Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 10, 2019

F. Xu
Tencent
Y. Gu
S. Zhuang
Z. Li
Huawei

March 9, 2019

BGP Route Policy and Attribute Trace Using BMP
draft-xu-grow-bmp-route-policy-attr-trace-00

Abstract

The generation of BGP adj-rib-in, local-rib or adj-rib-out comes from BGP protocol communication, and route policy processing. 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]. However, there lacks monitoring of how BGP routes are transformed from adj-rib-in into local-rib and then adj-rib-out (i.e., the BGP route policy processing procedures). This document describes a method of using BMP to trace the change of BGP routes in correlation with responsible 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 will expire on September 10, 2019.

Copyright Notice

Table of Contents

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 signle 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 collecte from each related device, the server is able to find the propagation path hop by hop. The identified path is helpful for

oeprators to better understand its network, and thus benefitting 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:

```
   +---------------------------------------------------------------+
   |                          Prefix length                        |
   +---------------------------------------------------------------+
   |                             Prefix                            |
   +---------------------------------------------------------------+
   |                       Route Distinguisher                     |
   +---------------------------------------------------------------+
   |                           Previous Hop                        |
   +---------------------------------------------------------------+
   |                          Event count                          |
   +---------------------------------------------------------------+
   |                       Total event length                      |
   +---------------------------------------------------------------+
   |                       Single event length (1st event)         |
   +---------------------------------------------------------------+
   |                           Event index                         |
   +---------------------------------------------------------------+
   |                       Timestamp(seconds)                      |
   +---------------------------------------------------------------+
   |                       Timestamp(microseconds)                 |
   +--------------------------------+------------------------------+
   |           Policy ID            |     Policy distinguisher     |
   +--------------------------------+------------------------------+
```

```
   |                            Peer ID                            |
   +---------------------------------------------------------------+
   |                            Peer AS                            |
   +---------------------------------------------------------------+
   |                         Peer VRF/Table name                   |
   +-------------------------------+-------------------------------+
   |           Peer  AFI           |          Peer SAFI            |
   +-------------------------------+-------------------------------+
   |                     Total attribute length                    |
   +---------------------------------------------------------------+
   |                        Attribute TLVs                         |
   +---------------------------------------------------------------+
   ~                                                               ~
   +                                                               +
   ~                                                               ~
   +---------------------------------------------------------------+
   |                        Single event length (Last event)       |
   +---------------------------------------------------------------+
   ~                                                               ~
   +                                                               +
   ~                                                               ~
   +---------------------------------------------------------------+
   |                        Attribute TLVs                         |
   +---------------------------------------------------------------+
```

Figure 2: Route Policy and Attribute Trace Message format

- o Prefix Length (1 Byte): indicates the length of the prefix.

- o Prefix (Variable): indicates the monitored prefix, with the length defined by Prefix Length field.

- o Route Distinguisher (8 Bytes): If the route is an IPv4 route, this field is zero- filled. If the peer is a VPNv4 route, it is set to the route distinguisher (RD) of the route.

- o Previous Hop (4 Bytes): indicates the BGP peer ID where this route is learnt from. If the route is locally generated, then field is set to the local BGP router ID (global or VRF specific).

- o Event Count (1 Byte): indicates the total number of policy processing event recorded in this message.

- o Total event lenght (1 Byte): indicates the total length of the following fields including all events, where the total number is indicated by the Event Count field.

- Single event lenght (1 Byte): indicates the total length of a single policy process event, including the following fields that belong to this event.

- Event index (1 Byte): indicates the sequence number of this event, staring from 1 and increases by 1 for each event recorded in order.

- Timestamp (4 Bytes): indicates the time when the policy of this event starts excution, expressed in seconds and microseconds since midnight (zero hour), January 1, 1970 (UTC).

- Policy ID (Variable): indicates the ID of the route policy of this event, which is user specific or vendor specific. It consists of the Route Policy Name and the Route Policy Item/Node ID. The Policy name and Item/Node ID is in the format of ASCII string, the length of both fields are indicated by the Policy length and Item/Node length fields, respectively

-

```
+--------------------+---------------------+
|   Policy length    |   Policy name       |
+------------------------------------------+
|   Item/node length|   Item/Node ID       |
+------------------------------------------+
```

- Policy Distinguisher (4 Bits): indicates the category of the policy. Currently 3 policy categories are defined: "0000" indicating the inbound policy, "0001" indicating the outbound policy, "0010" indicating the redistribution policy. More categories to be defined.

- Peer ID (4 Bytes): indicates the BGP Peer ID where this policy is configured under. This field is used in combination with the Policy Direction field. If the Policy Direction field is set to "0000", meaning inbound policy, then this field is set to the BGP Peer ID where the route is received from; if the Policy Direction field is set to "0001", meaning outbound policy, then this field is set to the BGP Peer ID where the route is distributed to; If the Policy Direction field is set to "0010", meaning redistribution policy, then this field is set to the local BGP router ID (global or VRF specific).

- Peer AS (4 Bytes): indicates the AS number of the BGP Peer that defined the Peer ID field.

- 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. The VRF/Table name information varies for the same route under different policy processing event. For example, an IPv4 route is received from a CE router at the PE router through iGBP, an RD is attached to this IPv4 route (under VRF name A) and making it a VPNv4 route, and then this VPNv4 route (under the Global routing table) is distributed to the RR. During this process, the VRF/Table name information changes from VRF A to the Global routing Table name at the inbound and outbound policy process.

- AFI/SAFI (2 Bytes): indicates the AFI/SAFI of the route. The AFI/SAFI information varies for the same route under different policy processing event. For example, an IPv4 route is received from a CE router at the PE router through iGBP, an RD is attached to this IPv4 route and making it a VPNv4 route, and then this VPNv4 route is distributed to the RR. During this process, the AFI information changes from IPv4 to VPNv4 at the inbound and outbound policy process.

- Total attribute length (2 Bytes): indicates the total length of the following route attribute TLVs.

- Attribute TLVs: include atttributes that are currently carried in BGP Update messages (e.g., Community, Ext-community, Next Hop, AS path, MED...) and those that are not (to be defined).

3. Implementation Example

```
                           +----------+
                   +------>+BMP server+<-------+
                   |       +--+-------+        |
                   +          ^    ^           |
           Event 1,2,3  +-----+    |           +
                   +    +          +       Event 9,10,11
                   | Event 4,5  Event 6,7,8    +
                   |    +          +           |
                   |    |          |           |
10.1.1.1/24    ****|****|**********|***********|*****
+--------->    *   |    |          |           | AS0*
+-----+        *   |    |       +--+--+        |    *
| CE1 ++eBGP+  *   | +--------->+ RR  +------+ |    *
+-----+     |  *   | |  |       ++----+      | |    *
  AS1       |  *   | |  |        ^    iBGP   | |    *
            |  *   | | ++----+   | 65000:10  | |    *     10.1.1.1/24
10.1.1.1/24 +----------> PE2 +---+10.1.1.1/24| |    *    +----------->
+--------->    *   | | +-----+               | |    *         +-----+
+-----+        *   | |                 iBGP  | |    *  +eBGP+>+ CE4 |
| CE2 ++eBGP+  *   | |  iBGP        65000:10 + |    *  |      +-----+
+-----+     |  *   | +65000:10     10.1.1.1/24 |    *  |        AS4
  AS2       |  *   | 10.1.1.1/24             + |    *  |
            |  *  ++----+                 +--v-++   *  |
+-----+     +-----> PE1 |                 | PE3 +------+      +-----+
| CE3 ++eBGP+----->     +---------------->+     +------+eBGP+>+ CE5 |
+-----+        *  +-----+                 +-----+   *         +-----+
10.1.1.1/24    *                                    *     10.1.1.1/24
+-------->     **************************************    +----------->
  AS3                                                         AS5
```

Figure 3: Route Policy and Attribute Trace record implementation example

We take the network shown in Figure 2 as an example to show how to use Route Policy and Attribute Trace Messages to recover the footprint of the route propagation. Notice that only basic events required for footprint recovery are listed here.

Suppose a prefix 10.1.1.1/24 is sent from both CE2 and CE3 to PE1 through eBGP peering, PE1 processes the two Update messages with inbound policies. Such procedure is recorded as two events, namely Event 1 and Event 2. Then PE1 selects the route from CE2 as the best route, add it to VRF 1, and then distribute the VPNv4 route to RR. The distribution procedure is recorded by PE1 as Event 3. As an example, the Route Policy and Attribute Trace Message of Event 1, 2, 3 is listed as follows. Only fields related to footprint recovery are listed in the message shown below. Specifically, the Previous Hop information is carried in Event 3 when outbounding the route, indicating that the outbounded route is learnt from CE2. The same

prefix is sent from CE1 to PE2, added to VRF 1 and then distributed to RR in the form of VPNv4 route. Two events, Event 4 (inbound) and Event 5 (outbound) are recorded by PE2. Now for RR, prefix 10.1.1.1/24 is received from both PE1 and PE2 in the form of VPNv4 route. RR selects the route from CE2 as the best route, and distribute it to PE3. Three events, Event 6 (PE2 inbound), Event 7 (PE1 inbound), Event 8 (PE3 outbound) are recorded in this case. PE3 receives the VPNv4 route from RR, adds it to VRF 1 and then distribute the IPv4 route to CE4 and CE5, respectively. Here, three events are recorded, Event 9 (RR inbound), Event 10 (CE4 outbound) and Event 11 (CE5 outbound).

```
+---------------------------------------------------------------+
|                         RD: 65000:10                          |
+---------------------------------------------------------------+
|                      Prefix:  10.1.1.1/24                     |
+---------------------------------------------------------------+
|                             Event 1                           |
+---------------------------------------------------------------+
|                            Timestamp 1                        |
+--------------------------------+------------------------------+
|    Policy ID: WC1, node 101    |          Inbound policy      |
+--------------------------------+------------------------------+
|                           Peer ID: CE1                        |
+---------------------------------------------------------------+
|                           Peer AS: AS1                        |
+---------------------------------------------------------------+
|                       VRF/Table name: VRF 1                   |
+---------------------------------------------------------------+
|                            AFI: IPv4                          |
+---------------------------------------------------------------+
|                        Previous Hop: CE1                      |
+---------------------------------------------------------------+
|                             Event 2                           |
+---------------------------------------------------------------+
|                            Timestamp 2                        |
+--------------------------------+------------------------------+
|   Policy ID: WC1, node 102     |          Inbound policy      |
+--------------------------------+------------------------------+
|                           Peer ID: CE2                        |
+---------------------------------------------------------------+
|                           Peer AS: AS2                        |
+---------------------------------------------------------------+
|                       VRF/Table name: VRF 1                   |
+---------------------------------------------------------------+
|                            AFI: IPv4                          |
+---------------------------------------------------------------+
|                         Previous Hop: CE2                     |
```

```
   +---------------------------------------------------------------+
   |                             Event 3                           |
   +---------------------------------------------------------------+
   |                            Timestamp 3                        |
   +--------------------------------+------------------------------+
   |    Policy ID: RR1, node 103    |         Outbound policy      |
   +--------------------------------+------------------------------+
   |                           Peer ID: RR                         |
   +---------------------------------------------------------------+
   |                           Peer AS: AS0                        |
   +---------------------------------------------------------------+
   |                       VRF/Table name: VRF 1                   |
   +---------------------------------------------------------------+
   |                            AFI: VPNv4                         |
   +---------------------------------------------------------------+
   |                         Previous Hop: CE1                     |
   +---------------------------------------------------------------+
```

Figure 4: Event 1,2,3 data partial view

The BMP server can use the collected events to recover the route footprint. The key information required from recovery is the Timestamp of each event, and the Previous Hop of the route. The Timestamp allows the server to identify the order of each event, while the Previous Hop information, combined with the outbound peer information, allows the server to recover the route propagation hop by hop.

4. Implementation Considerations

Considering the data amount of monitoring the route and policy trace of all routes from all BMP clients, the Route Policy and Attribute Trace monitoring MAY be triggered by user at any user-specific time, and MAY be applied to user-specific routes as well as all routes. Successive recored 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.

5. Acknowledgements

TBD.

6. IANA Considerations

TBD.

7. Security Considerations

TBD.

8. 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-03 (work in progress), December 2018.

[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-02 (work in progress), September 2018.

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

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

[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>.

Authors' Addresses

Feng Xu
Tencent
Guangzhou
China

Email: oliverxu@tencent.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