Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 9, 2020

F. Xu
Tencent
T. Graf
Swisscom
Y. Gu
S. Zhuang
Z. Li
Huawei

January 6, 2020

BGP Route Policy and Attribute Trace Using BMP
draft-xu-grow-bmp-route-policy-attr-trace-04

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 [I-D.ietf-grow-bmp-local-rib] and BGP adj-rib-out [I-D.ietf-grow-bmp-adj-rib-out]. 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 July 9, 2020.

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 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:

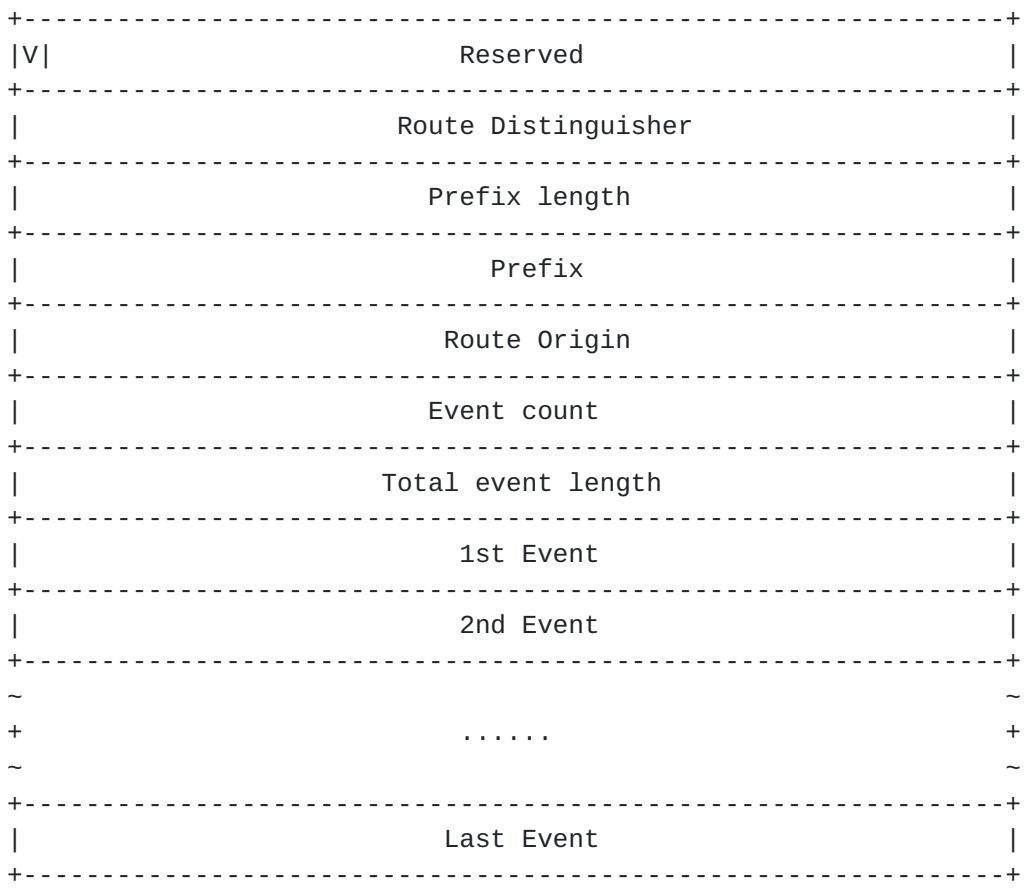

Figure 1: Route Policy and Attribute Trace Message format

- Flags (1 Byte): The V flag indicates that the Peer address is an IPv6 address. For IPv4 peers, this is set to 0.

- Route Distinguisher (8 Bytes): indicates the route distinguisher (RD) related to the route.

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

- 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.

- 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.

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

- 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.

- 1 ~ Last event: indicates each event, stacked one by one in order of time. The event format is further defined as follows.

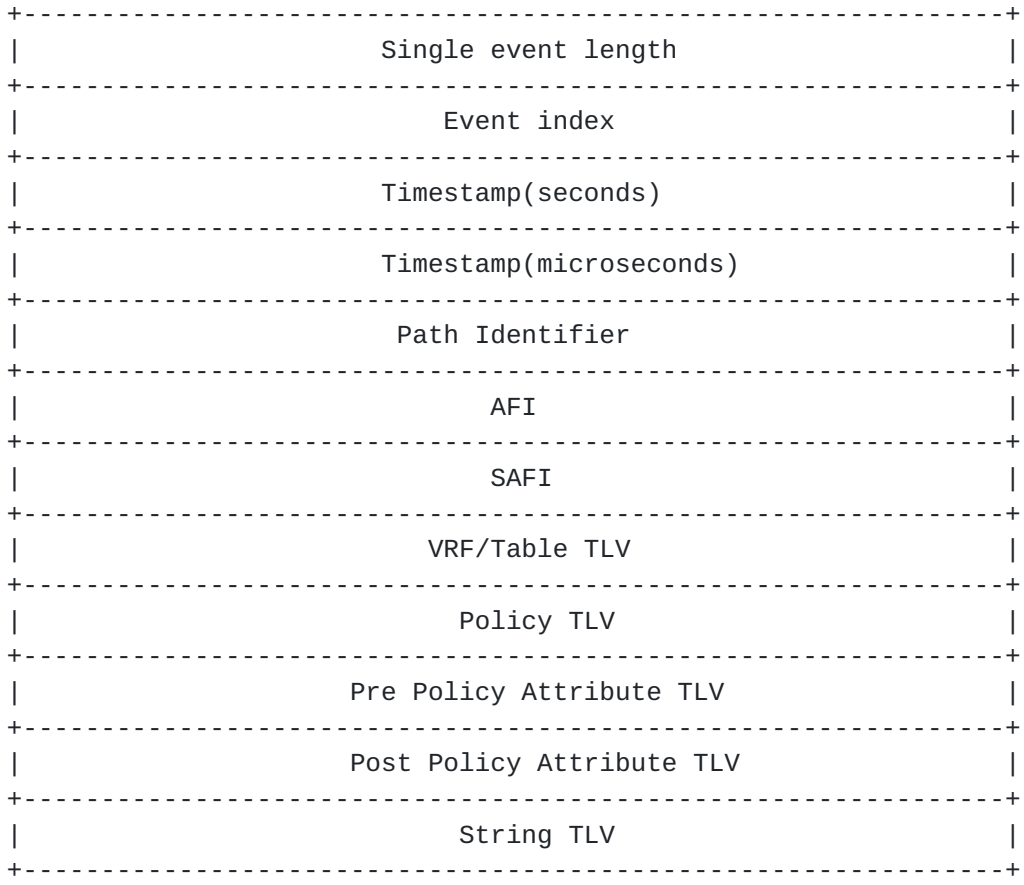

Figure 2: Event format

- Single event length (2 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 (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).

- 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.

- AFI (2 Bytes)/SAFI (1 Byte): indicates the AFI/SAFI of the route.

- VRF/Table TLV (Variable): indicates the VRFinformation of the route. The format of the VRF/Table TLV is further defined in Figure 3. The VRF/Table TLV is optional.

- 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.

- 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.

- 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.

- String TLV (Variable): leaves for future extension. The String TLV is optional.

2.3.1. VRF/Table TLV

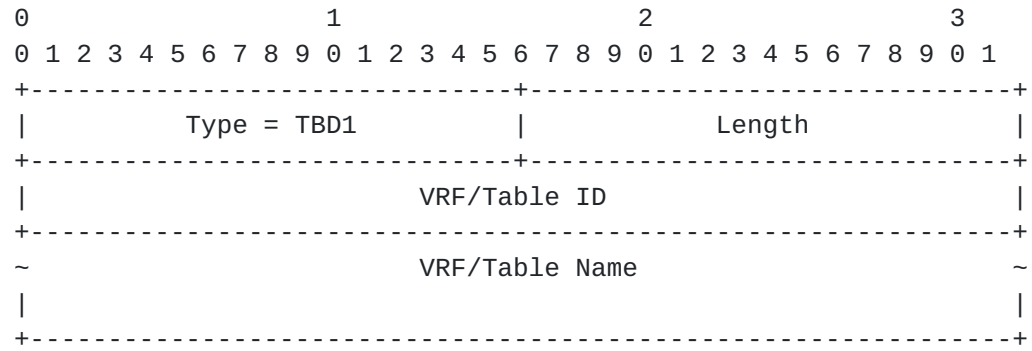

Figure 3: VRF/Table TLV

- Type = TBD1 (2 Byte): VRF/Table TLV.

- Length (2 Byte): indicates the length of the VRF/Table name field.

- VRF/Table ID (4 Bytes): indicates the VRF or table ID of this route.

- 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

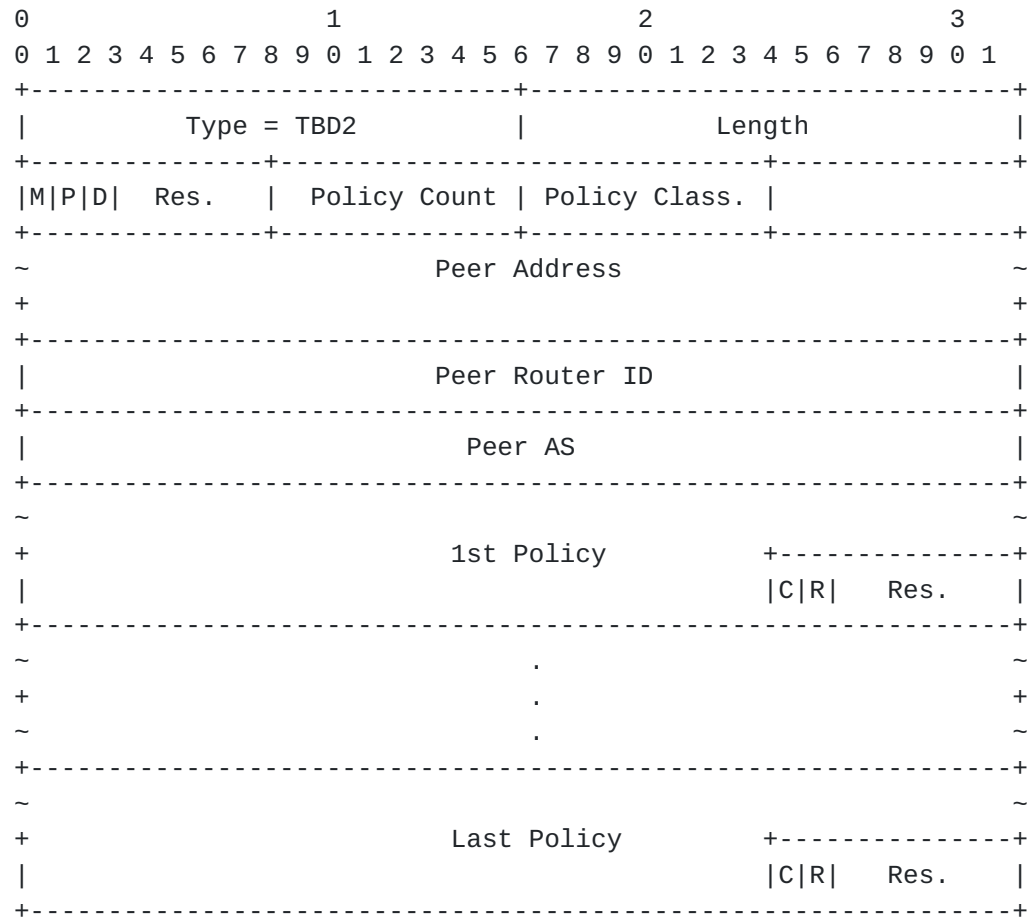

Figure 4: Policy TLV

- Type = TBD2 (2 Byte): Policy TLV.

- 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.

- 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.

- Policy Count (1 Byte): indicates the number of policies (in the format of Policy name + Item ID) carried in this event.

- 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.

-

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

- 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.

- 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.

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

- 1st ~ Last Policy (Variable): indicates the Policy name and the Item ID of each policy match.

- 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".

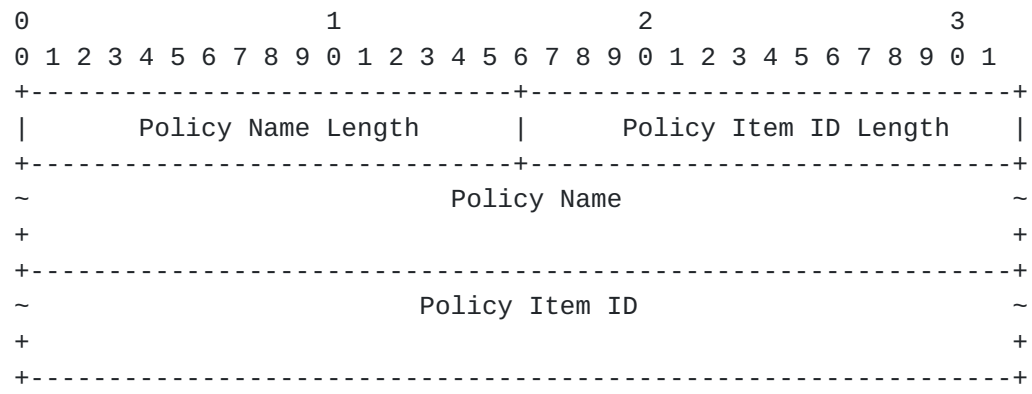

Figure 5: Policy field format

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

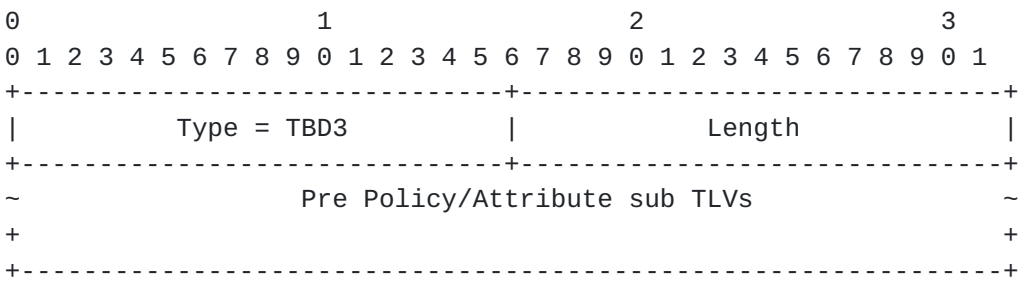

Figure 6: Pre Policy Attribute TLV

- Type = TBD3 (2 Byte): Pre Policy Attribute TLV.

- Pre Policy Attribute length (2 Byte): indicates the total length of the following Pre Policy Attribute sub TLVs.

- Pre Policy Attribute sub TLVs (Variable): include the BGP route attributes before the policy is executed.

2.3.4. Post Policy Attribute TLV

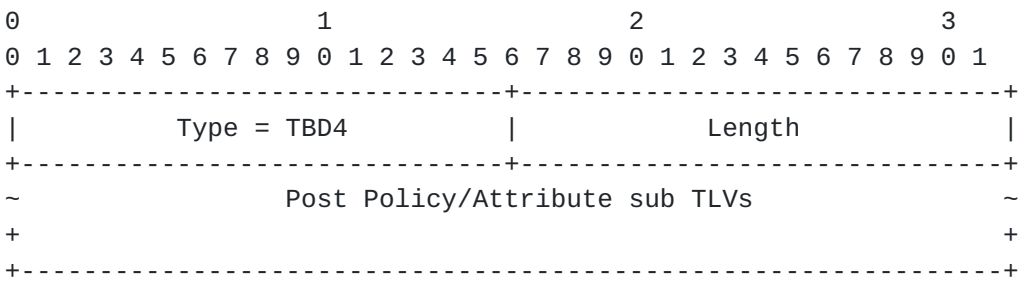

Figure 7: Post Policy Attribute TLV

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

2.3.5. String TLV

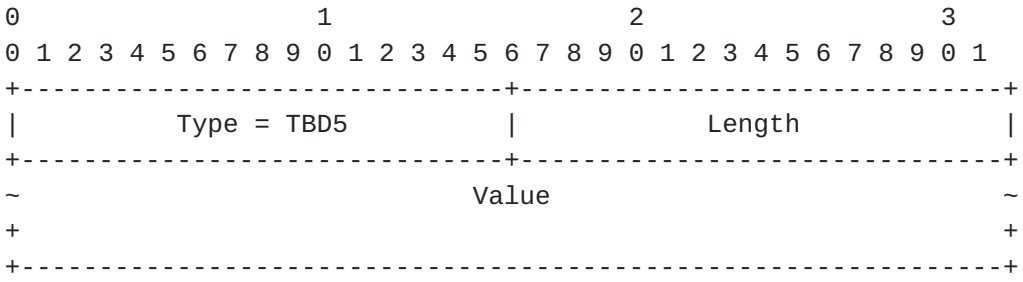

Figure 8: String TLV

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

One or more Optional String TLVs can be used. 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 Optional TLV.

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.

4. Acknowledgments

TBD.

5. IANA Considerations

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

- 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).

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

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-07 (work in progress), August 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-06 (work in progress), November 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>.

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