGP. BGP implementation SHOULD provide a configuration-time option to enable the EBGP-OAD session type. If the session type is changed once the BGP connection has been established, the BGP speaker MUST readvertise its entire Adj-RIB-Out to its peer. Requesting a route refresh [RFC7313] is RECOMMENDED.

The requirement that Import and Export Policies exist [RFC8212] SHOULD be disabled if both peers are configured with the EBGP-OAD session type.

If multiple peerings exist between two autonomous systems that belong to OAD, all SHOULD be configured consistently. Improper configuration may result in inconsistent or unexpected forwarding. The inconsistent use of EBGP-OAD sessions is out of scope of this document.

BGP Confederations [RFC5065] provide similar behavior, on a session by session basis, as what is specified in this document. The use of confederations with an EBGP-OAD peering is out of scope of this document.

The consideration of the ASN of the EBGP-OAD peer to determine the neighboring AS for MED comparison Section 3.2 may result in the creation persistent route oscillations, similar to the Type II Churn described in [RFC3345]. [RFC7964] provides solutions and recommendations to address this issue.

## 5. IANA Considerations

This memo includes no request to IANA.

## 6. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP protocol, such as those described in [RFC4271] and [RFC4272].

This document defines a new BGP session type which combines the path attribute propagation rules for EBGP and IBGP peering. Any existing security considerations related to existing path attributes apply to the new EBGP-OAD session type.

By combining the path attribute propagation rules, IBGP information may now be propagated to another autonomous system. However, it is expected that the new session type will only be enabled when peering with a router that also belongs to OAD. If misconfigured, the impact is minimal due to the fact that both [RFC4271] and [RFC7606] define mechanisms to deal with unexpected path attributes.

## 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>.

[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>.

[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>.

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

[RFC7313] Patel, K., Chen, E., and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", RFC 7313, DOI 10.17487/RFC7313, July 2014, <https://www.rfc-editor.org/info/rfc7313>.

[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>.

[RFC8212] Mauch, J., Snijders, J., and G. Hankins, "Default External BGP (EBGP) Route Propagation Behavior without Policies", RFC 8212, DOI 10.17487/RFC8212, July 2017, <https://www.rfc-editor.org/info/rfc8212>.

## 7.2. Informative References

[RFC3345] McPherson, D., Gill, V., Walton, D., and A. Retana, "Border Gateway Protocol (BGP) Persistent Route Oscillation Condition", RFC 3345, DOI 10.17487/RFC3345, August 2002, <https://www.rfc-editor.org/info/rfc3345>.

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

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

[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>.

[RFC7964] Walton, D., Retana, A., Chen, E., and J. Scudder, "Solutions for BGP Persistent Route Oscillation", RFC 7964, DOI 10.17487/RFC7964, September 2016, <https://www.rfc-editor.org/info/rfc7964>.

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

## Acknowledgements

TBD

## Authors' Addresses

Jim Uttaro
AT&T
Email: ju1738@att.com

Avinash Lingala
AT&T
Email: ar977m@att.com

Keyur Patel
Arrcus, Inc.
Email: keyur@arrcus.com

Dhananjaya Rao
Cisco Systems
Email: dhrao@cisco.com

Bin Wen
Comcast
Email: bin_wen@comcast.com

Alvaro Retana
Futurewei Technologies, Inc.
Email: alvaro.retana@futurewei.com

Srihari Sangli
Juniper Networks
Email: ssangli@juniper.net

Pradosh Mohapatra
Sproute Networks
Email: pradosh@sproute.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: 7 September 2023

Z. Wu
H. Wang
L. Wang
Z. Tan
X. Ding
Huawei Technologies

6 March 2023

# BGP Flowspec Redirect Load Balancing Group Community
draft-wu-idr-flowspec-redirect-group-01

## Abstract

This document defines an extension to "BGP Community Container Attribute" [draft-ietf-idr-wide-bgp-communities], which allows flowspec redirection to multiple paths. This extended community serves to redirect traffic to a load balancing group and supports both equal-cost multi-path(ECMP) and unequal-cost multi-path(UCMP) scenarios.

## 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 7 September 2023.

## Copyright Notice

Table of Contents

## 1. Introduction

"Redirect to IP Extended Community", defined in [I-D.ietf-idr-flowspec-redirect-ip], allows traffic to be redirected to a specific IPv4 or IPv6 address, and [I-D.ietf-idr-ts-flowspec-srv6-policy] defines the redirection action to a SRv6 tunnel by additionally carrying the "Color Extended Community" [RFC8955].

However, scenarios involving redirection load balancing are not described in both documents. Although in some implementations, Equal-cost multi-path(ECMP) of "Redirect to IP" action can be achieved by encoding multiple redirect Extended Communities, the current set of mechanisms can hardly support neither ECMP of SRv6 tunnels nor unequal-cost multi-path(UCMP) of either types.

This document defines an extension to "BGP Community Container Attribute" [I-D.ietf-idr-wide-bgp-communities], the "Redirect Load Balancing Group" community. It is a new type of wide community container attribute with encoding format of multiple redirection path TLVs. Each of these TLVs represents a different redirection action. It allows traffic redirection to a load balancing group and supports both ECMP and UCMP scenarios.

The "Redirect Load Balancing Group" community is intended to be used within flowspec-v1 scenarios, the compatibility and interactions with flowspec-v2 is outside the scope of this document.

### 1.1. Terminology

This document introduces the following terms:

ECMP: Equal-Cost Multi-Path

UCMP: Unequal-Cost Multi-Path

Redirect Group: Redirect Load Balancing Group Community, a new type of BGP Community Container Attribute defined by this document

Path-tlv: Sub-tlv of the BGP Wide Community Parameter TLV, each represents a redirection path

## 2. Redirect Load Balancing Group Community

This document defines a new type of "BGP Community Container Attribute", the "Redirect Load Balancing Group" community type. The format complies with "BGP Community Container Attribute" [I-D.ietf-idr-wide-bgp-communities] and is shown below:

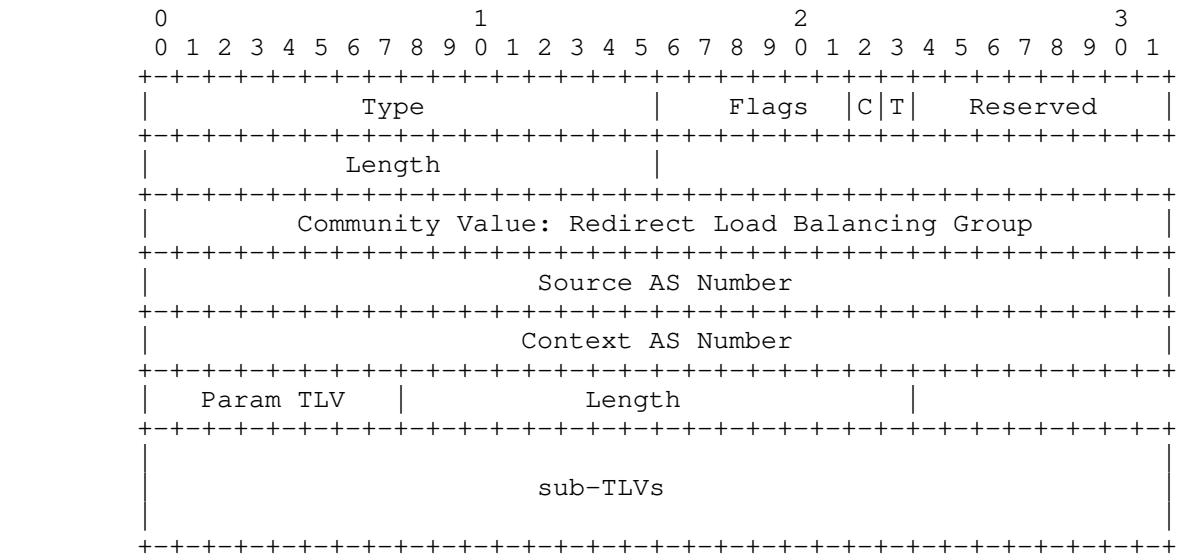



Figure 1: Redirect Load Balancing Group Community Format

The Type, Flags, Reserved and Length fields comply with the "BGP Community Container Attribute Common Header" definition.

The container type MUST be 1, which represents BGP Wide Community.

The Length field represents the total length of the container's contents in octets.

## 2.1. Community Value

The Community Value, Source AS Number and Context AS Number fields comply with the corresponding definition in "BGP Community Container Attribute".

Community Value: 4 octets value that represents the "Redirect Load Balancing Group" community type. The value is TBD and requires IANA registration; See Section 5.1.

## 2.2. Param TLV

The BGP Wide Community Parameter TLV (Sub-Type 3) contains a list of path-tlvs, comply with "BGP Wide Community Parameter(s) TLV" section of "BGP Community Container Attribute".

The Parameter TLV MUST present and SHOULD appear only once in a "Redirect Load Balancing Group" community container, no or multiple present SHOULD be considered malformed.

Sub-Type: Type 3 (BGP Wide Community Parameter TLV)

Length: Length of all the sub-TLVs in octets.

## 2.3. Sub-TLVs(Path-tlvs)

The list of path-tlvs that Param Tlv contains. Each path-tlv represents a different redirection path.

The general format of the sub-TLVs comply with path-tlvs' format defined in "BGP Community Container Attribute", as below:

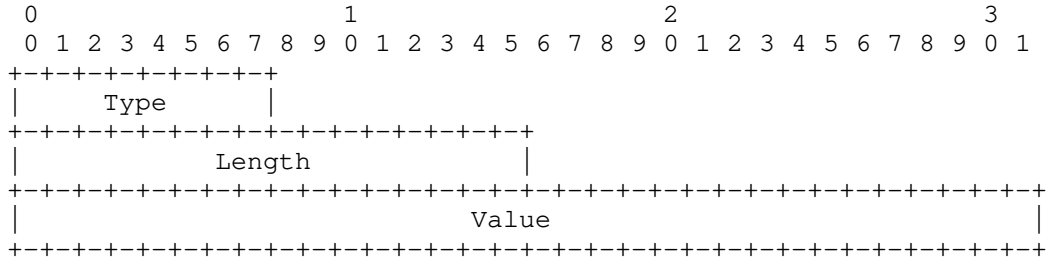



Figure 2: Param Sub-TlV Format

The Type field is an octet from 1~254 (0 and 255 are reserved). Supported type of the sub-TLVs includes:

Type 1: IPv4 Prefix Only

Type 2: IPv4 Prefix with Weight

Type 3: IPv4 Prefix with Color

Type 4: IPv4 Prefix with Color and Weight

Type 5: IPv6 Prefix Only

Type 6: IPv6 Prefix with Weight

Type 7: IPv6 Prefix with Color

Type 8: IPv6 Prefix with Color and Weight

These sub-TLV types SHOULD be used exclusively within "Redirect Load Balancing Group" community containers.

The Length represents the length of the "Value" field in octets, and it is fixed for each specific sub-TLV.

If the length and type of a sub-TLV do not match, the "Redirect Load Balancing Group" community container SHOULD be considered malformed.

If a sub-TLV is a total dupilication of a previous one, the latter sub-TLV MUST be ignored.

In principle, sub-TLVs of different types may be combined in any mode. The supported combinations depend on the specific implementation.

### 2.3.1. Path-tlv Type 1: IPv4 Prefix Only

Indicating the redirection path is unweighted and to a IPv4 address. The format is shown below:

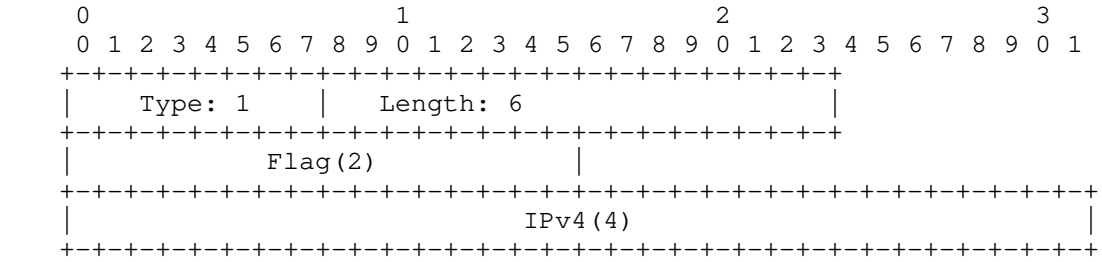



Figure 3: Path-tlv Type 1: IPv4 Prefix Only

Length: MUST be 6.

Flags: 2 octets, reserved for future use, MUST be set to 0 upon the sender and MUST be ignored upon the receiver.

IPv4: 4-octet IPv4 address, redirection destination

### 2.3.2. Path-tlv Type 2: IPv4 Prefix with Weight

Indicating the redirection path is weighted and to a IPv4 address. The format is shown below:

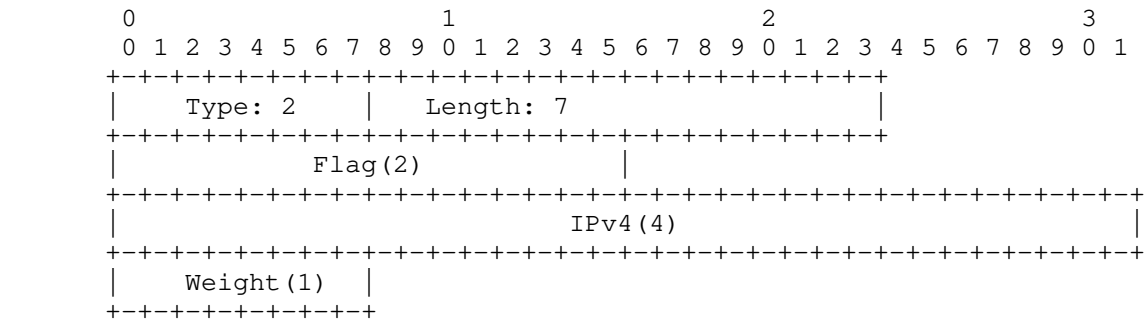



Figure 4: Path-tlv Type 2: IPv4 Prefix with Weight

Length: MUST be 7.

Flags: 2 octets, reserved for future use, MUST be set to 0 upon the sender and MUST be ignored upon the receiver.

IPv4: 4-octet IPv4 address, redirection destination

Weight: 1 octet, values from 1~255, load balancing weight

### 2.3.3. Path-tlv Type 3: IPv4 Prefix with Color

Indicating the redirection path is unweighted and to a SR-TE tunnel. The format is shown below:

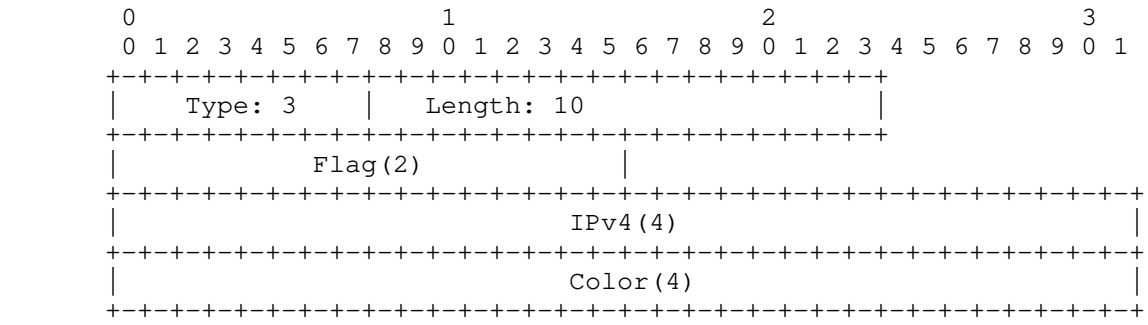



Figure 5: Path-tlv Type 3: IPv4 Prefix with Color

Length: MUST be 10.

Flags: 2 octets, reserved for future use, MUST be set to 0 upon the sender and MUST be ignored upon the receiver.

IPv4: 4-octet IPv4 address, SR-TE tunnel Endpoint for redirection

Color: 4 octets, SR-TE tunnel Color for redirection

### 2.3.4. Path-tlv Type 4: IPv4 Prefix with Color and Weight

Indicating the redirection path is weighted and to a SR-TE tunnel. The format is shown below:

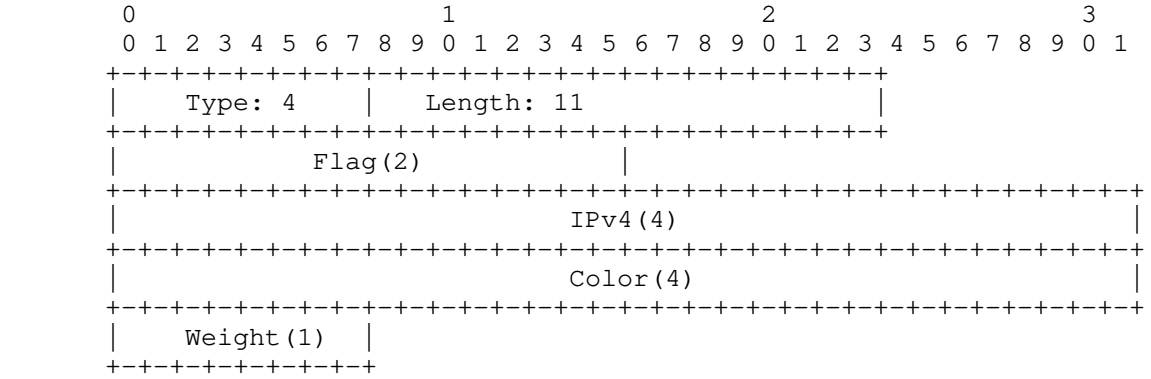



Figure 6: Path-tlv Type 4: IPv4 Prefix with Color and Weight

Length: MUST be 11.

Flags: 2 octets, reserved for future use, MUST be set to 0 upon the sender and MUST be ignored upon the receiver.

IPv4: 4-octet IPv4 address, SR-TE tunnel Endpoint for redirection

Color: 4 octets, SR-TE tunnel Color for redirection

Weight: 1 octet, values from 1~255, load balancing weight

### 2.3.5. Path-tlv Type 5: IPv6 Prefix Only

Indicating the redirection path is unweighted and to a IPv6 address. The format is shown below:

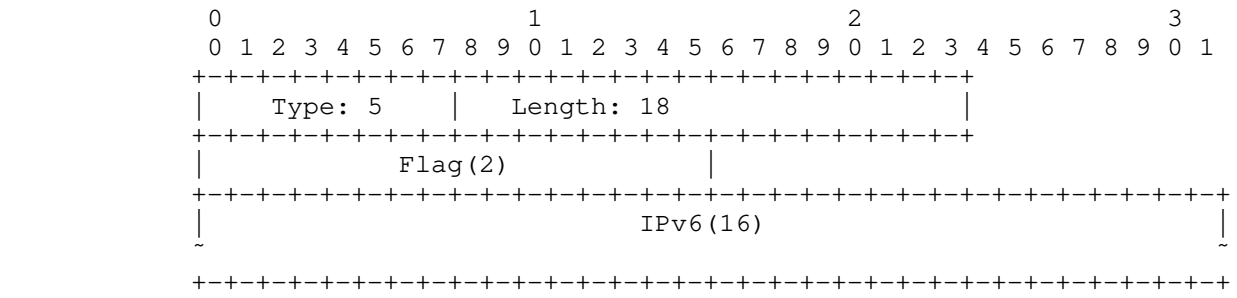



Figure 7: Path-tlv Type 5: IPv6 Prefix Only

Length: MUST be 18.

Flags: 2 octets, reserved for future use, MUST be set to 0 upon the sender and MUST be ignored upon the receiver.

IPv6: 16-octet IPv6 address, redirection destination

### 2.3.6. Path-tlv Type 6: IPv6 Prefix with Weight

Indicating the redirection path is weighted and to a IPv6 address. The format is shown below:

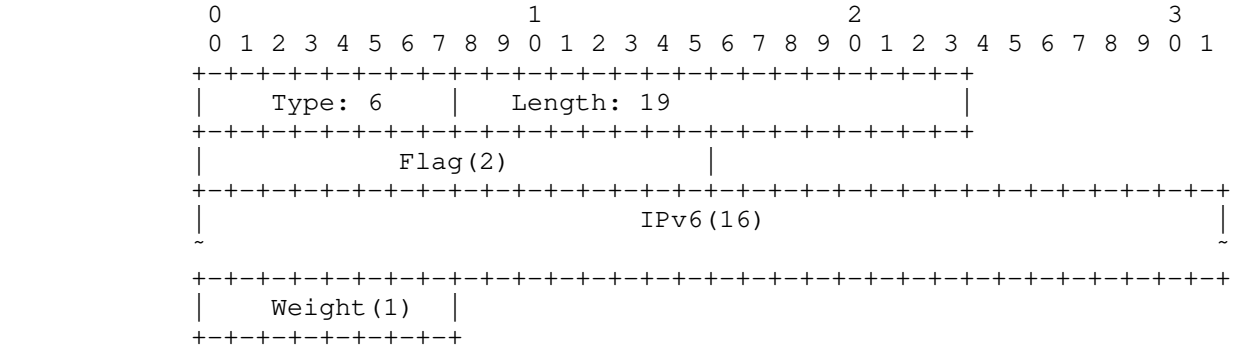



Figure 8: Path-tlv Type 6: IPv6 Prefix with Weight

Length: MUST be 19.

Flags: 2 octets, reserved for future use, MUST be set to 0 upon the sender and MUST be ignored upon the receiver.

IPv6: 16-octet IPv6 address, redirection destination

Weight: 1 octet, values from 1~255, load balancing weight

### 2.3.7. Path-tlv Type 7: IPv6 Prefix with Color

Indicating the redirection path is unweighted and to a SRv6 tunnel. The format is shown below:

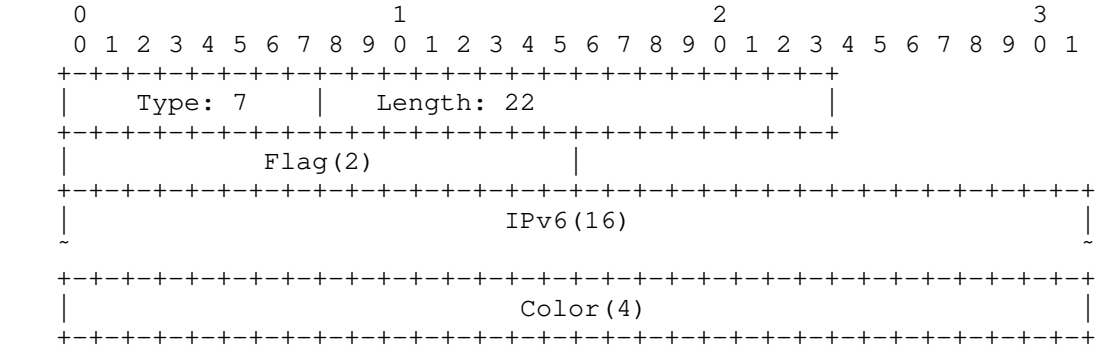



Figure 9: Path-tlv Type 7: IPv6 Prefix with Color

Length: MUST be 22.

Flags: 2 octets, reserved for future use, MUST be set to 0 upon the sender and MUST be ignored upon the receiver.

IPv6: 16-octet IPv6 address, SRv6 tunnel Endpoint for redirection

Color: 4 octets, SRv6 tunnel Color for redirection

### 2.3.8. Path-tlv Type 8: IPv6 Prefix with Color and Weight

Indicating the redirection path is weighted and to a SRv6 tunnel. The format is shown below:

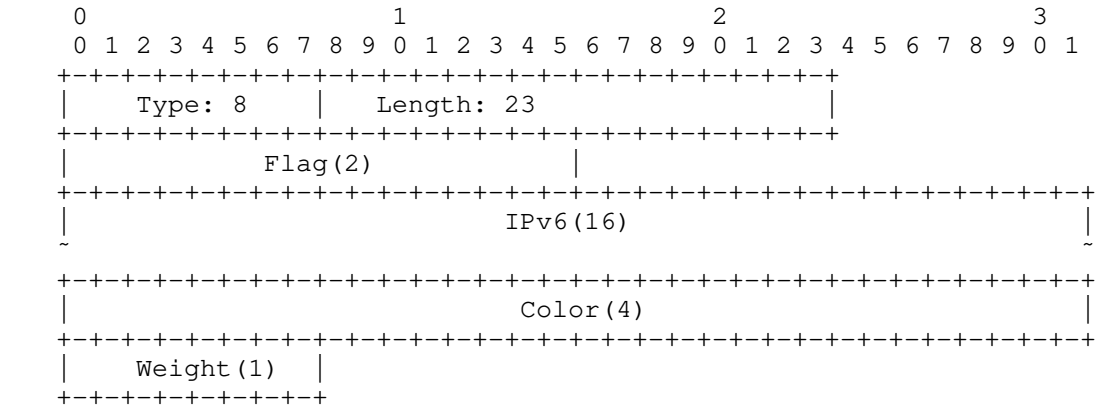


Figure 10: Path-tlv Type 8: IPv6 Prefix with Color and Weight

Length: MUST be 23.

Flags: 2 octets, reserved for future use, MUST be set to 0 upon the sender and MUST be ignored upon the receiver.

IPv6: 16-octet IPv6 address, SRv6 tunnel Endpoint for redirection

Color: 4 octets, SRv6 tunnel Color for redirection

Weight: 1 octet, values from 1~255, load balancing weight

## 3. Scenarios

This section describes a few use-case scenarios when deploying "Redirect Load Balancing Group" community type.

Weighted path-tlv types: Path-tlvs contain a Weight field, such as Type 2, 4, 6, 8

Unweighted path-tlv types: Path-tlvs do not contain a Weight field, such as Type 1, 3, 5, 7

### 3.1. ECMP

A system that originates a flowspec route with a "Redirect Load Balancing Group" community, among which its parameter TLV contains more than 1 path-tlvs. If not all path-tlvs are of a weighted type, these path-tlvs will form a ECMP group.

Implementations MUST be prepared to accept a Parameter TLV with both weighted and unweighted path-tlvs. In this case, the Weight field of the weighted path-tlv SHOULD be ignored.

### 3.2. UCMP

A system that originates a flowspec route with a "Redirect Load Balancing Group" community, among which its parameter TLV contains more than 1 path-tlvs. If all path-tlvs are of a weighted type, these path-tlvs will form a UCMP group.

In this case, the Weight field value of these path-tlvs SHOULD NOT be ignored, and the values are used as the ratio of the UCMP group.

## 4. Validation Procedure

In the absence of explicit configuration, a Redirect Group attribute MUST be validated before it is used for redirection action or sent to a BGP peer.

The validation procedure for a Redirect Group attribute follows the following rules:

* Each Path-tlv of the Redirect Group attribute SHOULD be validated separately. The validation of each path follows the validation procedure of Redirect to IP Action [I-D.ietf-idr-flowspec-redirect-ip].

* A Redirect Group attribute SHOULD be considered verified, only after all path-tlvs in the Redirect Group attribute are verified.

* If any path-tlvs are invalid, these paths SHOULD NOT participate in load-balance calculation and used for redirection actions.

* If any path-tlvs are invalid, the Redirect Group attribute SHOULD NOT be sent to a BGP peer.

## 5. Error Handling

Comply with Error Handling Procedure in "BGP Community Container Attribute" [I-D.ietf-idr-wide-bgp-communities].

In addition:

### 5.1. Redirect Group Wide Community Parameter TLV

A "Redirect Load Balancing Group" community container with no or multiple parameter TLVs SHOULD be considered malformed, and a "treat as withdraw" behavior is expected.

### 5.2. Redirect Group Wide Community Parameter Sub-TLVs

If the length and type of a sub-TLV do not match, the "Redirect Load Balancing Group" community container SHOULD be considered malformed, and a "treat as withdraw" behavior is expected.

## 6. Operational Considerations

The Extended Community attributes for redirection mentioned in this section include:

* Redirect to IP Extended Community [I-D.ietf-idr-flowspec-redirect-ip]

* Redirect to IPv6 Extended Community [I-D.ietf-idr-flowspec-redirect-ip]

* Redirect to SRv6 Policy [I-D.ietf-idr-ts-flowspec-srv6-policy]

### 6.1. Configuration Control

There SHOULD be an explicit configuration to control whether the Redirect Group attribute is used for redirection actions. In the absence of the explicit configuration(by default), the Redirect Group attribute MAY NOT take precedence over Extended Community attribute. With the explicit configuration, the Redirect Group attribute MAY take precedence over Extended Community attribute for redirection.

For clarity, the first scenario, in which the Redirect Group attribute does not take precedence, is called configuration situation A. And the second scenario is called configuration situation B.

### 6.2. Parsing

While receiving a flowspec route with Redirect Group attribute from a BGP peer:

* In configuration situation A, the Redirect Group attribute SHOULD NOT be used for redirection actions. If the route carries Extended Community attributes for redirection, these attributes MAY be used to generate the redirection actions. The Redirect Group attribute SHOULD still be saved locally and advertised with the fowspec route to other appropriate peers.

* In configuration situation B, the Redirect Group attribute SHOULD take precedence over Extended Community attribute for redirection. If the route carries Extended Community attributes for redirection, these attributes SHOULD NOT be used to generate the redirection actions, but SHOULD still be saved locally and advertised with the fowspec route to other appropriate peers.

### 6.3. Formating

While encoding a local-generated flowspec route:

* In configuration situation A, a Redirect Group attribute SHOULD NOT be encoded. Appropriate Extended Community attributes MAY be used for specifying redirection actions.

* In configuration situation B, the Redirect Group attribute SHOULD be encoded for specifying redirection actions, despite of there is one or more paths. For the sake of compatibility, we MAY select the path with the lowest IP address from the paths of the Redirect Group attribute and encode it with appropriate Extended Community attributes. During this selection, an IPv4 address is preferred over an IPv6 address.

While encoding a flowspec route learned from other BGP peers:

* In configuration situation A, the Redirect Group attribute MUST be encoded without modification.

* In configuration situation B, the Redirect Group attribute MUST pass the validation procedure before it is encoded and sent to a BGP peer.

## 7. IANA Considerations

### 7.1. BGP Wide Communities Community Type : Redirect Group

This document requests a new community value under "Registered Type 1 BGP Wide Community Community Types" registery. This registery is defined and requested in "BGP Community Container Attribute" [I-D.ietf-idr-wide-bgp-communities].

Requested value:

| Name | Type Value |
|---|---|
| Redirect Load Balancing Group | TBD |

## 8. Security Considerations

A system that originates a flowspec route with a "Redirect Load Balancing Group" BGP wide community can cause many receivers of that route to redirect traffic to a single next-hop, overwhelming that next-hop and resulting in inadvertent or deliberate denial-of-service. This is also a concern about the "redirect to IP" extended community, therefore this document introduces no additional security considerations than those already covered in [RFC8955].

## 9. References

### 9.1. Normative References

[I-D.ietf-idr-flowspec-redirect-ip]
Uttaro, J., Haas, J., Texier, M., akarch@cisco.com, Ray, S., Simpson, A., and W. Henderickx, "BGP Flow-Spec Redirect to IP Action", Work in Progress, Internet-Draft, draft-ietf-idr-flowspec-redirect-ip-02, 5 February 2015, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-flowspec-redirect-ip-02>.

[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-10, 2 March 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-wide-bgp-communities-10>.

[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>.

## 9.2. References

[I-D.ietf-idr-ts-flowspec-srv6-policy]
Wenying, J., Liu, Y., Zhuang, S., Mishra, G. S., and S. Chen, "Traffic Steering using BGP FlowSpec with SRv6 Policy", Work in Progress, Internet-Draft, draft-ietf-idr-ts-flowspec-srv6-policy-01, 9 October 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-ts-flowspec-srv6-policy-01>.

[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>.

## Authors' Addresses

Zhiwen Wu
Huawei Technologies
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: wuzhiwen1@huawei.com

Haibo Wang
Huawei Technologies
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: rainsword.wang@huawei.com

Lili Wang
Huawei Technologies
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: lily.wong@huawei.com

Zhen Tan
Huawei Technologies
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: tanzhen6@huawei.com

Xiangfeng Ding
Huawei Technologies
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: dingxiangfeng@huawei.com

IDR
Internet-Draft
Intended status: Standards Track
Expires: 28 August 2023

X. Yi, Ed.
T. He, Ed.
China Unicom
H. Shi, Ed.
X. Ding
H. Wang
Huawei Technologies
24 February 2023

# Distribution of Service Metadata in BGP FlowSpec
draft-yi-idr-bgp-fs-edge-service-metadata-00

## Abstract

In edge computing, a service may be deployed on multiple instances within one or more sites, called edge service. The edge service is associated with an ANYCAST IP address, and the route of it along with service metadata can be collected by a central controller. The controller may process the metadata and distribute the result to ingress routers using BGP FlowSpec. The service metadata can be used by ingress routers to make path selections not only based on the routing cost but also the running environment of the edge services. This document describes a mechanism to distribute the information of the service routes and related service metadata using BGP FlowSpec.

Table of Contents



## 1. Introduction

Many modern services deploy their service instances in multiple sites to get better response time and resource utilization. These sites are often geographically distributed to serve the user demand. For some services such as VR/AR and intelligent transportation, the QoE will depend on both the network metrics and the compute metrics. For example, if the nearest site is overloaded due to the demand fluctuation, then steer the user traffic to another light-loaded sites may improve the QoE. To steer the traffic to the best site, the computing metadata of the site needs to be collected.

[I-D.ietf-idr-5g-edge-service-metadata] describes the BGP extension of distributing service route with network and computing-related metrics. The router connected to the site will received the service routes and service metadata sent from devices inside the edge site, and then generates the corresponding routes and distributes them to ingress routers. However, the route with service metadata on the router connected to the site can be also collected by a central controller using BGP LS. Then the central controller may process the metadata and distributes the result to the ingress router using BGP FlowSpec.

This document defines an extension of BGP FlowSpec to carry the service metadata along with the service route which is received from the controller. Using the service metadata and the service route, the ingress router can calculate the best site for the traffic, giving each user the best QoE.

### 1.1. Terminology

### 1.2. Requirements Language

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

## 2. BGP FlowSpec Extension for Service Metadata

The goal of the BGP FlowSpec extension is to distribute the information of the service route and metadata. A service is identified by an prefix and this information is carried using the existing Destination Prefix Component specified in [RFC8955] and [RFC8956]. [I-D.ietf-idr-ts-flowspec-srv6-policy] defines that the Color Extended Community and BGP Prefix-SID attribute is carried in the context of the FlowSpec NLRI.

In addition to that, this document proposes to carry the service metadata attribute(See Figure 1). The ingress router can compare the compute metric of different sites and steer the traffic into the best one using the SR policy. The metadata can be original values defined in [I-D.ietf-idr-5g-edge-service-metadata] or an aggregated one calculated using original values.

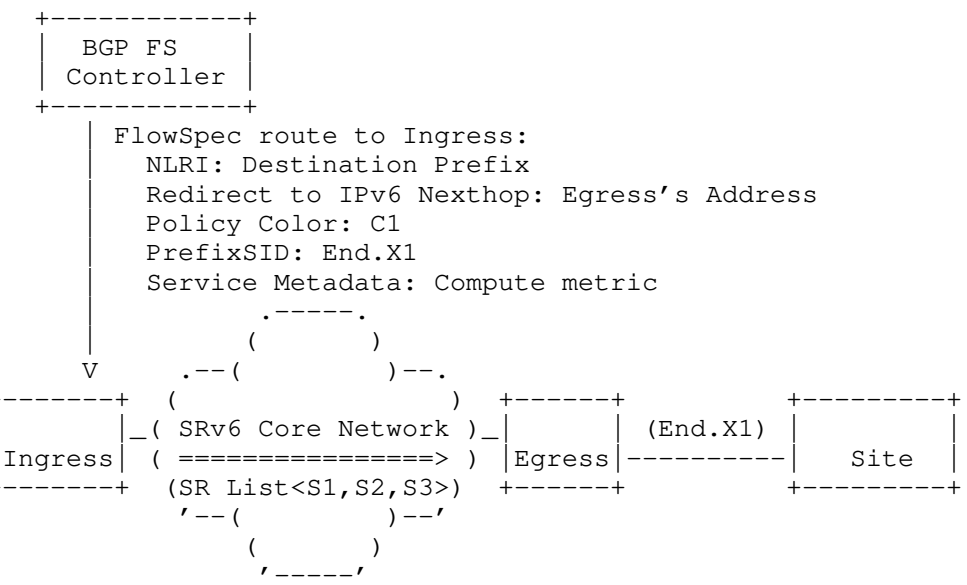



Figure 1: Example of using BGP FlowSpec to distribute the service route and metadata

## 2.1. Metadata Path Attribute TLV

The Metadata Path Attribute TLV is the same as defined in [I-D.ietf-idr-5g-edge-service-metadata], including the following three sub-TLVs:

1. Site Preference Index sub-TLV indicates the preference to choose the site.

2. Capacity Index sub-TLV indicates the capability of a site. One Edge Site can be in full capacity, reduced capacity, or completely out of service.

3. Load Measurement sub-TLV indicates the load level of the site.

## 2.2. Aggregated Metric Path Attribute TLV

The Aggregated Metric Path Attribute is a newly defined TLV(See Figure 2). It contains a single aggregated value which is calculated by the controller using the original metrics such as site preference, capacity and load measurement.

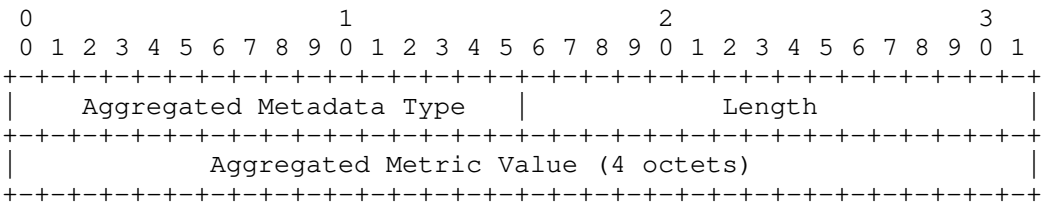



Figure 2: Aggregated Metric Path Attribute TLV format

* Type: identify the Aggregated Metadata Attribute, to be assigned by IANA.

* Length: the total number of the octets of the value field.

* Value: value of Aggregated Computing metric.

## 3. Security Considerations

TBD

## 4. IANA Considerations

This document requires IANA to assign the following code points from the registry called "BGP Path Attributes":

| Value | Description | Reference |
|---|---|---|
| TBD1 | Aggregated Metadata Type | Section 2.2 |

Table 1

## 5. Normative References

[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>.

[I-D.ietf-idr-5g-edge-service-metadata]
Dunbar, L., Majumdar, K., Wang, H., and G. S. Mishra, "BGP Extension for 5G Edge Service Metadata", Work in Progress, Internet-Draft, draft-ietf-idr-5g-edge-service-metadata-00, 2 December 2022, <https://www.ietf.org/archive/id/draft-ietf-idr-5g-edge-service-metadata-00.txt>.

[I-D.ietf-idr-ts-flowspec-srv6-policy]
Wenying, J., Liu, Y., Zhuang, S., Mishra, G. S., and S. Chen, "Traffic Steering using BGP FlowSpec with SRv6 Policy", Work in Progress, Internet-Draft, draft-ietf-idr-ts-flowspec-srv6-policy-01, 9 October 2022, <https://www.ietf.org/archive/id/draft-ietf-idr-ts-flowspec-srv6-policy-01.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>.

## Authors' Addresses

Xinxin Yi (editor)
China Unicom
Beijing
China
Email: yixx3@chinaunicom.cn

Tao He (editor)
China Unicom
Beijing
China
Email: het21@chinaunicom.cn

Hang Shi (editor)
Huawei Technologies
Beijing
China
Email: shihang9@huawei.com

Xiangfeng Ding
Huawei Technologies
Beijing
China
Email: dingxiangfeng@huawei.com

Haibo Wang
Huawei Technologies
Beijing
China
Email: rainsword.wang@huawei.com