

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 6, 2020

C. Cardona
P. Lucente
NTT
P. Francois
INSA
Y. Gu
Huawei
July 05, 2019

BMP Extension for Path Marking TLV
draft-cppy-grow-bmp-path-marking-tlv-00

Abstract

The BGP Monitoring Protocol (BMP) provides the monitoring of BGP adj-rib-in [RFC7854], BGP local-rib [I-D.ietf-grow-bmp-local-rib] and BGP adj-rib-out [I-D.ietf-grow-bmp-adj-rib-out] through Route Monitoring (RM) messages. With the capability of allowing optional data to be added to the RM Messages in the format of TLV draft-lucente-bmp-tlv [I-D.lucente-bmp-tlv], more information about the BGP Update message encapsulated in the RM can be revealed. This document proposes an extension to BMP to describe the BGP path status through the definition and use of Path Marking TLV.

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 RFC 2119 [RFC2119] RFC 8174 [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 January 6, 2020.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
2. Path Marking TLV for the RM Message	3
2.1. Path Type	3
2.2. Reason String	4
3. Acknowledgements	5
4. IANA Considerations	6
4.1. Path Marking TLV	6
4.2. Path Marking TLV Reason String	6
5. Security Considerations	6
6. Normative References	6
Authors' Addresses	7

1. Introduction

For a given prefix, multiple paths with different path status, e.g., the "best-path", the "best-external-path" and so on, may co-exist in the BGP module upon the local policy processing. In addition, during the whole process, from receiving a BGP route to advertising it, a path can also undergo various status in different processing states. Such path status information is currently not carried in the BGP Update Message RFC4271 [RFC4271]. However, they can be useful to enable a lot of applications. For example, for traffic steering purposes in a SDN environment, the operator/SDN controller needs the reachability information of multiple paths to ensure the selected optimized route is reachable.

This document defines a so-called Path Marking TLV to convey the BGP path status information to the BMP server. The BMP Path Marking is defined to be prepended in the BMP Route Monitoring (RM) Message.

2. Path Marking TLV for the RM Message

Per RFC4271 [RFC4271], the BMP RM Message consists of the Common Header, Per-Peer Header, and the BGP Update PDU. According to draft-lucente-bmp-tlv [I-D.lucente-bmp-tlv], optional trailing data in TLV format is allowed in the BMP RM Message to convey characteristics of transported NLRIs (ie. to help stateless parsing) or vendor-specific data. Such TLV types are to be defined per each application.

To include the path status along with each BGP path, we define the Path Marking TLV, shown as follows.

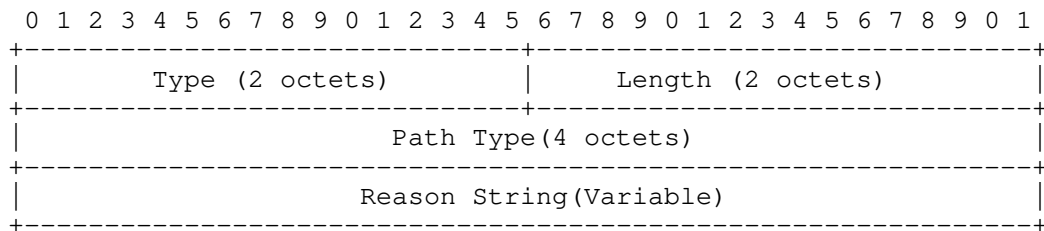


Figure 1: Path Marking TLV

- o Type = TDB1 (2 Octets): Path Marking.
- o Length (2 Octets): indicates the length of the value field of the Path Marking TLV. The value field further consists of the Path-Type field and Reason String field.
- o Path-Type (4 Octets): indicates the path status of the BGP Update PDU encapsulated in the RM Message. Currently 8 types of path status are defined, as shown in Table 1.
- o Reason String (Variable): indicates the reasons/explanations of the path status indicated in the Path Type field. The detailed Reason String format is defined in Figure 2.

2.1. Path Type

Value	Path type
0x0000	Unknown
0x0001	Best path
0x0002	Best external path
0x0004	Primary path
0x0008	Backup path
0x0010	Non-installed path
0x0020	Unreachable next-hop

Table 1: Path Type

The Path type field contains a bitfield where each bit encodes a specific role of the path. Multiple bits may be set when a path is used in multiple roles.

The best-path is defined in RFC4271 [RFC4271] and the best-external path is defined in draft-ietf-idr-best-external [I-D.ietf-idr-best-external].

A primary path is a recursive or non-recursive path that can be used all the time as long as a walk starting from this path can end to an adjacency draft-ietf-rtgwg-bgp-pic [I-D.ietf-rtgwg-bgp-pic]. A prefix can have more than one primary path. A best-path is also considered as a primary path.

A backup path is also installed in the RIB, but it is not used until some or all primary paths become unreachable. Backup paths are used for fast convergence in the event of failures.

All other reachable paths are marked as 'Non-installed'.

Lastly, all paths that are considered unreachable are marked as 'Unreachable next-hop'. Unreachable paths may be sent only in special cases.

2.2. Reason String

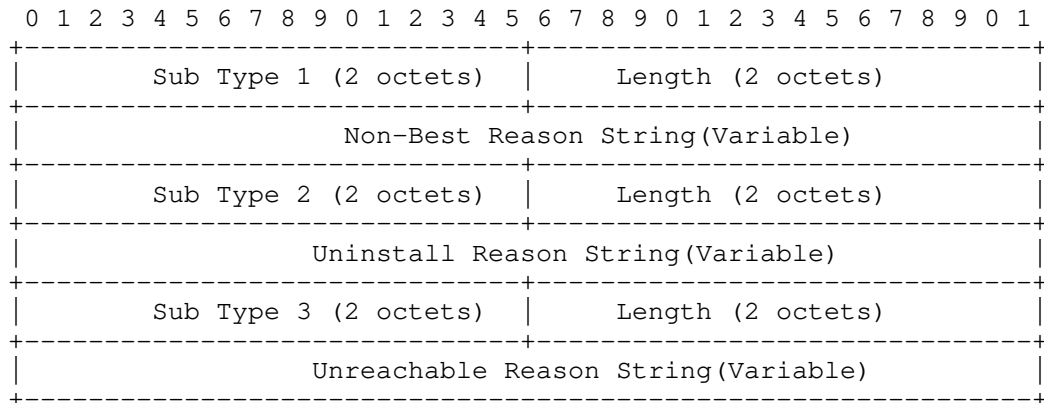


Figure 2: Reason String field

- o Sub Type 1 (2 Octets) = TDB2: Non-Best Reason String.
- o Length (2 Octets): indicates the length of the value field of the Non-Best Reason String.
- o Non-Best Reason String (Variable): includes user specific description of the non-best reason in the format of ASCII string.
- o Sub Type 2 (2 Octets) = TDB3: Uninstalled Reason String.
- o Length (2 Octets): indicates the length of the value field of the Non-Best Reason String.
- o Uninstalled Reason String (Variable): includes user specific description of the uninstalled (into RIB) reason in the format of ASCII string.
- o Sub Type 3 (2 Octets) = TDB4: Unreachable Reason String.
- o Length (2 Octets): indicates the length of the value field of the Non-Best Reason String.
- o Unreachable Reason String (Variable): includes user specific description of the unreachable reason in the format of ASCII string.

3. Acknowledgements

TBD.

4. IANA Considerations

This document requests that IANA assign the following new parameters to the BMP parameters name space.

4.1. Path Marking TLV

This document defines the Path Marking TLV with Type = TDB1: Path Marking (Section 2).

4.2. Path Marking TLV Reason String

This document defines three new sub types of the Reason String in the Path Marking TLV (Section 2.2).

Sub Type 1 = TDB2: Non-Best Reason String.

Sub Type 2 = TDB3: Uninstalled Reason String.

Sub Type 3 = TDB4: Unreachable Reason String.

5. Security Considerations

It is not believed that this document adds any additional security considerations.

6. Normative References

[I-D.ietf-grow-bmp-adj-rib-out]

Evens, T., Bayraktar, S., Lucente, P., Mi, K., and S. Zhuang, "Support for Adj-RIB-Out in BGP Monitoring Protocol (BMP)", draft-ietf-grow-bmp-adj-rib-out-06 (work in progress), June 2019.

[I-D.ietf-grow-bmp-local-rib]

Evens, T., Bayraktar, S., Bhardwaj, M., and P. Lucente, "Support for Local RIB in BGP Monitoring Protocol (BMP)", draft-ietf-grow-bmp-local-rib-04 (work in progress), June 2019.

[I-D.ietf-idr-best-external]

Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H. Gredler, "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-05 (work in progress), January 2012.

- [I-D.ietf-rtgwg-bgp-pic]
Bashandy, A., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", draft-ietf-rtgwg-bgp-pic-09 (work in progress), April 2019.
- [I-D.lucente-bmp-tlv]
Lucente, P., Gu, Y., and H. Smit, "TLV support for BMP Route Monitoring and Peer Down Messages", draft-lucente-bmp-tlv-00 (work in progress), July 2019.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<https://www.rfc-editor.org/info/rfc5492>>.
- [RFC7854] Scudder, J., Ed., Fernando, R., and S. Stuart, "BGP Monitoring Protocol (BMP)", RFC 7854, DOI 10.17487/RFC7854, June 2016, <<https://www.rfc-editor.org/info/rfc7854>>.
- [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

Camilo Cardona
NTT
164-168, Carrer de Numancia
Barcelona 08029
Spain

Email: camilo@ntt.net

Paolo Lucente
NTT
Siriusdreef 70-72
Hoofddorp, WT 2132
Netherlands

Email: paolo@ntt.net

Pierre Francois
INSA
Lyon
France

Email: Pierre.Francois@insa-lyon.fr

Yunan Gu
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: guyunan@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 11, 2022

Y. Gu
Huawei
H. Chen
China Telecom Co., Ltd.
D. Ma
ZDNS
S. Zhuang
Huawei
July 10, 2021

BMP for BGP Route Leak Detection
draft-gu-grow-bmp-route-leak-detection-05

Abstract

According to Route-Leak Problem Definition [RFC7908], Route leaks refer to the case that the delivery range of route advertisements is beyond the expected range. For many current security protection solutions, the ISPs (Internet Service Providers) are focusing on finding ways to prevent the happening of BGP [RFC4271] route leaks. However, the real-time route leak detection if any occurs is important as well, and serves as the basis for leak mitigation. This document extends the BGP Monitoring Protocol (BMP) [RFC7854] to provide a routing security scheme suitable for ISPs to detect BGP route leaks at the prefix level.

Requirements Language

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 11, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Terminology	2
2. Introduction	3
2.1. Actions Against Route Leaks	3
2.2. Challenges of the Current Actions against Route Leaks . .	4
3. Route Leak Detection (RLD) Design Considerations	5
4. BMP Support for RLD	5
4.1. RLD TLV Format	5
4.2. RLD TLV Usage	6
4.3. Coordination with iOTC and RLP	7
5. Acknowledgements	8
6. Contributors	8
7. IANA Considerations	8
8. Security Considerations	8
9. References	9
9.1. Normative References	9
9.2. Informative References	9
Authors' Addresses	10

1. Terminology

BMP: BGP Monitoring Protocol

BMS: BGP Monitoring Station

C2P: Customer to Provider

ISP: Internet Service Provider

P2C: Provider to Customer

P2P: Peer to Peer

RIB: Routing Information Base

RLP: Route Leak Protection

RLD: Route Leak Detection

2. Introduction

RFC7908 [RFC7908] defines "Route Leak" as: A route leak is the propagation of routing announcement(s) beyond their intended scope, which can result in possible situations such as eavesdropping, device overload, routing black hole and so on. More specifically, the intended scope of route announcements is usually defined by local route filtering/distribution policies within devices. These policies are designed to realise the pair-wise peering business relationships between ASes (autonomous systems), which include Customer to Provider (C2P), Peer to Peer (Peer to Peer), and Provider to Customer (P2C). In a C2P relationship, the customer pays the transit provider for traffic sent between the two ASes. In return, the customer gains access to the ASes that the transit provider can reach, including those which the transit provider reaches through its own transit providers. In a P2P relationship, the peering ASes gain access to each other's customers, typically without either AS paying the other AS Relationships, Customer Cones, and Validation [Luckie].

More precisely, the route leaks we discuss in this draft, referring to Type 1, 2, 3, and 4 Route Leaks defined in RFC7908 [RFC7908], can be summarized as: a route leak occurs when a route received from a transit provider or a lateral peer is propagated to another transit provider or a lateral peer.

2.1. Actions Against Route Leaks

There are several types of approaches against route leak from different perspectives. In this draft, we mainly discuss the following three types:

- o Route leak prevention: The approach to prevent route leak from happening. Commonly used methods include inbound/outbound prefix/peer/AS filtering policies configured at the ingress/egress nodes of ASes per the propagation of BGP routes.
- o Route leak detection: The approach to detect the existence of route leaks that happen at either the local AS, or upstream AS per the propagation of BGP routes. An intuitive way of detecting route leak is by checking the business relationship information at

both the ingress and egress nodes of the local AS along the BGP route propagation path with the route leak violation rules defined in RFC7908 [RFC7908]. Thus, it requires the knowledge of the actual route propagation trace, as well as the resulting business relationship information at the ingress and egress nodes. With the above information collected, the analysis can be done by the routing device or a centralized server. This draft specifies one such method.

- o Route leak mitigation: The approach to mitigate route leaks that already happened at either the local AS, or upstream AS per the propagation of BGP routes. Commonly used methods include reject, drop or stop propagating the invalid routes once detected the existence of leaks.

The above mentioned actions can be used alone or in combination, depending on the entities (routing devices, network manager) that execute the actions, and the relative positions of the executing entities from the leaking point (local or downstream).

2.2. Challenges of the Current Actions against Route Leaks

Route Leak Prevention [I-D.ietf-idr-bgp-open-policy] updates the BGP Open negotiation process with a new BGP capability to exchange the BGP Roles between two BGP speakers, and also proposes to use a new BGP attribute, called the iOTC (Internal Only To Customer) Path attribute to mark routes according to the BGP Roles established in Open Message. The iOTC attribute of the incoming route is set at the ingress node of the local AS, and is conveyed with the BGP Update to the egress node of the local AS for outbound filtering to prevent route leaks in the local AS. This attribute is removed at the egress node before the BGP Update is sent to eBGP neighbors. For representation simplification, we use iOTC to refer to the method specified in Route Leak Prevention [I-D.ietf-idr-bgp-open-policy].

Route-Leak Detection and Mitigation

[I-D.ietf-grow-route-leak-detection-mitigation] describes a route leak detection and mitigation solution based on conveying route-leak protection (RLP) information in a well-known transitive BGP community, called the RLP community. The RLP community helps with detection and mitigation of route leaks that happen at the upstream AS (per the BGP routes propagation), as an Inter-AS solution. For representation simplification, we use RLP to refer to the method specified in Route-Leak Detection and Mitigation [I-D.ietf-grow-route-leak-detection-mitigation].

The above two drafts provide solutions for route leak prevention, detection and mitigation. To summarize:

- o iOTC is used for route leak prevention of the local AS. It does not provide the detection or mitigation of route leaks of either local AS or upstream AS per the BGP routes propagation.
- o iOTC is peer/AS-level route leak prevention, due to the fact the BGP Role negotiation is peer-level. It does not provide prefix-level route leak prevention.
- o RLP is used for route leak detection and mitigation of route leak that happens in the upstream AS (per the BGP Update distribution). It is prefix-level detection and mitigation.

Thus, there lacks method for local AS route leak detection.

3. Route Leak Detection (RLD) Design Considerations

Considering the challenges facing the existing approaches, this draft proposes a method called Route Leak Detection (RLD). It utilizes the BGP Monitoring Protocol (BMP) to convey the RLD information from to the BMP server to realize centralized leak detection. BMP is currently deployed by OTT and carriers to monitor the BGP routes, such as monitoring BGP Adj-RIB-In using the process defined in RFC7854 [RFC7854], and monitoring BGP Adj-RIB-Out using the process defined in RFC8671 [RFC8671]. On the other hand, the RLD information is in fact a representation of the business relationships between the local AS and its neighboring AS. It does not involve any information disclosure issue regarding third parties. Thus, a single ISP can deploy RLD without relying on any information from either other ISPs or other third parties.

4. BMP Support for RLD

4.1. RLD TLV Format

A RLD TLV is defined for the BMP Route Monitoring Message. Considering that the AS relationships are sometimes per route based instead of per peer/AS based, this TLV is appended to each route, following the BGP Update Message. The order of placing the RLD TLV among other BMP supported TLVs is out of the scope of this draft. The TLV format is defined as follows:

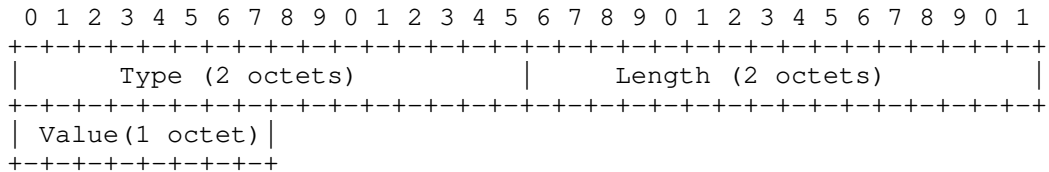


Figure 1: RLD TLV

- o Type (2 octets) = TBD1, the RLD TLV represents the prefix-level business relationship between the transmitter AS and the receiver AS. The local AS is a transmitter or a receiver, depending on if the route is an incoming route from a neighbor AS or an outgoing route to a neighbor AS.
- o Length (2 octets): Defines the length of the Value field. It SHOULD be set to 0x01, considering the Value field is of 1 octet fixed length.
- o Value (1 octet): Currently 4 values are defined to represent the business relationships, which are specified in Table 1.

Value	Business Relationship
0	P2C
1	C2P
2	P2P
3	I2I

Table 1: Business relationship value

4.2. RLD TLV Usage

The RLD TLV, presenting the business relationship between the neighbor AS and the local AS of the incoming route, SHOULD be prepended to the Adj-RIB-In at the ingress node of the local AS. The RLD TLV, representing the business relationship between the local AS and the neighbor AS of the outgoing route, SHOULD also be prepended to the Adj-RIB-Out at the egress node of the local AS. The BMP server, by analyzing the above two RLD TLVs of the same route, can use the rules defined in RFC7908 [RFC7908] to detect the existence of any route leak. As example is shown in Figure 2.

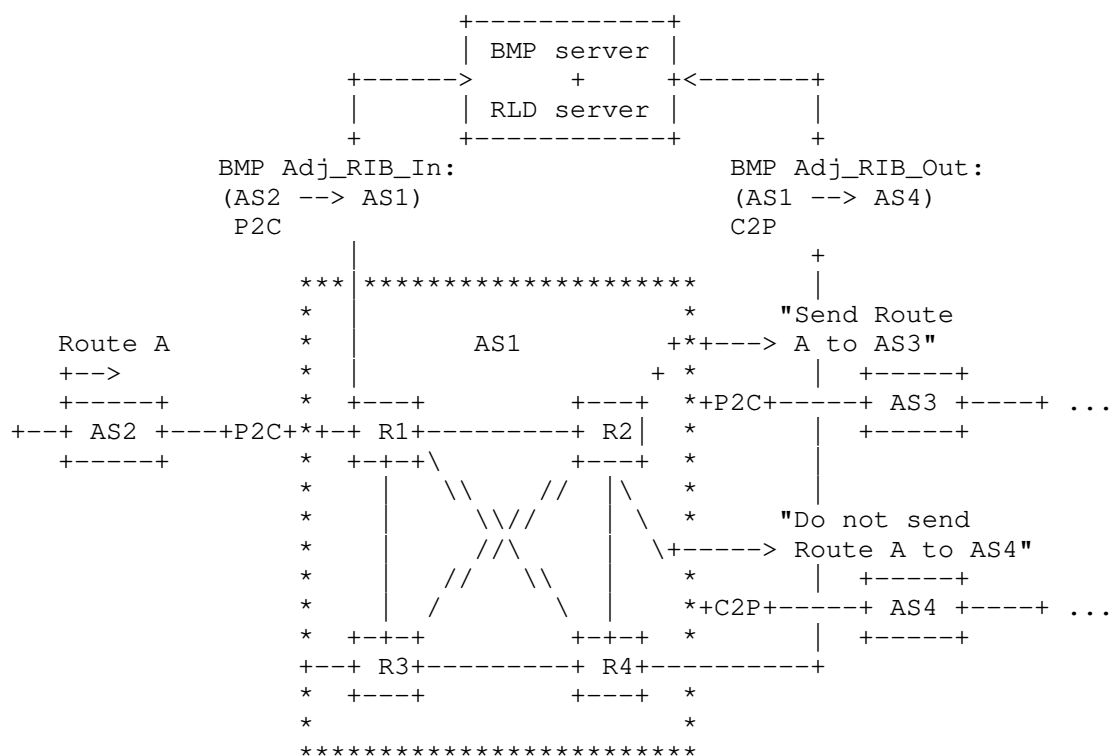


Figure 2: RLD depolyment by a single ISP

As shown in Figure 2, with the RLD TLV attached to each Route Monitoring Message, the RLD server (also working as the BMP server) combines the BMP Adj_RIB_In message collected from R1 and the BMP Adj_RIB_Out message collected from R4 to decide if there's a route leak. For example, if the RLD TLV in R1's Adj_RIB_In message indicates a value of "0", and the RLD TLV in R4's Adj_RIB_Out message indicates a value of "1", then the RLD server knows there exists a route leak.

4.3. Coordination with iOTC and RLP

RLD can be used as a complementary method to the existing methods against route leaks. More specifically, RLD can coordination with both iOTC and RLP.

- o With the settlement of the iOTC draft, the iOTC attribute is naturally included in the BGP Update and can be collected to the BMP server without BMP extension. With the RLD TLV collected also

by BMP (more specifically, the iBGP Adj-RIB-In of the ingress node), the BMP server can do validate the consistency of the iOTC attribute with the RLD. If contradiction detected, the BMP server may further check the bussiness contract for the actual business relationship.

- o For special prefixes that does not obey the peer/AS level business relationship negotiated through BGP Open Message, the BMP server can use the RLD TLV to detect such routes since the RLD TLV is set at prefix level.
- o For devices that do not support RLP, using RLD to collect the BGP routes, which conveys the RLD information from upstream ASes, allows the BMP server to detect and mitigate the route leaks that happen in the upstream AS. In other words, the detection and mitigation process can be also done in the BMP server, should the BMP server collects the BGP Update messages at the ingress or egress nodes.

5. Acknowledgements

6. Contributors

Haibo Wang

Huawei

Email: rainsword.wang@huawei.com

7. IANA Considerations

This document defines the following new BMP Route Monitoring message TLV type (Section 4.1):

- o Type = TBD1, the RLD TLV represents the prefix-level business relationship between the transmitter AS and the receiver AS. The local AS is a transmitter or a receiver, depending on if the route is an incoming route from a neighbor AS or an outgoing route to a neighbor AS.

8. Security Considerations

It is not believed that this document adds any additional security considerations.

9. References

9.1. Normative References

- [I-D.ietf-grow-route-leak-detection-mitigation]
Sriram, K. and A. Azimov, "Methods for Detection and Mitigation of BGP Route Leaks", draft-ietf-grow-route-leak-detection-mitigation-05 (work in progress), April 2021.
- [I-D.ietf-idr-bgp-open-policy]
Azimov, A., Bogomazov, E., Bush, R., Patel, K., and K. Sriram, "Route Leak Prevention using Roles in Update and Open messages", draft-ietf-idr-bgp-open-policy-15 (work in progress), January 2021.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [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>>.
- [RFC7854] Scudder, J., Ed., Fernando, R., and S. Stuart, "BGP Monitoring Protocol (BMP)", RFC 7854, DOI 10.17487/RFC7854, June 2016, <<https://www.rfc-editor.org/info/rfc7854>>.
- [RFC7908] Sriram, K., Montgomery, D., McPherson, D., Osterweil, E., and B. Dickson, "Problem Definition and Classification of BGP Route Leaks", RFC 7908, DOI 10.17487/RFC7908, June 2016, <<https://www.rfc-editor.org/info/rfc7908>>.
- [RFC8671] Evens, T., Bayraktar, S., Lucente, P., Mi, P., and S. Zhuang, "Support for Adj-RIB-Out in the BGP Monitoring Protocol (BMP)", RFC 8671, DOI 10.17487/RFC8671, November 2019, <<https://www.rfc-editor.org/info/rfc8671>>.

9.2. Informative References

- [Luckie] claffy, M. L. M. L. A. D. V. G. K., "AS Relationships, Customer Cones, and Validation", October 2013.

[Siddiqui]

Ramirez, M. S. S. D. M. M. Y. R. S. X. M. W., "Route Leak Detection Using Real-Time Analytics on local BGP Information", 2014.

Authors' Addresses

Yunan Gu
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: guyunan@huawei.com

Huanan Chen
China Telecom Co., Ltd.
109 Zhongshan W Ave
Guangzhou 510630
China

Email: chenhn8.gd@chinatelecom.cn

Di Ma
ZDNS
4 South 4th St. Zhongguancun
Beijing, Haidian
China

Email: madi@zdns.cn

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

Email: zhuangshunwan@huawei.com

Global Routing Operations
Internet-Draft
Updates: 7854 (if approved)
Intended status: Standards Track
Expires: 4 March 2022

T. Evens
S. Bayraktar
M. Bhardwaj
Cisco Systems
P. Lucente
NTT Communications
31 August 2021

Support for Local RIB in BGP Monitoring Protocol (BMP)
draft-ietf-grow-bmp-local-rib-13

Abstract

The BGP Monitoring Protocol (BMP) defines access to local Routing Information Bases (RIBs). This document updates BMP (RFC 7854) by adding access to the Local Routing Information Base (Loc-RIB), as defined in RFC 4271. The Loc-RIB contains the routes that have been selected by the local BGP speaker's Decision Process.

Status of This Memo

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

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

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

This Internet-Draft will expire on 4 March 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. Alternative Method to Monitor Loc-RIB	4
2. Terminology	6
3. Definitions	6
4. Per-Peer Header	7
4.1. Peer Type	7
4.2. Peer Flags	7
5. Loc-RIB Monitoring	8
5.1. Per-Peer Header	8
5.2. Peer Up Notification	9
5.2.1. Peer Up Information	9
5.3. Peer Down Notification	10
5.4. Route Monitoring	10
5.4.1. ASN Encoding	10
5.4.2. Granularity	10
5.5. Route Mirroring	11
5.6. Statistics Report	11
6. Other Considerations	11
6.1. Loc-RIB Implementation	11
6.1.1. Multiple Loc-RIB Peers	11
6.1.2. Filtering Loc-RIB to BMP Receivers	12
6.1.3. Changes to existing BMP sessions	12
7. Security Considerations	12
8. IANA Considerations	12
8.1. BMP Peer Type	12
8.2. BMP Loc-RIB Instance Peer Flags	12
8.3. Peer Up Information TLV	13
8.4. Peer Down Reason code	13
8.5. Deprecated entries	13
9. Normative References	13
10. Informative References	14
Acknowledgements	14
Authors' Addresses	14

1. Introduction

This document defines a mechanism to monitor the BGP Loc-RIB state of remote BGP instances without the need to establish BGP peering sessions. BMP [RFC7854] does not define a method to send the BGP instance Loc-RIB. It does define in section 8.2 of [RFC7854] locally originated routes, but these routes are defined as the routes originated into BGP. For example, as defined by Section 9.4 of [RFC4271]. Loc-RIB includes all selected received routes from BGP peers in addition to locally originated routes.

Figure 1 shows the flow of received routes from one or more BGP peers into the Loc-RIB.

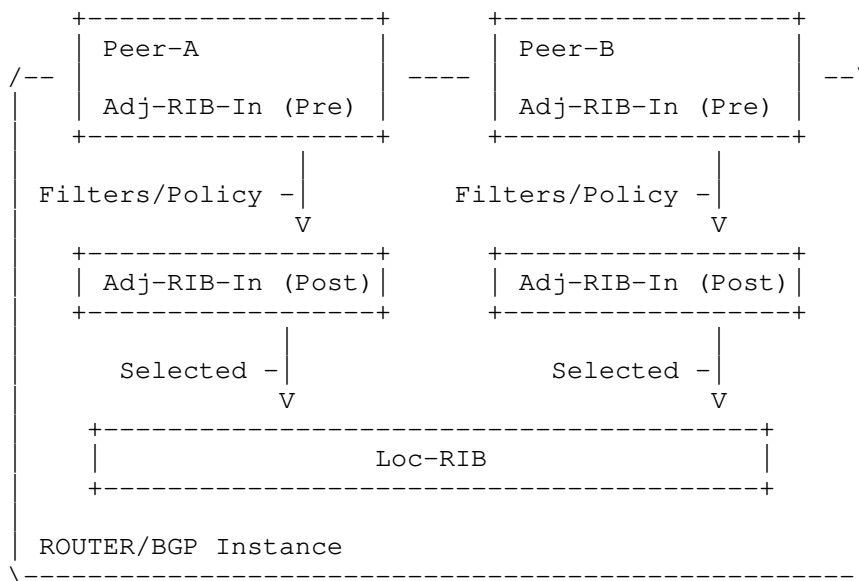


Figure 1: BGP peering Adj-RIBs-In into Loc-RIB

The following are some use-cases for Loc-RIB access:

- * The Adj-RIB-In for a given peer Post-Policy may contain hundreds of thousands of routes, with only a handful of routes selected and installed in the Loc-RIB after best-path selection. Some monitoring applications, such as ones that need only to correlate flow records to Loc-RIB entries, only need to collect and monitor the routes that are actually selected and used.

Requiring the applications to collect all Adj-RIB-In Post-Policy data forces the applications to receive a potentially large unwanted data set and to perform the BGP decision process selection, which includes having access to the interior gateway protocol (IGP) next-hop metrics. While it is possible to obtain the IGP topology information using BGP Link-State (BGP-LS), it requires the application to implement shortest path first (SPF) and possibly constrained shortest path first (CSPF) based on additional policies. This is overly complex for such a simple application that only needs to have access to the Loc-RIB.

- * It is common to see frequent changes over many BGP peers, but those changes do not always result in the router's Loc-RIB changing. The change in the Loc-RIB can have a direct impact on the forwarding state. It can greatly reduce time to troubleshoot and resolve issues if operators have the history of Loc-RIB changes. For example, a performance issue might have been seen for only a duration of 5 minutes. Post-facto troubleshooting this issue without Loc-RIB history hides any decision based routing changes that might have happened during those five minutes.
- * Operators may wish to validate the impact of policies applied to Adj-RIB-In by analyzing the final decision made by the router when installing into the Loc-RIB. For example, in order to validate if multi-path prefixes are installed as expected for all advertising peers, the Adj-RIB-In Post-Policy and Loc-RIB needs to be compared. This is only possible if the Loc-RIB is available. Monitoring the Adj-RIB-In for this router from another router to derive the Loc-RIB is likely to not show same installed prefixes. For example, the received Adj-RIB-In will be different if ADD-PATH [RFC7911] is not enabled or if maximum supported number of equal paths is different between Loc-RIB and advertised routes.

This document adds Loc-RIB to the BGP Monitoring Protocol and replaces Section 8.2 of [RFC7854] Locally Originated Routes.

1.1. Alternative Method to Monitor Loc-RIB

Loc-RIB is used to build Adj-RIB-Out when advertising routes to a peer. It is therefore possible to derive the Loc-RIB of a router by monitoring the Adj-RIB-In Pre-Policy from another router. This becomes overly complex and error prone when considering the number of peers being monitored per router.

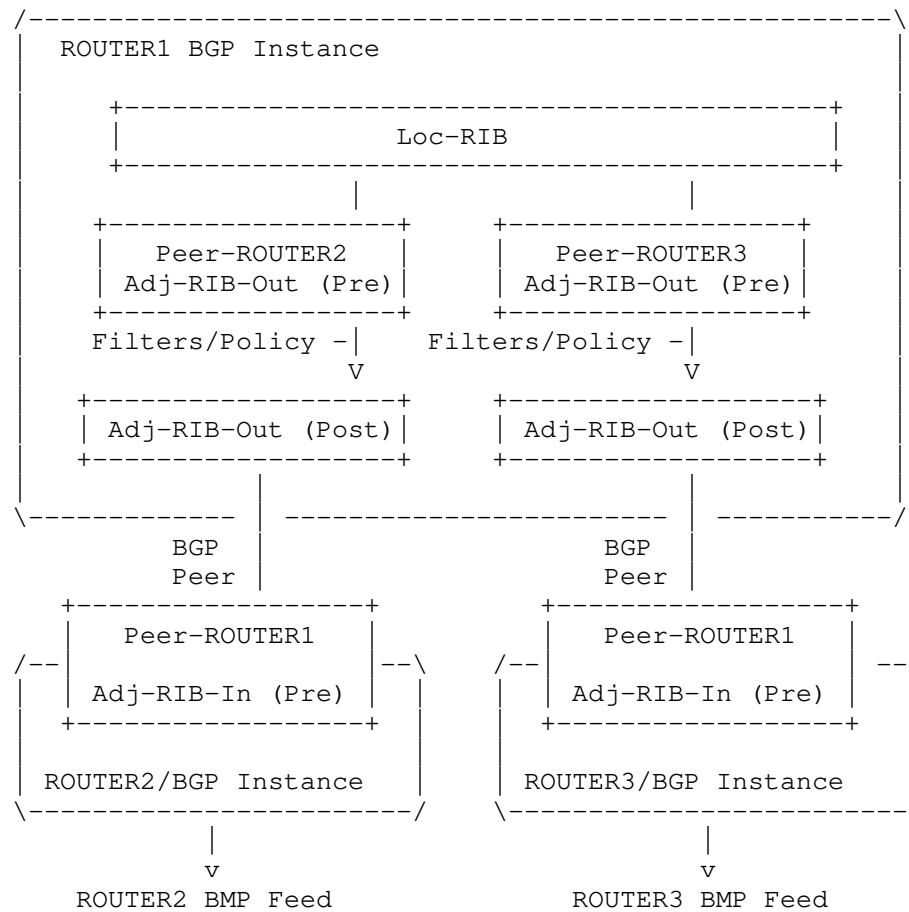


Figure 2: Alternative method to monitor Loc-RIB

The setup needed to monitor the Loc-RIB of a router requires another router with a peering session to the target router that is to be monitored. As shown in Figure 2, the target router Loc-RIB is advertised via Adj-RIB-Out to the BMP router over a standard BGP peering session. The BMP router then forwards Adj-RIB-In Pre-Policy to the BMP receiver.

BMP lacking access to Loc-RIB introduces the need for additional resources:

- * Requires at least two routers when only one router was to be monitored.

- * Requires additional BGP peering to collect the received updates when peering may have not even been required in the first place. For example, virtual routing and forwarding (VRF) tables with no peers, redistributed BGP-LS with no peers, and segment routing egress peer engineering where no peers have link-state address family enabled are all situations with no preexisting BGP peers.

Many complexities are introduced when using a received Adj-RIB-In to infer a router Loc-RIB:

- * Adj-RIB-Out received as Adj-RIB-In from another router may have a policy applied that filters, generates aggregates, suppresses more specific prefixes, manipulates attributes, or filters routes. Not only does this invalidate the Loc-RIB view, it adds complexity when multiple BMP routers may have peering sessions to the same router. The BMP receiver user is left with the error-prone task of identifying which peering session is the best representative of the Loc-RIB.
- * BGP peering is designed to work between administrative domains and therefore does not need to include internal system level information of each peering router (e.g., the system name or version information). In order to derive the Loc-RIB of a router, the router name or other system information is needed. The BMP receiver and user are forced to do some type of correlation using what information is available in the peering session (e.g., peering addresses, autonomous system numbers, and BGP identifiers). This leads to error-prone correlations.
- * Correlating BGP identifiers (BGP-ID) and session addresses to a router requires additional data, such as router inventory. This additional data provides the BMP receiver the ability to map and correlate the BGP-IDs and/or session addresses, but requires the BMP receiver to somehow obtain this data outside of BMP. How this data is obtained and the accuracy of the data directly affects the integrity of the correlation.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Definitions

- * BGP Instance: refers to an instance of BGP-4 [RFC4271] and considerations in section 8.1 of [RFC7854] do apply to it.
- * Adj-RIB-In: As defined in [RFC4271], "The Adj-RIBs-In contains unprocessed routing information that has been advertised to the local BGP speaker by its peers." This is also referred to as the pre-policy Adj-RIB-In in this document.
- * Adj-RIB-Out: As defined in [RFC4271], "The Adj-RIBs-Out contains the routes for advertisement to specific peers by means of the local speaker's UPDATE messages."
- * Loc-RIB: As defined in section 9.4 of [RFC4271], "The Loc-RIB contains the routes that have been selected by the local BGP speaker's Decision Process." Note that the Loc-RIB state as monitored through BMP might also contain routes imported from other routing protocols such as an IGP, or local static routes.
- * Pre-Policy Adj-RIB-Out: The result before applying the outbound policy to an Adj-RIB-Out. This normally represents a similar view of the Loc-RIB but may contain additional routes based on BGP peering configuration.
- * Post-Policy Adj-RIB-Out: The result of applying outbound policy to an Adj-RIB-Out. This MUST be what is actually sent to the peer.

4. Per-Peer Header

4.1. Peer Type

A new peer type is defined for Loc-RIB to distinguish that it represents the router Loc-RIB, which may have a route distinguisher (RD). Section 4.2 of [RFC7854] defines a Local Instance Peer type, which is for the case of non-RD peers that have an instance identifier.

This document defines the following new peer type:

- * Peer Type = 3: Loc-RIB Instance Peer

4.2. Peer Flags

If locally sourced routes are communicated using BMP, they MUST be conveyed using the Loc-RIB instance peer type.

The per-peer header flags for Loc-RIB Instance Peer type are defined as follows:

```

      0 1 2 3 4 5 6 7
    +--+--+--+--+--+--+--+
    |F|  |  |  |  |  |  |
    +--+--+--+--+--+--+--+

```

- * The F flag indicates that the Loc-RIB is filtered. This MUST be set when a filter is applied to Loc-RIB routes sent to the BMP collector.

The unused bits are reserved for future use. They MUST be transmitted as 0 and their values MUST be ignored on receipt.

5. Loc-RIB Monitoring

The Loc-RIB contains all routes selected by the BGP Decision Process as described in section 9.1 of [RFC4271]. These routes include those learned from BGP peers via its Adj-RIBs-In Post-Policy, as well as routes learned by other means as per section 9.4 of [RFC4271]. Examples of these include redistribution of routes from other protocols into BGP or otherwise locally originated (i.e., aggregate routes).

As described in Section 6.1.2, a subset of Loc-RIB routes MAY be sent to a BMP collector by setting the F flag.

5.1. Per-Peer Header

All peer messages that include a per-peer header as defined in section 4.2 of [RFC7854] MUST use the following values:

- * Peer Type: Set to 3 to indicate Loc-RIB Instance Peer.
- * Peer Distinguisher: Zero filled if the Loc-RIB represents the global instance. Otherwise set to the route distinguisher or unique locally defined value of the particular instance the Loc-RIB belongs to.
- * Peer Address: Zero-filled. Remote peer address is not applicable. The V flag is not applicable with Loc-RIB Instance peer type considering addresses are zero-filled.
- * Peer AS: Set to the primary router BGP autonomous system number (ASN).
- * Peer BGP ID: Set to the BGP instance global or RD (e.g., VRF) specific router-id section 1.1 of [RFC7854].

- * **Timestamp:** The time when the encapsulated routes were installed in the Loc-RIB, expressed in seconds and microseconds since midnight (zero hour), January 1, 1970 (UTC). If zero, the time is unavailable. Precision of the timestamp is implementation-dependent.

5.2. Peer Up Notification

Peer Up notifications follow section 4.10 of [RFC7854] with the following clarifications:

- * **Local Address:** Zero-filled, local address is not applicable.
- * **Local Port:** Set to 0, local port is not applicable.
- * **Remote Port:** Set to 0, remote port is not applicable.
- * **Sent OPEN Message:** This is a fabricated BGP OPEN message. Capabilities **MUST** include the 4-octet ASN and all necessary capabilities to represent the Loc-RIB route monitoring messages. Only include capabilities if they will be used for Loc-RIB monitoring messages. For example, if ADD-PATH is enabled for IPv6 and Loc-RIB contains additional paths, the ADD-PATH capability should be included for IPv6. In the case of ADD-PATH, the capability intent of advertise, receive or both can be ignored since the presence of the capability indicates enough that add-paths will be used for IPv6.
- * **Received OPEN Message:** Repeat of the same Sent Open Message. The duplication allows the BMP receiver to parse the expected received OPEN message as defined in section 4.10 of [RFC7854].

5.2.1. Peer Up Information

The following Peer Up information TLV type is added:

- * **Type = 3: VRF/Table Name.** The Information field contains a UTF-8 string whose value **MUST** be equal to the value of the VRF or table name (e.g., RD instance name) being conveyed. The string size **MUST** be within the range of 1 to 255 bytes.

The VRF/Table Name TLV is optionally included to support implementations that may not have defined a name. If a name is configured, it **MUST** be included. The default value of "global" **MUST** be used for the default Loc-RIB instance with a zero-filled distinguisher. If the TLV is included, then it **MUST** also be included in the Peer Down notification.

Multiple TLVs of the same type can be repeated as part of the same message, for example to convey a filtered view of a VRF. A BMP receiver should append multiple TLVs of the same type to a set in order to support alternate or additional names for the same peer. If multiple strings are included, their ordering MUST be preserved when they are reported.

5.3. Peer Down Notification

Peer Down notification MUST use reason code 6. Following the reason is data in TLV format. The following Peer Down information TLV type is defined:

- * Type = 3: VRF/Table Name. The Information field contains a UTF-8 string whose value MUST be equal to the value of the VRF or table name (e.g., RD instance name) being conveyed. The string size MUST be within the range of 1 to 255 bytes. The VRF/Table Name informational TLV MUST be included if it was in the Peer Up.

5.4. Route Monitoring

Route Monitoring messages are used for initial synchronization of the Loc-RIB. They are also used to convey incremental Loc-RIB changes.

As defined in section 4.6 of [RFC7854], "Following the common BMP header and per-peer header is a BGP Update PDU."

5.4.1. ASN Encoding

Loc-RIB route monitor messages MUST use 4-byte ASN encoding as indicated in Peer Up sent OPEN message (Section 5.2) capability.

5.4.2. Granularity

State compression and throttling SHOULD be used by a BMP sender to reduce the amount of route monitoring messages that are transmitted to BMP receivers. With state compression, only the final resultant updates are sent.

For example, prefix 192.0.2.0/24 is updated in the Loc-RIB 5 times within 1 second. State compression of BMP route monitor messages results in only the final change being transmitted. The other 4 changes are suppressed because they fall within the compression interval. If no compression was being used, all 5 updates would have been transmitted.

A BMP receiver should expect that Loc-RIB route monitoring granularity can be different by BMP sender implementation.

5.5. Route Mirroring

Section 4.7 of [RFC7854], defines Route Mirroring for verbatim duplication of messages received. This is not applicable to Loc-RIB as PDUs are originated by the router. Any received Route Mirroring messages SHOULD be ignored.

5.6. Statistics Report

Not all Stat Types are relevant to Loc-RIB. The Stat Types that are relevant are listed below:

- * Stat Type = 8: (64-bit Gauge) Number of routes in Loc-RIB.
- * Stat Type = 10: Number of routes in per-AFI/SAFI Loc-RIB. The value is structured as: 2-byte AFI, 1-byte SAFI, followed by a 64-bit Gauge.

6. Other Considerations

6.1. Loc-RIB Implementation

There are several methods for a BGP speaker to implement Loc-RIB efficiently. In all methods, the implementation emulates a peer with Peer Up and Down messages to convey capabilities as well as Route Monitor messages to convey Loc-RIB. In this sense, the peer that conveys the Loc-RIB is a locally emulated peer.

6.1.1. Multiple Loc-RIB Peers

There MUST be at least one emulated peer for each Loc-RIB instance, such as with VRFs. The BMP receiver identifies the Loc-RIB by the peer header distinguisher and BGP ID. The BMP receiver uses the VRF/ Table Name from the Peer Up information to associate a name to the Loc-RIB.

In some implementations, it might be required to have more than one emulated peer for Loc-RIB to convey different address families for the same Loc-RIB. In this case, the peer distinguisher and BGP ID should be the same since they represent the same Loc-RIB instance. Each emulated peer instance MUST send a Peer Up with the OPEN message indicating the address family capabilities. A BMP receiver MUST process these capabilities to know which peer belongs to which address family.

6.1.2. Filtering Loc-RIB to BMP Receivers

There may be use-cases where BMP receivers should only receive specific routes from Loc-RIB. For example, IPv4 unicast routes may include internal BGP (IBGP), external BGP (EBGP), and IGP but only routes from EBGP should be sent to the BMP receiver. Alternatively, it may be that only IBGP and EBGP that should be sent and IGP redistributed routes should be excluded. In these cases where the Loc-RIB is filtered, the F flag is set to 1 to indicate to the BMP receiver that the Loc-RIB is filtered. If multiple filters are associated to the same Loc-RIB, a Table Name MUST be used in order to allow a BMP receiver to make the right associations.

6.1.3. Changes to existing BMP sessions

In case of any change that results in the alteration of behavior of an existing BMP session, ie. changes to filtering and table names, the session MUST be bounced with a Peer Down/Peer Up sequence.

7. Security Considerations

The same considerations as in section 11 of [RFC7854] apply to this document. Implementations of this protocol SHOULD require that sessions are only established with authorized and trusted monitoring devices. It is also believed that this document does not add any additional security considerations.

8. IANA Considerations

This document requests that IANA assign the following new parameters to the BMP parameters name space (<https://www.iana.org/assignments/bmp-parameters/bmp-parameters.xhtml>).

8.1. BMP Peer Type

This document defines a new peer type (Section 4.1):

* Peer Type = 3: Loc-RIB Instance Peer

8.2. BMP Loc-RIB Instance Peer Flags

This document requests IANA to rename "BMP Peer Flags" to "BMP Peer Flags for Peer Types 0 through 2" and create a new registry named "BMP Peer Flags for Loc-RIB Instance Peer Type 3." This document defines that peer flags are specific to the Loc-RIB instance peer type. As defined in (Section 4.2):

- * Flag 0: The F flag indicates that the Loc-RIB is filtered. This indicates that the Loc-RIB does not represent the complete routing table.

Flags 0 through 3 and 5 through 7 are unassigned. The registration procedure for the registry is "Standards Action".

8.3. Peer Up Information TLV

This document requests that IANA rename "BMP Initiation Message TLVs" registry to "BMP Initiation and Peer Up Information TLVs." section 4.4 of [RFC7854] defines that both Initiation and Peer Up share the same information TLVs. This document defines the following new BMP Peer Up information TLV type (Section 5.2.1):

- * Type = 3: VRF/Table Name. The Information field contains a UTF-8 string whose value MUST be equal to the value of the VRF or table name (e.g., RD instance name) being conveyed. The string size MUST be within the range of 1 to 255 bytes.

8.4. Peer Down Reason code

This document defines the following new BMP Peer Down reason code (Section 5.3):

- * Type = 6: Local system closed, TLV data follows.

8.5. Deprecated entries

This document also requests that IANA marks as "deprecated" the F Flag entry in the "BMP Peer Flags for Peer Types 0 through 2" registry.

9. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<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>>.

- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <<https://www.rfc-editor.org/info/rfc5226>>.
- [RFC7854] Scudder, J., Ed., Fernando, R., and S. Stuart, "BGP Monitoring Protocol (BMP)", RFC 7854, DOI 10.17487/RFC7854, June 2016, <<https://www.rfc-editor.org/info/rfc7854>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

10. Informative References

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

Acknowledgements

The authors would like to thank John Scudder, Jeff Haas and Mukul Srivastava for their valuable input.

Authors' Addresses

Tim Evens
Cisco Systems
2901 Third Avenue, Suite 600
Seattle, WA 98121
United States of America

Email: tievens@cisco.com

Serpil Bayraktar
Cisco Systems
3700 Cisco Way
San Jose, CA 95134
United States of America

Email: serpil@cisco.com

Manish Bhardwaj
Cisco Systems
3700 Cisco Way
San Jose, CA 95134
United States of America

Email: manbhard@cisco.com

Paolo Lucente
NTT Communications
Siriusdreef 70-72
2132 Hoofddorp
Netherlands

Email: paolo@ntt.net

Network Working Group
Internet-Draft
Updates: 7854 (if approved)
Intended status: Standards Track
Expires: December 5, 2019

J. Scudder
Juniper Networks
June 3, 2019

Revision to Registration Procedures for Multiple BMP Registries
draft-ietf-grow-bmp-registries-change-01.txt

Abstract

This document updates RFC 7854, BGP Monitoring Protocol (BMP) by making a change to the registration procedures for several registries. Specifically, any BMP registry with a range of 32768-65530 designated "Specification Required" has that range re-designated as "First Come First Served".

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 December 5, 2019.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	2
2. IANA Considerations	2
3. Security Considerations	3
4. Acknowledgements	3
5. Normative References	3
Author's Address	3

1. Introduction

[RFC7854] creates a number of IANA registries that include a range of 32768-65530 designated "Specification Required". Each such registry also has a large range designated "Standards Action". Subsequent experience has shown two things. First, there is less difference between these two policies in practice than there is in theory (consider that [RFC8126] explains that for Specification Required, "Publication of an RFC is an ideal means of achieving this requirement"). Second, it's desirable to have a very low bar to registration, to avoid the risk of conflicts introduced by use of unregistered code points (so-called "code point squatting").

Accordingly, this document revises the registration procedures, as given in Section 2.

2. IANA Considerations

IANA is requested to revise the following registries within the BMP group:

- o BMP Statistics Types
- o BMP Initiation Message TLVs
- o BMP Termination Message TLVs
- o BMP Termination Message Reason Codes
- o BMP Peer Down Reason Codes
- o BMP Route Mirroring TLVs
- o BMP Route Mirroring Information Codes

For each of these registries, the ranges 32768-65530 whose registration procedures were "Specification Required" are revised to have the registration procedures "First Come First Served".

3. Security Considerations

This revision to registration procedures does not change the underlying security issues inherent in the existing [RFC7854].

4. Acknowledgements

Thanks to Jeff Haas for review and encouragement.

5. Normative References

- [RFC7854] Scudder, J., Ed., Fernando, R., and S. Stuart, "BGP Monitoring Protocol (BMP)", RFC 7854, DOI 10.17487/RFC7854, June 2016, <<https://www.rfc-editor.org/info/rfc7854>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.

Author's Address

John Scudder
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: jgs@juniper.net

Network Working Group
Internet-Draft
Updates: 1997 (if approved)
Intended status: Standards Track
Expires: December 15, 2019

J. Borkenhagen
AT&T
R. Bush
IIJ & Arrcus
R. Bonica
Juniper Networks
S. Bayraktar
Cisco Systems
June 13, 2019

Well-Known Community Policy Behavior
draft-ietf-grow-wkc-behavior-08

Abstract

Well-Known BGP Communities are manipulated differently across various current implementations; resulting in difficulties for operators. Network operators should deploy consistent community handling across their networks while taking the inconsistent behaviors from the various BGP implementations into consideration.. This document recommends specific actions to limit future inconsistency, namely BGP implementors must not create further inconsistencies from this point forward.

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 December 15, 2019.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
2. Manipulation of Communities by Policy	3
3. Community Manipulation Policy Differences	3
4. Documentation of Vendor Implementations	3
4.1. Note on an Inconsistency	4
5. Note for Those Writing RFCs for New Community-Like Attributes	5
6. Action Items	5
7. Security Considerations	5
8. IANA Considerations	5
9. Acknowledgments	6
10. Normative References	6
Authors' Addresses	6

1. Introduction

The BGP Communities Attribute was specified in [RFC1997] which introduced the concept of Well-Known Communities. In hindsight, [RFC1997] did not prescribe as fully as it should have how Well-Known Communities may be manipulated by policies applied by operators. Currently, implementations differ in this regard, and these differences can result in inconsistent behaviors that operators find difficult to identify and resolve.

This document describes the current behavioral differences in order to assist operators in generating consistent community-manipulation policies in a multi-vendor environment, and to prevent the introduction of additional divergence in implementations.

This document recommends specific actions to limit future inconsistency, namely BGP implementors MUST NOT create further inconsistencies from this point forward.

2. Manipulation of Communities by Policy

[RFC1997] says:

"A BGP speaker receiving a route with the COMMUNITIES path attribute may modify this attribute according to the local policy."

One basic operational need is to add or remove one or more communities to the set. The focus of this document is another common operational need, to replace all communities with a new set. To simplify this second case, most BGP policy implementations provide syntax to "set" community that operators use to mean "remove any/all communities present on the route, and apply this set of communities instead."

Some operators prefer to write explicit policy to delete unwanted communities rather than using "set;" i.e. using a "delete community *:*" and then "add community x:y ..." configuration statements in an attempt to replace all communities. The same community manipulation policy differences described in the following section exist in both "set" and "delete community *:*" syntax. For simplicity, the remainder of this document refers only to the "set" behaviors, which we refer to collectively as each implementation's "set" directive.'

3. Community Manipulation Policy Differences

Vendor implementations differ in the treatment of certain Well-Known communities when modified using the syntax to "set" the community. Some replace all communities including the Well-Known ones with the new set, while others replace all non-Well-Known Communities but do not modify any Well-Known Communities that are present.

These differences result in what would appear to be identical policy configurations having very different results on different platforms.

4. Documentation of Vendor Implementations

In this section we document the syntax and observed behavior of the "set" directive in several popular BGP implementations to illustrate the severity of the problem operators face.

In Juniper Networks' Junos OS, "community set" removes all communities, Well-Known or otherwise.

In Cisco IOS XR, "set community" removes all communities except for the following:

Numeric	Common Name
0:0	internet
65535:0	graceful-shutdown
65535:1	accept-own rfc7611
65535:65281	NO_EXPORT
65535:65282	NO_ADVERTISE
65535:65283	NO_EXPORT_SUBCONFED (or local-AS)

Communities not removed by Cisco IOS XR

Table 1

Cisco IOS XR does allow Well-Known communities to be removed only by explicitly enumerating one at a time, not in the aggregate; for example, "delete community accept-own". Operators are advised to consult Cisco IOS XR documentation and/or Cisco support for full details.

On Extreme networks' Brocade NetIron: "set community X" removes all communities and sets X.

In Huawei's VRP product, "community set" removes all communities, Well-Known or otherwise.

In OpenBGPD, "set community" does not remove any communities, Well-Known or otherwise.

Nokia's SR OS has several directives that operate on communities. Its "set" directive is called using the "replace" keyword, replacing all communities, Well-Known or otherwise, with the specified communities.

4.1. Note on an Inconsistency

The IANA publishes a list of Well-Known Communities [IANA-WKC].

Cisco IOS XR's set of Well-Known communities that "set community" will not overwrite diverges from the IANA's list of Well-Known communities. Quite a few Well-Known communities from IANA's list do not receive special treatment in Cisco IOS XR, and at least one community on Cisco IOS XR's special treatment list, internet == 0:0,

is not formally a Well-Known Community as it is not in [IANA-WKC]; but taken from the Reserved range [0x00000000-0x0000FFFF].

This merely notes an inconsistency. It is not a plea to 'protect' the entire IANA list from "set community."

5. Note for Those Writing RFCs for New Community-Like Attributes

When establishing new [RFC1997]-like attributes (large communities, wide communities, etc.), RFC authors should state explicitly how the new attribute is to be handled.

6. Action Items

Network operators are encouraged to limit their use of the "set" directive (within reason), to improve consistency across platforms.

Unfortunately, it would be operationally disruptive for vendors to change their current implementations.

Vendors MUST clearly document the behavior of "set" directive in their implementations.

Vendors MUST ensure that their implementations' "set" directive treatment of any specific community does not change if/when that community becomes a new Well-Known Community through future standardization. For most implementations, this means that the "set" directive MUST continue to remove the community; for those implementations where the "set" directive removes no communities, that behavior MUST continue.

Given the implementation inconsistencies described in this document, network operators are urged never to rely on any implicit understanding of a neighbor ASN's BGP community handling. I.e., before announcing prefixes with NO_EXPORT or any other community to a neighbor ASN, the operator should confirm with that neighbor how the community will be treated.

7. Security Considerations

Surprising defaults and/or undocumented behaviors are not good for security. This document attempts to remedy that.

8. IANA Considerations

The IANA is requested to list this document as an additional reference for the [IANA-WKC] registry.

9. Acknowledgments

The authors thank Martijn Schmidt, Qin Wu for the Huawei data point, Greg Hankins, Job Snijders, David Farmer, John Heasley, and Jakob Heitz.

10. Normative References

- [IANA-WKC] IANA, "Border Gateway Protocol (BGP) Well-Known Communities", <<https://www.iana.org/assignments/bgp-well-known-communities>>.
- [RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <<http://www.rfc-editor.org/info/rfc1997>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<http://www.rfc-editor.org/info/rfc8174>>.

Authors' Addresses

Jay Borkenhagen
AT&T
200 Laurel Avenue South
Middletown, NJ 07748
United States of America

Email: jayb@att.com

Randy Bush
IIJ & Arrcus
5147 Crystal Springs
Bainbridge Island, WA 98110
US

Email: randy@psg.com

Ron Bonica
Juniper Networks
2251 Corporate Park Drive
Herndon, VA 20171
US

Email: rbonica@juniper.net

Serpil Bayraktar
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America

Email: serpil@cisco.com

Global Routing Operations
Internet-Draft
Intended status: Standards Track
Expires: September 9, 2019

J. Snijders
NTT Communications
M. Aelmans
Juniper Networks
March 8, 2019

BGP Maximum Prefix Limits
draft-sa-grow-maxprefix-02

Abstract

This document describes mechanisms to limit the negative impact of route leaks [RFC7908] and/or resource exhaustion in BGP [RFC4271] implementations.

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 September 9, 2019.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

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

Table of Contents

1. Introduction	2
2. Inbound Maximum Prefix Limits	2
2.1. Type A: Pre-Policy Inbound Maximum Prefix Limits	3
2.2. Type B: Post-Policy Inbound Maximum Prefix Limits	3
3. Outbound Maximum Prefix Limits	3
4. Considerations for Operations with Multi-Protocol BGP	4
5. Considerations for soft thresholds	4
6. Security Considerations	4
7. IANA Considerations	4
8. Acknowledgments	5
9. Implementation status - RFC EDITOR: REMOVE BEFORE PUBLICATION	5
10. Appendix: Implementation Guidance	6
11. References	7
11.1. Normative References	7
11.2. Informative References	7
Authors' Addresses	7

1. Introduction

This document describes mechanisms to reduce the negative impact of certain types of misconfigurations and/or resource exhaustions in BGP [RFC4271] operations. While [RFC4271] already described a method to tear down BGP sessions when certain thresholds are exceeded, some nuances in this specification were missing resulting in inconsistencies between BGP implementations. In addition to clarifying "inbound maximum prefix limits", this document also introduces a specification for "outbound maximum prefix limits".

2. Inbound Maximum Prefix Limits

An operator MAY configure a BGP speaker to terminate its BGP session with a neighbor when the number of address prefixes received from that neighbor exceeds a locally configured upper limit. The BGP speaker then MUST send the neighbor a NOTIFICATION message with the Error Code Cease and the Error Subcode "Threshold reached: Maximum Number of Prefixes Received", and MAY support other actions. Reporting when thresholds have been exceeded is an implementation specific consideration, but SHOULD include methods such as Syslog

[RFC5424]. Inbound Maximum Prefix Limits can be applied in two distinct places in the conceptual model: before or after the application of routing policy.

2.1. Type A: Pre-Policy Inbound Maximum Prefix Limits

The Adj-RIBs-In stores routing information learned from inbound UPDATE messages that were received from another BGP speaker Section 3.2 [RFC4271]. The Type A pre-policy limit uses the number of NLRIs per Address Family Identifier (AFI) per Subsequent Address Family Identifier (SAFI) as input into its threshold comparisons. For example, when an operator configures the Type A pre-policy limit for IPv4 Unicast to be 50 on a given EBGp session, and the other BGP speaker announces its 51st IPv4 Unicast NLRI, the session MUST be terminated.

Type A pre-policy limits are particularly useful to help dampen the effects of full table route leaks and memory exhaustion when the implementation stores rejected routes.

2.2. Type B: Post-Policy Inbound Maximum Prefix Limits

RFC4271 describes a Policy Information Base (PIB) that contains local policies that can be applied to the information in the Routing Information Base (RIB). The Type B post-policy limit uses the number of NLRIs per Address Family Identifier (AFI) per Subsequent Address Family Identifier (SAFI), after application of the Import Policy as input into its threshold comparisons. For example, when an operator configures the Type B post-policy limit for IPv4 Unicast to be 50 on a given EBGp session, and the other BGP speaker announces a hundred IPv4 Unicast routes of which none are accepted as a result of the local import policy (and thus not considered for the Loc-RIB by the local BGP speaker), the session is not terminated.

Type B post-policy limits are useful to help prevent FIB exhaustion and prevent accidental BGP session teardown due to prefixes not accepted by policy anyway.

3. Outbound Maximum Prefix Limits

An operator MAY configure a BGP speaker to terminate its BGP session with a neighbor when the number of address prefixes to be advertised to that neighbor exceeds a locally configured upper limit. The BGP speaker then MUST send the neighbor a NOTIFICATION message with the Error Code Cease and the Error Subcode "Threshold reached: Maximum Number of Prefixes Send", and MAY support other actions. Reporting when thresholds have been exceeded is an implementation specific

consideration, but SHOULD include methods such as Syslog [RFC5424]. By definition, Outbound Maximum Prefix Limits are Post-Policy.

The Adj-RIBs-Out stores information selected by the local BGP speaker for advertisement to its neighbors. The routing information stored in the Adj-RIBs-Out will be carried in the local BGP speaker's UPDATE messages and advertised to its neighbors Section 3.2 [RFC4271]. The Outbound Maximum Prefix Limit uses the number of NLRI's per Address Family Identifier (AFI) per Subsequent Address Family Identifier (SAFI), after application of the Export Policy, as input into its threshold comparisons. For example, when an operator configures the Outbound Maximum Prefix Limit for IPv4 Unicast to be 50 on a given EBGP session, and were about to announce its 51st IPv4 Unicast NLRI to the other BGP speaker as a result of the local export policy, the session MUST be terminated.

Outbound Maximum Prefix Limits are useful to help dampen the negative effects of a misconfiguration in local policy. In many cases, it would be more desirable to tear down a BGP session rather than causing or propagating a route leak.

4. Considerations for Operations with Multi-Protocol BGP

5. Considerations for soft thresholds

describe soft and hard limits (warning vs teardown)

6. Security Considerations

Maximum Prefix Limits are an essential tool for routing operations and SHOULD be used to increase stability.

7. IANA Considerations

This memo requests that IANA updates the name of subcode "Maximum Number of Prefixes Reached" to "Threshold exceeded: Maximum Number of Prefixes Received" in the "Cease NOTIFICATION message subcodes" registry under the "Border Gateway Protocol (BGP) Parameters" group.

This memo requests that IANA assigns a new subcode named "Threshold exceeded: Maximum Number of Prefixes Send" in the "Cease NOTIFICATION message subcodes" registry under the "Border Gateway Protocol (BGP) Parameters" group.

8. Acknowledgments

The authors would like to thank Saku Ytti and John Heasley (NTT Communications), Jeff Haas, Colby Barth and John Scudder (Juniper Networks), Martijn Schmidt (i3D.net), Teun Vink (BIT), Sabri Berisha (eBay), Martin Pels (Quanza), Steven Bakker (AMS-IX), Aftab Siddiqui (ISOC) and Yu Tianpeng for their support, insightful review, and comments.

9. Implementation status - RFC EDITOR: REMOVE BEFORE PUBLICATION

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

The below table provides an overview (as of the moment of writing) of which vendors have produced implementation of inbound or outbound maximum prefix limits. Each table cell shows the applicable configuration keywords if the vendor implemented the feature.

Vendor	Type A Pre-Policy	Type B Post-Policy	Outbound
Cisco IOS XR		maximum-prefix	
Cisco IOS XE		maximum-prefix	
Juniper Junos OS	prefix-limit	accepted-prefix-limit, or prefix-limit combined with 'keep none'	
Nokia SR OS	prefix-limit		
NIC.CZ BIRD	'import keep filtered' combined with 'receive limit'	'import limit' or 'receive limit'	export limit
OpenBSD OpenBGPD	max-prefix		
Arista EOS	maximum-routes	maximum-accepted-routes	
Huawei VRPv5	peer route-limit		
Huawei VRPv8	peer route-limit	peer route-limit accept-prefix	

First presented by Snijders at [RIPE77]

Table 1: Maximum prefix limits capabilities per implementation

10. Appendix: Implementation Guidance

1) make it clear who does what: if A sends too many prefixes to B A should see "ABC" in log B should see "DEF" in log to make it clear which of the two parties does what 2) recommended by default automatically restart after between 15 and 30 minutes

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

11.2. Informative References

- [RFC5424] Gerhards, R., "The Syslog Protocol", RFC 5424, DOI 10.17487/RFC5424, March 2009, <<https://www.rfc-editor.org/info/rfc5424>>.
- [RFC7908] Sriram, K., Montgomery, D., McPherson, D., Osterweil, E., and B. Dickson, "Problem Definition and Classification of BGP Route Leaks", RFC 7908, DOI 10.17487/RFC7908, June 2016, <<https://www.rfc-editor.org/info/rfc7908>>.
- [RIPE77] Snijders, J., "Robust Routing Policy Architecture", May 2018, <https://ripe77.ripe.net/wp-content/uploads/presentations/59-RIPE77_Snijders_Routing_Policy_Architecture.pdf>.

Authors' Addresses

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

Email: job@ntt.net

Melchior Aelmans
Juniper Networks
Boeing Avenue 240
Schiphol-Rijk 1119 PZ
The Netherlands

Email: maelmans@juniper.net

GROW
Internet-Draft
Updates: 7854 (if approved)
Intended status: Standards Track
Expires: June 17, 2019

J. Scudder
Juniper Networks
December 14, 2018

BMP Peer Up Message Namespace
draft-scudder-grow-bmp-peer-up-00.txt

Abstract

RFC 7854, BMP, uses different message types for different purposes. Most of these are Type, Length, Value (TLV) structured. One message type, the Peer Up message, lacks a set of TLVs defined for its use, instead sharing a namespace with the Initiation message. Subsequent experience has shown that this namespace sharing was a mistake, as it hampers the extension of the protocol.

This document updates RFC 7854 by creating an independent namespace for the Peer Up message. The changes in this document are formal only, compliant implementations of RFC 7854 also comply with this specification.

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 June 17, 2019.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

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

Table of Contents

1. Introduction	2
1.1. Requirements Language	2
2. String Definition	3
3. Changes to RFC 7854	3
3.1. Revision to Information TLV, Renamed as Initiation Information TLV	3
3.2. Revision to Peer Up Notification	3
3.3. Definition of Peer Up Information TLV	4
4. IANA Considerations	4
5. Security Considerations	5
6. Acknowledgements	5
7. Normative References	5
Author's Address	5

1. Introduction

[RFC7854] defines a number of different BMP message types. With the exception of the Route Monitoring message type, these messages are TLV-structured. Most message types have distinct namespaces and IANA registries. However, the namespace of the Peer Up message overlaps that of the Initiation message. As the BMP protocol has been extended, this oversight has become problematic. In this document, we create a distinct namespace for the Peer Up message to eliminate this overlap, and create the corresponding missing registry.

The changes in this document are formal only, compliant implementations of [RFC7854] also comply with this specification.

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. String Definition

A string TLV is a free-form sequence of UTF-8 characters whose length is given by the TLV's Length field. There is no requirement to terminate the string with a null (or any other particular) character -- the Length field gives its termination.

3. Changes to RFC 7854

We update [RFC7854] as follows:

- o The "Information TLV" of section 4.4, that was shared between the Initiation and Peer Up message types, is renamed as the "Initiation Information TLV", and is only relevant to the Initiation message type.
- o A "Peer Up Information TLV" is defined, and is relevant to the Peer Up message type.
- o A "Peer Up TLVs" registry is created, seeded with the Peer Up Information TLV.

Other than as summarized above, and detailed below, there are no other changes.

3.1. Revision to Information TLV, Renamed as Initiation Information TLV

The Information TLV defined in section 4.4 of [RFC7854] is renamed "Initiation Information TLV". It is used only by the Initiation message, not by the Peer Up message.

The definition of Type = 0 is revised to be:

- o Type = 0: String. The Information field contains a string (Section 2). The value is administratively assigned. If multiple strings are included, their ordering MUST be preserved when they are reported.

3.2. Revision to Peer Up Notification

The final paragraph of section 4.10 of [RFC7854] references the Information TLV (which is revised above (Section 3.1)). That paragraph is replaced by the following:

- o Information: Information about the peer, using the Peer Up Information TLV format defined below (Section 3.3). The String type may be repeated. Inclusion of the Information field is

OPTIONAL. Its presence or absence can be inferred by inspection of the Message Length in the common header.

3.3. Definition of Peer Up Information TLV

The Peer Up Information TLV is used by the Peer Up message.

```

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

```

- o Information Type (2 bytes): Type of information provided. Defined types are:

- * Type = 0: String. The Information field contains a string (Section 2). The value is administratively assigned. If multiple strings are included, their ordering MUST be preserved when they are reported.

- o Information Length (2 bytes): The length of the following Information field, in bytes.
- o Information (variable): Information about the monitored router, according to the type.

4. IANA Considerations

IANA is requested to create a registry within the BMP group, named "BMP Peer Up Message TLVs", reference this document.

Registration procedures for this registry are:

Range	Registration Procedures
0-32767	Standards Action
32768-65530	First Come, First Served
65531-65534	Experimental
65535	Reserved

Initial values for this registry are:

Type	Description	Reference
0	String	this document
65535	Reserved	this document

5. Security Considerations

This rearrangement of deck chairs does not change the underlying security issues inherent in the existing [RFC7854].

6. Acknowledgements

TBD

7. 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>>.
- [RFC7854] Scudder, J., Ed., Fernando, R., and S. Stuart, "BGP Monitoring Protocol (BMP)", RFC 7854, DOI 10.17487/RFC7854, June 2016, <<https://www.rfc-editor.org/info/rfc7854>>.
- [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>>.

Author's Address

John Scudder
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: jgs@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 9, 2022

F. Xu
Tencent
T. Graf
Swisscom
Y. Gu
S. Zhuang
Z. Li
Huawei
March 8, 2022

BGP Route Policy and Attribute Trace Using BMP
draft-xu-grow-bmp-route-policy-attr-trace-06

Abstract

The generation of BGP adj-rib-in, local-rib or adj-rib-out comes from BGP route exchange and route policy processing. BGP Monitoring Protocol (BMP) provides the monitoring of BGP adj-rib-in [RFC7854], BGP local-rib [RFC9069] and BGP adj-rib-out [RFC8671]. By monitoring these BGP RIB's the full state of the network is visible, but how route-policies affect the route propagation or changes BGP attributes is still not. This document describes a method of using BMP to record the trace data on how BGP routes are (not) changed under the process of route 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 September 9, 2022.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
1.1. BGP Route Policy and Attribute Trace Overview	3
1.2. Use cases	3
2. Extension of BMP for Route Policy and Attribute Trace	4
2.1. Common Header	4
2.2. Per Peer Header	4
2.3. Route Policy and Attribute Trace Message	4
2.3.1. VRF/Table TLV	7
2.3.2. Policy TLV	8
2.3.3. Pre Policy Attribute TLV	12
2.3.4. Post Policy Attribute TLV	12
2.3.5. String TLV	13
3. Implementation Considerations	13
4. Acknowledgments	13
5. IANA Considerations	13
6. Security Considerations	14
7. Normative References	14
Authors' Addresses	15

1. Introduction

The typical processing procedure after receiving a BGP Update Message at a routing device is as follows: 1. Adding the pre-policy routes into the pre-policy adj-rib-in (if any); 2. Filtering the pre-policy routes through inbound route policies; 3. Selecting the BGP best routes from the post-policy routes; 4. Adding the selected routes into the BGP local-rib; 5-a. Adding the BGP best routes from local-rib to the core routing table manager for selection; 5-b. Filtering the routes from BGP local-rib through outbound route policies w.r.t. per peer or peer groups; 6. Sending the BGP adj-rib-out to the target peer or peer groups. Details may vary by vendors. The BGP

Monitoring Protocol (BMP) can be utilized to monitor BGP routes in forms of adj-rib-in, local-rib and adj-rib-out. However, the complete procedure from inbound to outbound policy processing, including other policies, e.g., route redistribution, route selection and so on, is currently unobserved. For example, there are 10 policy items (or nodes) configured under one outbound route policy per a specific peer. By collecting the local-rib and adj-rib-out through BMP, the operator finds that the outbound policy didn't work as expected. However, it's hard to distinguish which one of the 10 policy items/nodes is responsible for the failure.

1.1. BGP Route Policy and Attribute Trace Overview

This document describes a method that records and reports how each policy item/node processes the routes (e.g., changes the route attribute). Each policy item/node processing is called an event thereafter in this document. Compared with conventional BGP rib entry, which consists of prefix/mask, route attributes, e.g., next hop, MED, local preference, AS path, and so on, the event record discussed in this document includes extra information, such as event index, timestamp, policy information, and so on. For example, if a route is processed by 5 policy items/nodes, there can be 5 event records for the same prefix/mask. Each event is numbered in order of time (e.g., the time of policy execution). The policy information includes the policy name and item/node ID/name so that the server/controller can map to the exact policy either directly from the device or from the configurations collected at the server side.

This document defines a new BMP message type to carry the recorded policy and route data. More detailed message format is defined in Section 2. The message is called the BMP Route Policy and Attribute Trace Message thereafter in this document.

1.2. Use cases

There are cases that a new policy is configured incorrectly, e.g., setting an incorrect community value, or policy placed in incorrect order among other policies. These may result in incorrect route attribute modification, best route selection mistake, or route distribution mistake. With the correlated record of policy and route, the server/controller is able to identify the unexpected route change and its responsible policy. Considering the fact that the BGP route policy impacts not only the route processing within the individual device but also the route distribution to its peers, the route trace data of a single device is always analyzed in correlation with such data collected from its peer devices.

Apart from the policy validation application, the route trace data can also be analyzed to discover the route propagation path within the network. With the route's inbound and outbound event records collect from each related device, the server is able to find the propagation path hop by hop. The identified path is helpful for operators to better understand its network, and thus benefiting both network troubleshooting and network planning.

2. Extension of BMP for Route Policy and Attribute Trace

2.1. Common Header

This document defines a new BMP message type to carry the Route Policy and Attribute Trace data.

- o Type = TBD: Route Policy and Attribute Trace Message

The new defined message type is indicated in the Message Type field of the BMP common header.

2.2. Per Peer Header

The Route Policy and Attribute Trace Message is not per peer based, thus it does not require the Per Peer Header.

2.3. Route Policy and Attribute Trace Message

The Route Policy and Attribute Trace Message format is defined as follows:

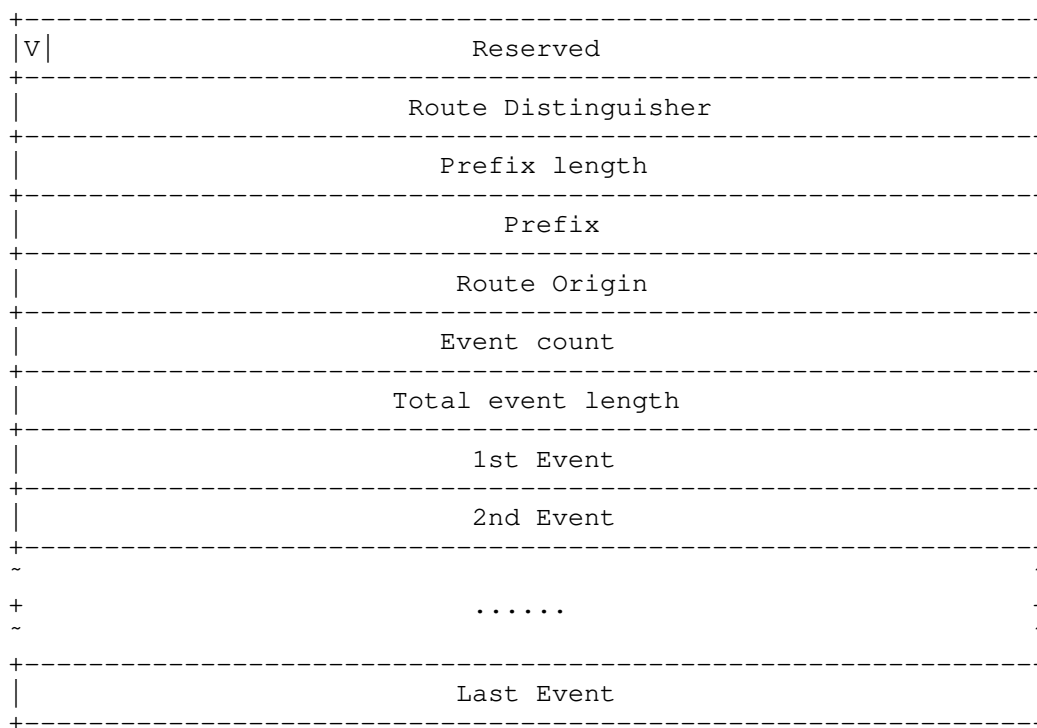


Figure 1: Route Policy and Attribute Trace Message format

- o Flags (1 Byte): The V flag indicates that the Peer address is an IPv6 address. For IPv4 peers, this is set to 0.
- o Route Distinguisher (8 Bytes): indicates the route distinguisher (RD) related to the route.
- o Prefix Length (1 Byte): indicates the length of the prefix.
- o Prefix (16 Bytes): indicates the monitored prefix, with mask defined by Prefix Length field. It is 4 bytes long if an IPv4 address is carried in this field (with the 12 most significant bytes zero-filled) and 16 bytes long if an IPv6 address is carried in this field.
- o Route Origin (4 Bytes): indicates the BGP router ID where this route is learned from. If the route is locally generated, this field is zero filled.
- o Event Count (1 Byte): indicates the total number of policy processing event recorded in this message.

- o Total event length (2 Byte): indicates the total length of the following fields including all events, where the total number is indicated by the Event Count field.
- o 1 ~ Last event: indicates each event, stacked one by one in order of time. The event format is further defined as follows.

Single event length
Event index
Timestamp(seconds)
Timestamp(microseconds)
Path Identifier
AFI
SAFI
VRF/Table TLV
Policy TLV
Pre Policy Attribute TLV
Post Policy Attribute TLV
String TLV

Figure 2: Event format

- o Single event length (2 Byte): indicates the total length of a single policy process event, including the following fields that belong to this event.
- o Event index (1 Byte): indicates the sequence number of this event, starting from 1 and increases by 1 for each event recorded in order.
- o Timestamp (8 Bytes): indicates the time when the policy of this event starts execution, expressed in seconds and microseconds since midnight (zero hour), January 1, 1970 (UTC).

- o Path Identifier (4 Bytes): used to distinguish multiple BGP paths for the same prefix. If there's no path ID, this field is zero filled.
- o AFI (2 Bytes)/SAFI (1 Byte): indicates the AFI/SAFI of the route.
- o VRF/Table TLV (Variable): indicates the VRF information of the route. The format of the VRF/Table TLV is further defined in Figure 3. The VRF/Table ID TLV is optional. At most one VRF/Table TLV can be included in each Route Policy and Attribute Trace Message.
- o Policy TLV (Variable): indicates the ID of the route policy of this event, which is user specific or vendor specific, which can be used for mapping to the actual policy content. The policy content data retrieval is out of the scope of this document. The format of the Policy ID TLV is further defined in Figure 4. The Policy ID TLV is optional. At most one Policy TLV can be included in each Route Policy and Attribute Trace Message.
- o Pre Policy Attribute TLV (Variable): include the BGP route attributes before the policy is executed. The format of the Pre-policy Attribute TLV is further defined in Figure 4. The Pre-policy Attribute TLV is optional. At most one Pre Policy Attribute TLV can be included in each Route Policy and Attribute Trace Message.
- o Post Policy Attribute TLV (Variable): include the BGP route attributes after the policy is executed. The format of the Post-policy Attribute TLV is further defined in Figure 5. The Post-policy Attribute TLV is optional. At most one Post Policy Attribute TLV can be included in each Route Policy and Attribute Trace Message.
- o String TLV (Variable): leaves for future extension. The String TLV is optional. One or more String TLVs can be included in each Route Policy and Attribute Trace Message.

2.3.1. VRF/Table TLV

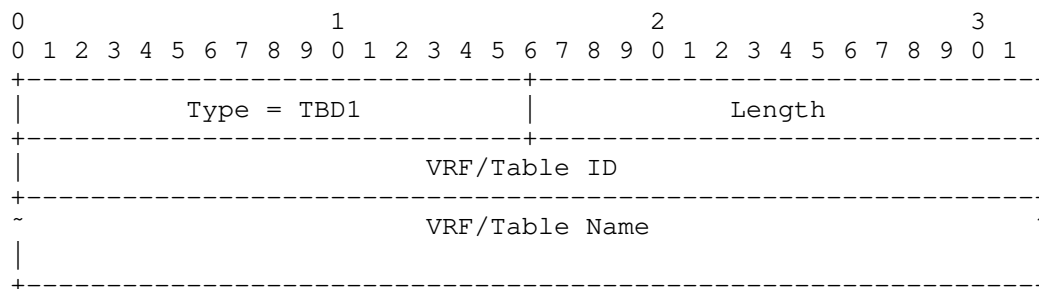


Figure 3: VRF/Table TLV

- o Type = TBD1 (2 Byte): VRF/Table TLV.
- o Length (2 Byte): indicates the total length of the VRF/Table ID field and the VRF/Table Name field.
- o VRF/Table ID (4 Bytes): indicates the VRF or table ID of this route.
- o VRF/Table name (Variable): indicates the VRF or table name of this route in the format of ASCII string. The string size MUST be within the range of 1 to 255 bytes.

2.3.2. Policy TLV

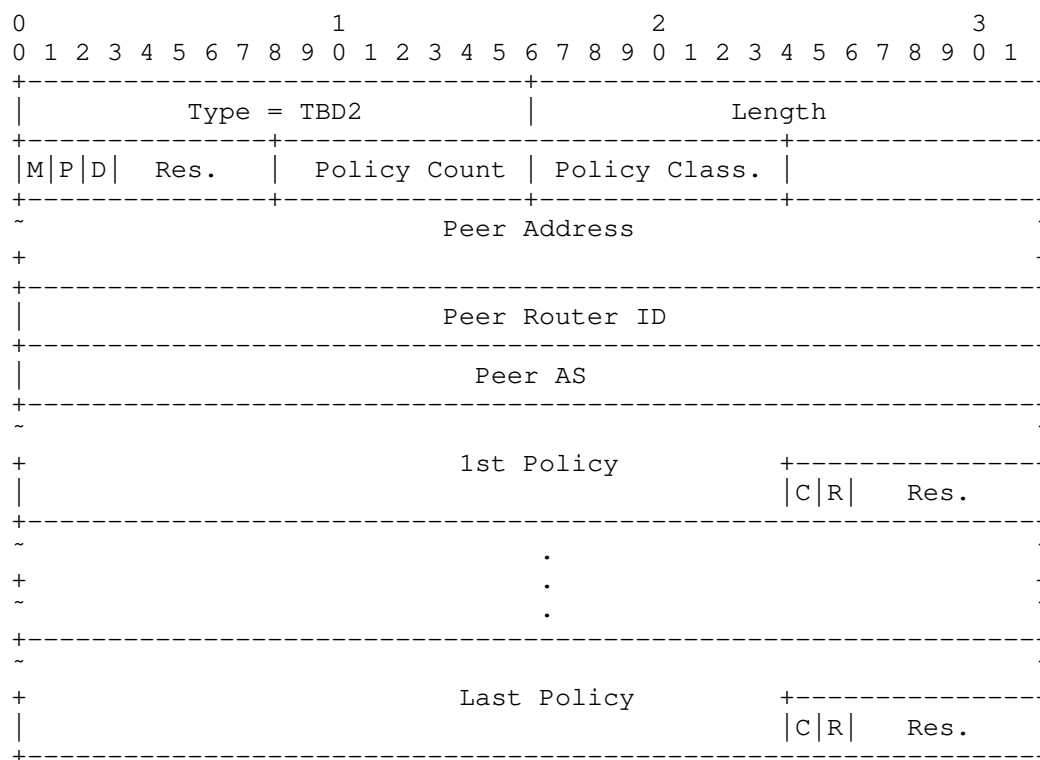


Figure 4: Policy TLV

- o Type = TBD2 (2 Byte): Policy TLV.
- o Length (2 Byte): indicates the length of the Policy Value field that follows it. The Policy value field includes the reserved Flag Byte, Policy Count field, Policy Classification field, Peer Router ID field, Peer AS field, and each Policy field.
- o Flag Byte (1 Byte): the M bit (the left most bit) indicates if the route in this event is matched (once or multiple times) or not by any policies. "0" means no match and "1" means else wise. When the M bit is set to "0", the Post Policy Attribute TLV SHALL not be included in the Message. The P bit (the second left bit) indicates if the matched result is Permit or Deny. "0" means Deny, and "1" means Permit. When the M bit is set to "0", any value of the P bit SHOULD be ignored. When the P bit is set to "0", the Post Policy Attribute TLV SHALL not be included in the Message. The D bit (the third left bit) indicates if there exists any difference between the pre-policy attributes and the post policy attributes. "0" means no difference, and "1" means difference

exists. When the D bit is set to "0", the Post Policy Attribute TLV SHALL not be included in the Message.

- o Policy Count (1 Byte): indicates the number of policies carried in this event.
- o Policy Classification (1 Byte): indicates the category of the policy. Currently 8 policy categories are defined: "00000000" indicates the Inbound policy; "00000001" indicates the Outbound policy; "00000010" indicating the Multi-protocol Redistribute policy (including routes import from other protocols, like ISIS/ OSPF and static routes), "00000011" indicates the Cross-VRF Redistribute policy (route import between VRF and global table and between VRFs); "00000100" indicates VRF Import policy (e.g., an IPv4 route within a VRF transformed from a VPNv4 route), "00000101" indicates VRF Export policy (e.g., a VPNv4 route transformed from an IPv4 route within an VRF); "00000110" indicates the Network policy (BGP network installment and advertisement), "00000111" indicates the Aggregation policy; "00001000" indicating the Route Withdraw (triggered by BGP Update or local actions, e.g., route aggregation). Specifications regarding each category can be included in the String TLV. For the route update, i.e., route creation and withdrawal, that is not processed by any route policy, the Policy Category field is set per the route update point. In addition, the Policy ID field in the Policy ID TLV SHOULD be set to 0.

o

Value	Policy Classification
00000000	Inbound policy
00000001	Outbound policy
00000010	Multi-protocol Redistribute
00000011	Cross-VRF Redistribute
00000100	VRF import
00000101	VRF export
00000110	Network
00000111	Aggregation
00001000	Route Withdraw

Table 1: Policy Classification

- o Peer Address: The remote IP address associated with the TCP session over which the encapsulated PDU was received. It is 4 bytes long if an IPv4 address is carried in this field (with the

12 most significant bytes zero-filled) and 16 bytes long if an IPv6 address is carried in this field.

- o Peer Router ID (4 Bytes): indicates the BGP Router ID where this policy is configured under. This field is used in combination with the Policy Classification field. If the Policy Classification field is set to "00000000", meaning Inbound policy, then this field is set to the BGP router ID where the route is received from; if the Policy Classification field is set to "00000001", meaning Outbound policy, then this field is set to the BGP router ID where the route is distributed to; If the Policy Direction field is set to any other values, then this field is set to all zeros.
- o Peer AS (4 Bytes): indicates the AS number of the BGP Peer that defined the Peer ID field.
- o 1st ~ Last Policy (Variable): indicates the Policy name and the Item ID of each policy match.
- o Flag Byte (1 Byte): the C bit (left most bit) indicates if the next subsequent policy has chaining relationship to the current policy. "1" means it's chaining relationship and "0" means else wise. For the flag byte following the Last Policy field, the C bit SHALL be set to "0". The R bit (second left bit) indicates if the next subsequent policy has recursion to the current policy. "1" means it's recursion and "0" means else wise. For the flag byte following the Last Policy field, the R bit SHALL be set to "0".

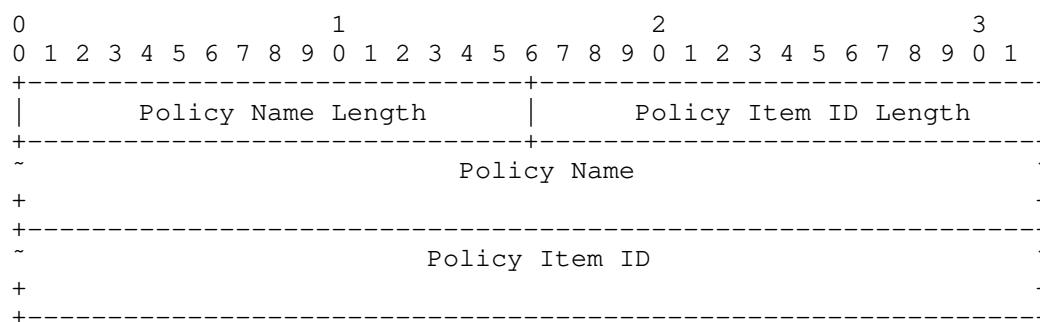


Figure 5: Policy field format

The Policy field consists of the Policy Name (Variable) and the Policy Item ID (Variable). The Policy Name and Policy Item ID fields are in the format of ASCII string. The length of Policy Name is indicated by the Policy Name Length (2 Bytes) field. The length of

Policy Item ID is indicated by the Policy Item ID Length (2 Bytes) field.

2.3.3. Pre Policy Attribute TLV

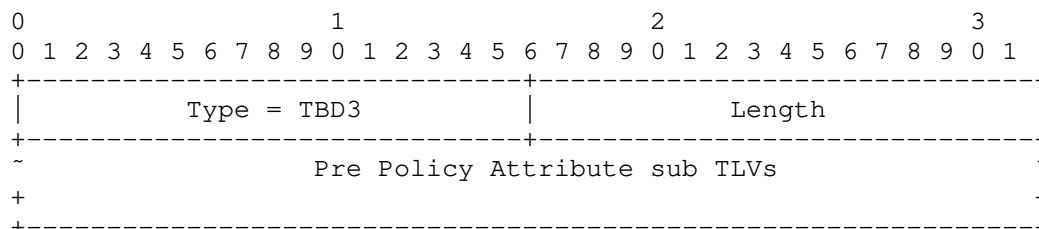


Figure 6: Pre Policy Attribute TLV

- o Type = TBD3 (2 Byte): Pre Policy Attribute TLV.
- o Pre Policy Attribute length (2 Byte): indicates the total length of the following Pre Policy Attribute sub TLVs.
- o Pre Policy Attribute sub TLVs (Variable): include the BGP route attributes before the policy is executed.

2.3.4. Post Policy Attribute TLV

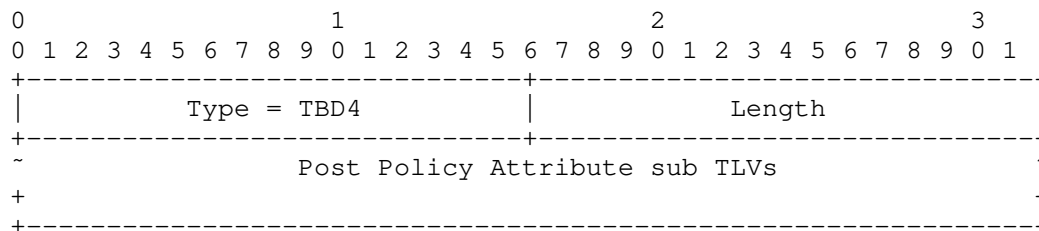


Figure 7: Post Policy Attribute TLV

- o Type = TBD4 (2 Byte): Post Policy Attribute TLV.
- o Post Policy Attribute length (2 Byte): indicates the total length of the following Post Policy Attribute sub TLVs.
- o Post Policy Attribute sub TLVs (Variable): include the BGP route attributes after the policy is executed.

2.3.5. String TLV

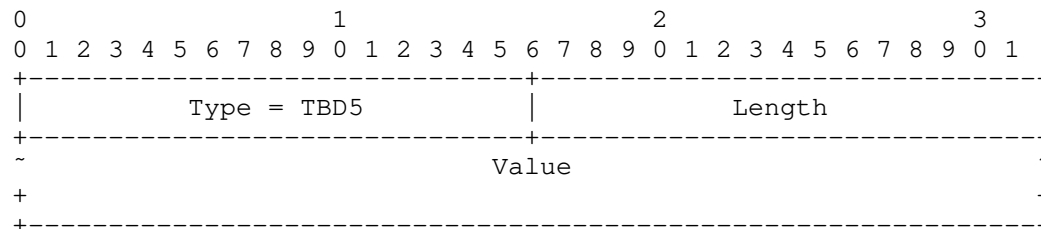


Figure 8: String TLV

- o Type = TBD5 (2 Byte): String TLV.
- o Length (2 Byte): indicates the length of the following value field.
- o Value (Variable): the textual expression of user-specific information in ASCII format.

An example of using the String TLV is expressing the route policy xpath information instead of using the Policy TLV.

3. Implementation Considerations

Considering the data amount of monitoring the route and policy trace of all routes from all BMP clients, users MAY trigger the monitoring at any user-specific time. Users MAY configure locally at the BMP client to monitor only user-specific routes or all the routes. In addition, users MAY configure locally at the BMP client whether to report the TLVs that are optional according to their own requirements, i.e., the Pre Policy Attribute TLV, Post Policy Attribute TLV, Policy ID TLV, and String TLV.

Successive recorded events from one device MAY be encapsulated in one Route Policy and Attribute Trace Message or multiple Route Policy and Attribute Trace Messages per the user configuration.

4. Acknowledgments

TBD.

5. IANA Considerations

This document defines the following new BMP Message type (Section 2.1).

- o Type = TBD: Route Policy and Attribute Trace Message.

This document defines the following new TLV types for the Route Policy and Attribute Trace Message (Section 2.3).

- o Type = TBD1 (2 Byte): VRF/Table TLV.
- o Type = TBD2 (2 Byte): Policy TLV.
- o Type = TBD3 (2 Byte): Pre Policy Attribute TLV.
- o Type = TBD4 (2 Byte): Post Policy Attribute TLV.
- o Type = TBD5 (2 Byte): String TLV.

6. Security Considerations

TBD.

7. 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>>.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <<https://www.rfc-editor.org/info/rfc5492>>.
- [RFC7854] Scudder, J., Ed., Fernando, R., and S. Stuart, "BGP Monitoring Protocol (BMP)", RFC 7854, DOI 10.17487/RFC7854, June 2016, <<https://www.rfc-editor.org/info/rfc7854>>.
- [RFC8671] Evens, T., Bayraktar, S., Lucente, P., Mi, P., and S. Zhuang, "Support for Adj-RIB-Out in the BGP Monitoring Protocol (BMP)", RFC 8671, DOI 10.17487/RFC8671, November 2019, <<https://www.rfc-editor.org/info/rfc8671>>.

[RFC9069] Evens, T., Bayraktar, S., Bhardwaj, M., and P. Lucente,
"Support for Local RIB in the BGP Monitoring Protocol
(BMP)", RFC 9069, DOI 10.17487/RFC9069, February 2022,
<<https://www.rfc-editor.org/info/rfc9069>>.

Authors' Addresses

Feng Xu
Tencent
Guangzhou
China

Email: oliverxu@tencent.com

Thomas Graf
Swisscom
Binzring 17
Zuerich 8045
Switzerland

Email: thomas.graf@swisscom.com

Yunan Gu
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: guyunan@huawei.com

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

Email: zhuangshunwan@huawei.com

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

Email: lizhenbin@huawei.com