Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2019

Y. Gu
Huawei
J. Chen
Tencent
P. Mi
S. Zhuang
Z. Li
Huawei

March 11, 2019

VPN Traffic Engineering Using BMP
draft-gu-grow-bmp-vpn-te-00

Abstract

The BGP Monitoring Protocol (BMP) is designed to monitor BGP running status, such as BGP peer relationship establishment and termination and route updates. This document provides a traffic engineering (TE) method in the VPN (Virtual Private Network) scenario using BMP.

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].

Table of Contents

1. Introduction

The Border Gateway Protocol (BGP) [RFC4271], as an inter-Autonomous (AS) routing protocol, is used to exchange network reachability information between BGP systems. Later on, RFC4760 [RFC4760] extends BGP to carry not only the routing information for BGP, but also for multiple Network Layer protocols (e.g., IPv6, Multicast, etc.), known as the MP-BGP (Multiprotocol BGP). The MP-BGP is currently widely deployed in case of MPLS L3VPN, to exchange VPN labels learned for the routes from the customer sites over the MPLS network. BGP routes are needed for both intra-domain and inter-domain route optimization. Before BGP Monitoring Protocol (BMP) [RFC7854] was introduced, BGP routes could be only obtained through manual query, such as screen scraping. The introduction of BMP greatly improves the BGP route monitoring efficiency and accuracy.Currently, it 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].

In the MPLS (Multiprotocol Label Switching) VPN traffic egnieering scenario, the controller distributes optimized route entries with MPLS VPN labels (inner labels) to the target devices. The target devices use the inner MPLS VPN labels to find the corresponding VRF (Virtual routing and forwarding) instance, and then add the optimized route entries into the target VRF table. Techically, it's workable to extract the labels from VPNv4 routes by monitoring the VPNv4 routes exchanged between two PE (provider edge) devices, i.e., by monitoring the adj-rib-out of and adj-rib-in of both PEs. However, unlike the public BGP routes and IGP routes, VPNv4 routes are not usually used for either the inter-domain or intra-domain traffic optmization. Thus, it's not very cost efficient, from the perspective of CPU and network bandwidth consumption, to monitor the VPNv4 routes only for the purpose of label extraction.

Depending on the implementation scenarios, there are typically different ways of allocating the VPN route labels: per route per label, per VRF per label, per next hop per label, and so on. For example, in the Multi-AS VPN case, the redistribution of labeled VPNv4 routes from one AS to another can be realized through setting up the EBGP peering between ASBRs (Autonomous System Border Routers). In this case, the per route per label allocation method is preferred. However, per route per label allocation can be very consuming as for the label space, thus, in many cases the per VRF/next hop per label assignment modes are adopted.

This document descrbes a method using BMP to collect the MPLS VPN label information. A new BMP message type is proposed to carry the label information. More specifically, in the per route per label case, the VRF nformation, route prefix and label are included in the newly defined BMP Label Message. In the per instance per label case, the VRF information and label are included in the newly defined BMP Label Message, while in the per next hop per label case, the VRF information, next hop and label are included in the newly defined BMP Label Message. The report of BMP Label Message is triggered by the label assignment chnage.

There are several merits of using the BMP Label Message type to collect the MPLS VPN labels compared with extracting labels from the monitored VPNv4 routes:

o It saves work of extracting the label information from the VPNv4 routes, and saves network bandwidth considering that VPNv4 routes includes all route attributes that are not necessary in this case.

o In the per instance/next hop per label assignment cases, the same VPN label is used for multiple VPNv4 routes. The BMP Label Message only report the label information once (if no change), and

thus saves network resources compared with the repeated label report by monitoring VPNv4 routes.

- The label assignments are typically less dynamic compared with the VPNv4 routes. Thus, acquiring the label information through the real-time monitoring of VPNv4 routes is not quite necessary.

All in all, it's more efficient to collect the MPLS VPN label independently than extracting it from VPNv4 routes. In Section 2, the BMP Label Message format is defined, and in Section 3, two specific implementation examples are provided to show case the usage of BMP Label Message.

2. VPN TE Using BMP

This document defines a new BMP message type called the Label Message to carry the VPN label.

2.1. Common Header

This document defines a new BMP message type to carry the VPN label data.

- Type = TBD: Label Message

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

2.2. Per Peer Header

The Label Message is not per peer based, thus it does not require the Per Peer Header.

2.3. Label Message

```
+--------------------------------+-----------------------------+
|     Label Assignment Mode      |          Reserved           |
+--------------------------------+-----------------------------+
|                     Label Mapping Information                |
+--------------------------------------------------------------+
|                               Label                          |
+--------------------------------------------------------------+
```

Figure 1: BMP Label Message

- Label Assignment Mode (4 Bits): indicates how label is assigned. Curerntly, 3 types of label assignment mode are defined: "0000" indicating the per instance per label assignment mode, "0001"

indicating the per next hop per label assignment mode, "0010" indicating the per instance per label assignment mode. More modes can be defined per requirement.

o Reserved (1 Byte): reserved for future use.

o Label Mapping Information (Variable): is interpreted in combination with the Label Assignment Mode field. If the Label Assignment Mode field is set to "0000", meaning per instance per label assignment mode, then this field is set to VRF Route Distinguisher; If the Label Assignment Mode field is set to "0001", meaning per next hop per label assignment mode, then this field is set to the next hop address; If the Label Assignment Mode field is set to "0010", meaning per route per label assignment mode, then this field is set to the route prefix.

o Label (3 Bytes): indicates the label value with 20 bits label and 4 bits zero padding.

More specifically, the Label Mapping Information field is defined as follows. Regarding different values indicated in the Label Assignment Mode field,

```
+-----------------------------------------------------------------+
|                              Length                             |
+-----------------------------------------------------------------+
|                              VRF RD                             |
+-----------------------------------------------------------------+
|                         Next Hop/Prefix                         |
+-----------------------------------------------------------------+
```

Figure 2: Label Mapping Information

o Length (2 Bytes): indicates the length of the following Label Mapping Information value fields. The Length field value SHALL be set in accordance with the Label Assignment Mode field. If the Label Assignment Mode is set to "0000", the Length field is set to the length of the VRF RD field (i.e., 8 Bytes); If the Label Assignment Mode is set to "0001", the Length field is set to the length of the VRF RD field (8 Bytes) + the length of the Next Hop field (variable); If the Label Assignment Mode is set to "0010", the Length field is set to the length of the VRF RD field (8 Bytes) + the length of the Prefix field (variable).

o VRF RD (8 Bytes): indicates the route distinguisher (RD) of the VRF. In either the "per instance per label" case, or "per next hop per label" case, or "per route per label" case, the VRF information (i.e., RD) SHALL be indicated in this field.

- o Next Hop/Prefix (Variable): is interpreted in combination with the Label Assignment Mode field and the Length field. If the Label Assignment Mode is set to "0000", this field SHALL be set empty; If the Label Assignment Mode is set to "0001", this field SHALL be set to the next hop address (i.e., the CE's address), with length indicated by the Length field (i.e., Length value - 8 Bytes); If the Label Assignment Mode is set to "0010", this field SHALL be set to the prefix of the route, with length indicated by the Length field (i.e., Length value - 8 Bytes)

3. Implementation Examples

In this section, we use two examples to more specifically explain how to use BMP for VPN traffic engineering.

```
                                    +-------------+
                 Option 1:          |  BMP server |          Option2:
                 10.2.1.0/24 +-------+     +     +---------+10.2.1.0/24
                 NH:CE1      |      |  Controller |         |NH:PE1
                 Label:100   |      +-+----------++         |Label:100
  10.2.0.0/24                |  VRF1  ^  ^VRF1   |          |
  10.1.0.0/24    10.1.1.0/24 |  R1:100|  |R1:500 |          |10.1.1.0/24
      +++        NH:PE2      |  R2:200|  |R2:600 |          |NH:PE2
       |         Label:600   |  R3:300|  |R3:700 |          |Label:600
       |                     |  R4:400|  |R4:800 |          |
       |                     |  ******|**|*******|********  |
  +----+---+ R1:10.2.0.0/16  v  *     |  |       +   AS0 *  |
  |  CE1   | R2:10.1.0.0/16 ++-----+  |  |  Option 1:    *  |
  | (ISP1  +---------------->+  PE1 +--+  |  10.2.1.0/24  *  |
  |  AS1)  +------------|   | VRF1 |     |  NH:PE1       *  |
  +--------+ R1,R2   +----->+      |     |  Label:100    *  |
  R3:10.2.0.0/17     |  |   +-------+    |  10.1.1.0/24  *  v
  R4:10.1.0.0/17     |  |       *        |  NH:CE1    +-----++   +---+
       +             |  |       *        |  Label:600 | PE3  +---+AS4|
       v             |  |       *        |       +    | VRF1 |   +---+
  +----+---+ R3,R4   |  |   +------+-----+       |    |      |
  |   CE2  +---------+  +-->+  PE2 |             |    +-------+
  |  (ISP2 |                | VRF1 +<------------+       *
  |   AS2) +---------------->+      |                     *
  +--------+    R3,R4       +-------+                     *
                                ************************** 
```

Figure 3: VPN TE using BMP example: per route per label

Two prefixes 10.2.0.0/24 and 10.1.0.0/24 are generated from ISP1 (AS1), advertised to ISP 2 (AS2) in the format of R3: 10.2.0.0/17, and R4: 10.1.0.0/17, and also advertised to AS0 in the format of R1:

10.2.0.0/16, and R2: 10.1.0.0/16. R1, R2 are advertised to both PE1 and PE2 in AS0, and so are R3 and R4. By the rule of the longest prefix match, any traffic, with the destination address within the subnets of 10.2.0.0/16 or 10.1.0.0/16, coming from AS4 that traverses AS0 will exit from PE2. This may cause unbalanced traffic loads on PE2 and PE1. In addition, the costs of traversing through AS1 and AS2 might be different due to business contracts assigned between different ISPs. Now suppose for traffic and cost optimization purposes, the operator wants to: 1) steer the traffic, with the destination address within the subnets of 10.2.0.0/16, to exit from PE1 and then traverse AS1 (ISP1) to its destination; 2) steer the traffic, with the destination address within the subnets of 10.1.0.0/16, to exit from PE2 and then traverse AS1 (ISP1) to its destination.

In the example shown in Figure 2, the VPN label assignement mode is per route per label. Thus, PE1 assigns R1, R2, R3, R4 with label 100, 200, 300, 400, respectively, under VRF1. PE2 assigns R1, R2, R3, R4 with label 600, 700, 800, 900, respectively, under VRF1. Using the BMP Label Message, PE1 and PE2 reports to the BMP server with the per-route labels, which also includes the VRF RD information. Then the TE controller (suppose it's colocated with the BMP server) combines the label information with routes, and distribute the optimized routes with label to either the ingress or egress devices. There are typically two options:

o Option 1: The controller distributes the optimized route to the Egress devices, i.e., PE1 and PE2. For optimizing 10.2.0.0/16 traffic, controller distributes 10.2.0.0/24 with next hop as CE1, label as 100, RT as 100:1 to PE1, so that when traffic, with the destination address within the subnets of 10.2.0.0/16, arrives at PE1 will exit from PE1 and choose CE1 (ISP1) as its next hop. Controller also distributes 10.2.0.0/24 with next hop as PE1, label as 100, RT as 100:1 to PE1, so that when traffic, with the destination address within the subnets of 10.2.0.0/16, arrives at PE2 will exit from PE1 and choose CE1 (ISP1) as its next hop. For optimizing 10.1.0.0/16 traffic, controller distributes 10.1.0.0/24 with next hop as PE2, label as 600, RT as 100:1 to PE1, so that when traffic, with the destination address within the subnets of 10.1.0.0/16, arrives at PE1 will exit from PE2 and choose CE1 (ISP1) as its next hop. Controller also distributes 10.1.0.0/24 with next hop as CE1, label as 600, RT as 100:1 to PE2, so that when traffic, with the destination address within the subnets of 10.1.0.0/16, arrives at PE2 will exit from PE2 and choose CE1 (ISP1) as its next hop.

o Option 2: The controller distributes a more specific route to the Ingress device, i.e., PE3. Controller distributes 10.2.0.0/24

with next hop as PE1, label as 100, RT as 100:1 to PE3, so that when traffic, with the destination address within the subnets of 10.2.0.0/16, arrives at PE3 will exit from PE1 and choose CE1 (ISP1) as its next hop. Controller also distributes 10.1.0.0/24 with next hop as PE2, label as 600, RT as 100:1 to PE3, so that when traffic, with the destination address within the subnets of 10.2.0.0/16, arrives at PE3 will exit from PE2 and choose CE1 (ISP1) as its next hop.

```
                                    +-------------+
                 Option 1:          |  BMP server |          Option2:
                 10.2.1.0/24 +------+      +      +-+ +------+10.2.1.0/24
                 NH:CE1      |      |  Controller |         |NH:PE1
                 Label:1000  |      +-+----------++         |Label:1000
                             |  VRF1  ^  ^VRF1   +          |
  10.2.0.0/24    10.1.1.0/24 |CE1:1000|  |CE1:3000          |10.1.1.0/24
  10.1.0.0/24    NH:PE2      |CE2:2000|  |CE2:4000          |NH:PE2
      +++        Label:3000  |        |  |       +          |Label:3000
       |                     |        |  |       |          |
       |                     |  ******|**|*******|********  |
  +----+---+ R1:10.2.0.0/16  v  *     |  |       +   AS0 *  |
  |  CE1   | R2:10.1.0.0/16 ++-----+  |  |  Option 1:    *  |
  | (ISP1  +--------------->+  PE1 +--+  |  10.2.1.0/24  *  |
  |  AS1)  +------------|   | VRF1 |     |  NH:PE1       *  |
  +--------+ R1,R2   +----->+      |     |  Label:1000   *  |
 R3:10.2.0.0/17      |  |   +------+     |  10.1.1.0/24  *  v
 R4:10.1.0.0/17      |  |       *        |  NH:CE1    +-----++   +---+
       |             |  |       *        |  Label:3000| PE3  +---+AS4|
       v             |  |       *        |       +    | VRF1 |   +---+
  +----+---+ R3,R4   |  |   +------+-----+       |    |      |
  |   CE2  +---------+  +-->+  PE2 |             |    +------+
  |  (ISP2 |                | VRF1 +<------------+       *
  |   AS2) +--------------->+      |                     *
  +--------+    R3,R4       +------+                     *
                                ********************************
```

Figure 4: VPN TE using BMP example: per next hop per label

In the example shown in Figure 3, he VPN label assignement mode is per next hop per label. Comparing the two examples in Figure 2 and Figure 3, less label information are reported though BMP if the label is allocated per next hop.

4. Acknowledgements

TBD.

5. IANA Considerations

TBD.

6. Security Considerations

TBD.

7. 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>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[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

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

Email: guyunan@huawei.com

Jie Chen
Tencent

Email: jasonjchen@tencent.com

Penghui Mi
Huawei
Shenzhen, Guangdong
China

Email: mipenghui@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