Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2020

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

July 7, 2019

BGP Route Policy and Attribute Trace Using BMP
draft-xu-grow-bmp-route-policy-attr-trace-01

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 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 8, 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 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

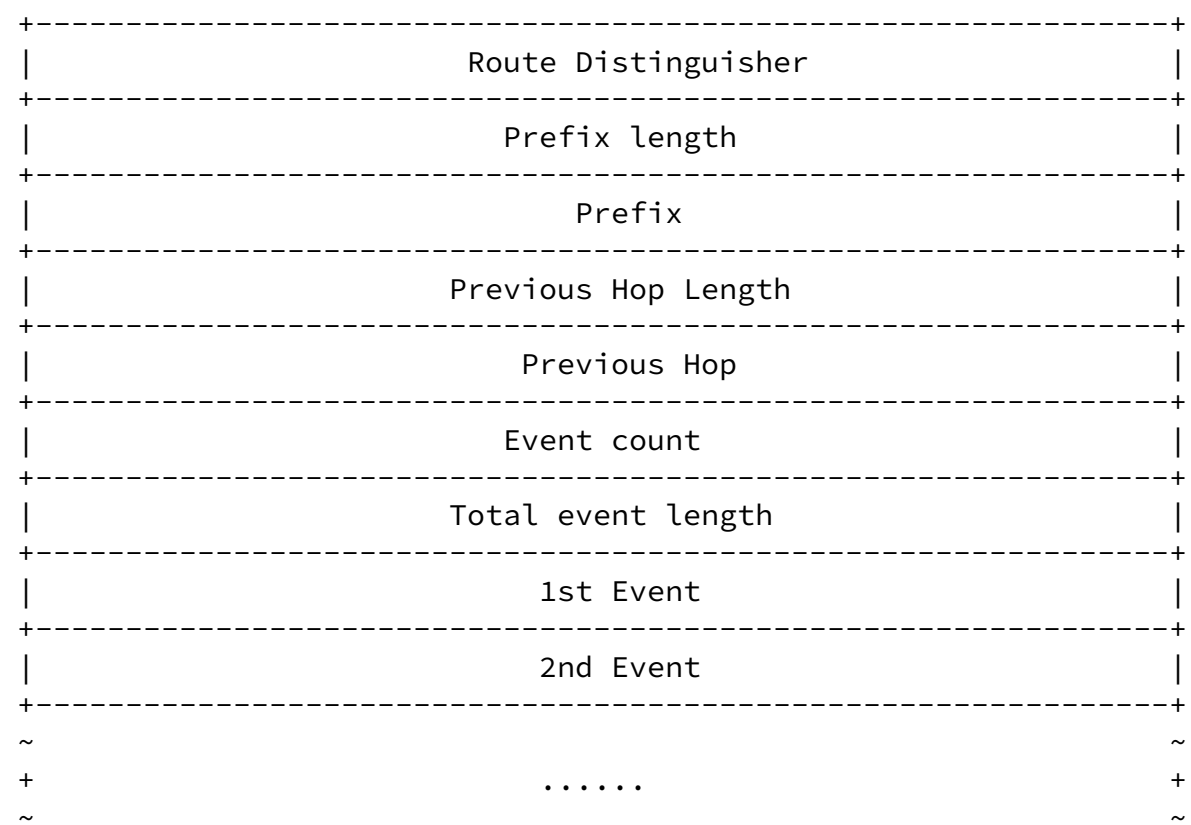
The new defined message type is indicated in the Message Type field of the BMP common header.

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:



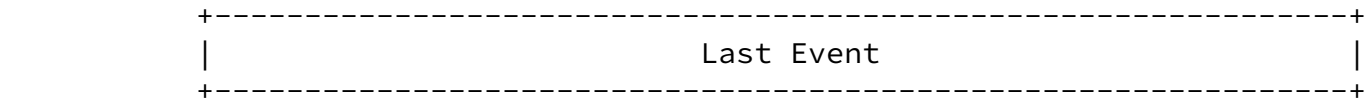
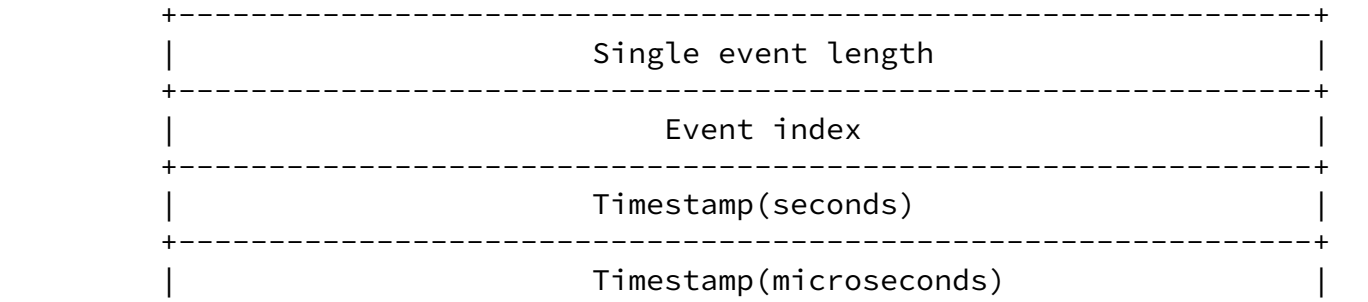


Figure 1: Route Policy and Attribute Trace Message format

- Route Distinguisher (8 Bytes): indicates the route distinguisher (RD) related to the route.
- Prefix Length (1 Byte): indicates the length of the prefix.
- Prefix (Variable): indicates the monitored prefix, with the length defined by Prefix Length field.
- Previous Hop Length (1 Byte): indicates the length of the following Previous Hop field. If the BGP peer ID of previous hop is IPv4, it is set to 4, and if the BGP peer ID of the previous hop is an IPv6, it is set to 16.
- Previous Hop (Variable): indicates the BGP peer ID where this route is learnt from. If the route is locally generated, then 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.



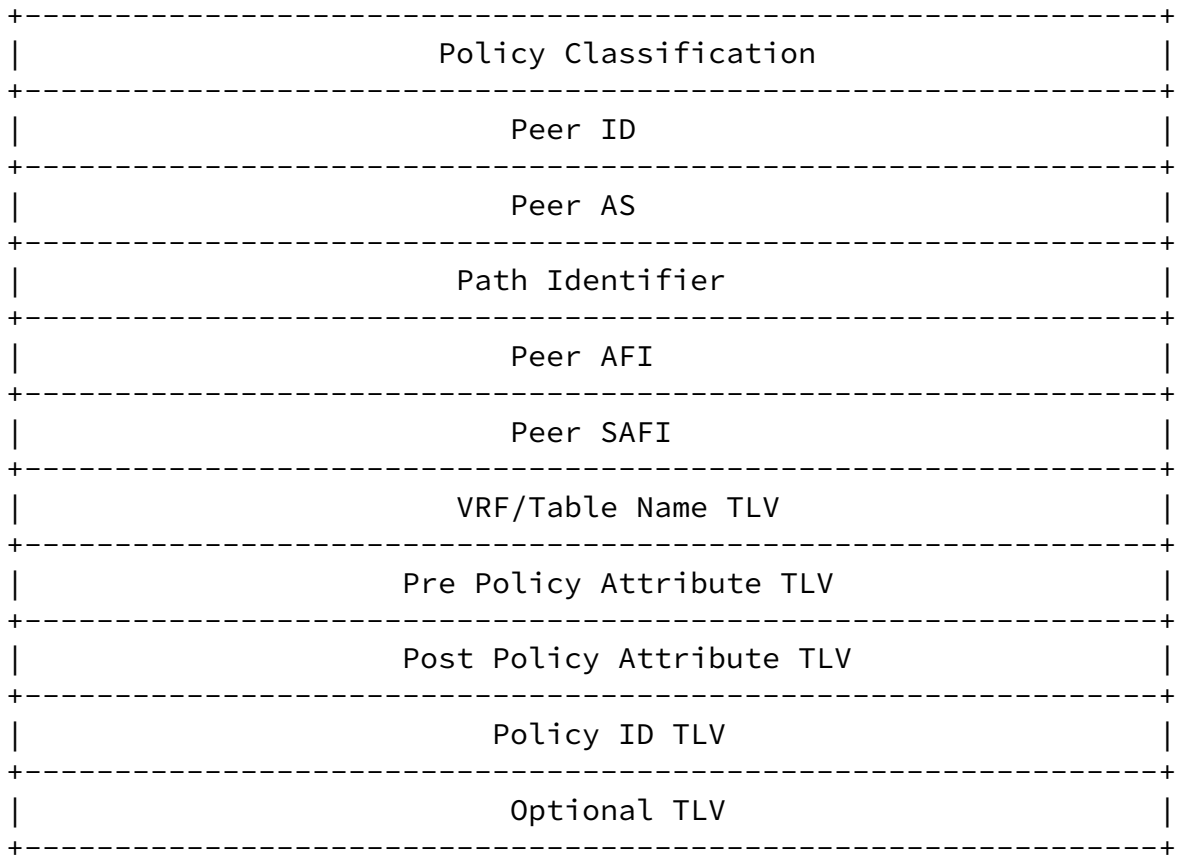


Figure 2: Event format

- 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).

- 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", "0010","0100" meaning Redistribution/Network/Aggregation policy, 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.

- Policy Classification (1 Byte): indicates the category of the policy. Currently 5 policy categories are defined: "0000" indicating the Inbound policy, "0001" indicating the Outbound policy, "0010" indicating the Redistribution policy (e.g., route import from other sources, like ISIS/OSPF), "0011" indicates the Route Leak policy (route leaking from the global routing table to a VRF or from a VRF to the global routing table, or between VRFs), "0100" indicates the Network policy (BGP network installment and advertisement), "0101" indicating the Aggregation policy. More categories can be defined.

-

Value	Policy Classification
00000000	Inbound policy
00000001	Outbound policy
00000010	Redistribution
00000011	Route Leak
00000100	Network
00000101	Aggregation

Table 1: Policy Classification

- 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.

- Peer AFI (2 Bytes)/Peer SAFI (1 Byte): 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 Unicast route is received from a CE router at the PE router through eBGP, an RD is attached to this IPv4 Unicast route and making it a VPNv4 route, and then this VPNv4 route is distributed to the RR. During this process, the AFI/SAFI information changes from IPv4 Unicast (1/1) to VPNv4 (1/128) at the inbound policy and outbound policy.

- VRF/Table Name TLV (Variable): indicates the VRF name or table name of the route. The format of the VRF/Table Name TLV is further defined in Figure 3. The VRF/Table Name TLV is non-optional.

- Pre-policy Attribute TLV (Variable): include the BGP route atttributes 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 atttributes 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.

- Policy ID 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 6. The Policy ID TLV is optional.

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

2.3.1. VRF/Table Name TLV

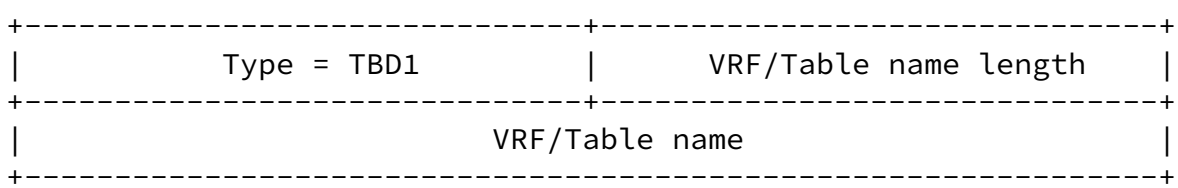


Figure 3: VRF/Table name TLV

- Type = TBD1 (2 Byte): indicates the type of VRF/Table name TLV.

- VRF/Table name length (2 Byte): indicates the length of the VRF/ Table name 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 varies for the same route under different events. For example, an IPv4 Unicast route is received from a CE router at the PE router through iBGP, an RD is attached to this IPv4 route (under VRF A) and making it a VPNv4 route, and then this VPNv4 route (under the Global routing table) is distributed to the RR. During the whole process, the VRF/Table name changes from VRF A to the Global routing Table name at the inbound event and outbound event.

2.3.2. Pre Policy Attribute TLV

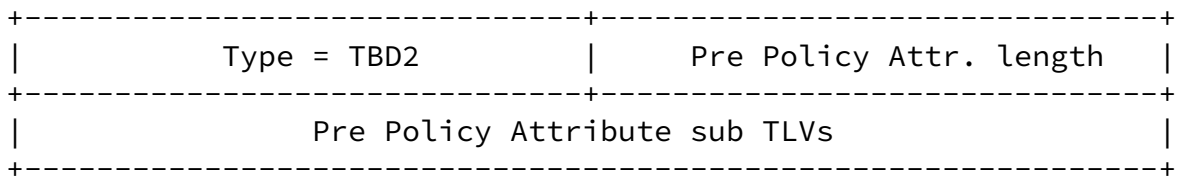


Figure 4: Pre Policy Attribute TLV

- Type = TBD2 (2 Byte): indicates the type of 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.

2.3.3. Post Policy Attribute TLV

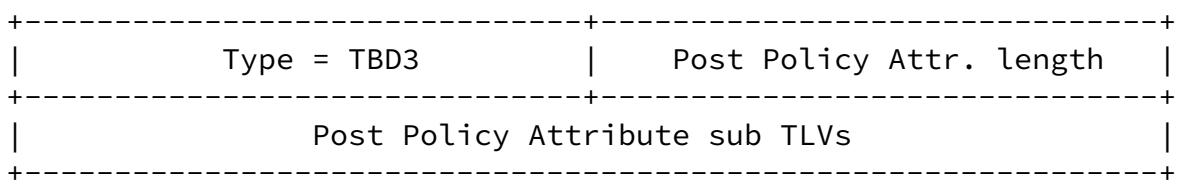


Figure 5: Post Policy Attribute TLV

- Type = TBD3 (2 Byte): indicates the type of 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. Policy ID TLV

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

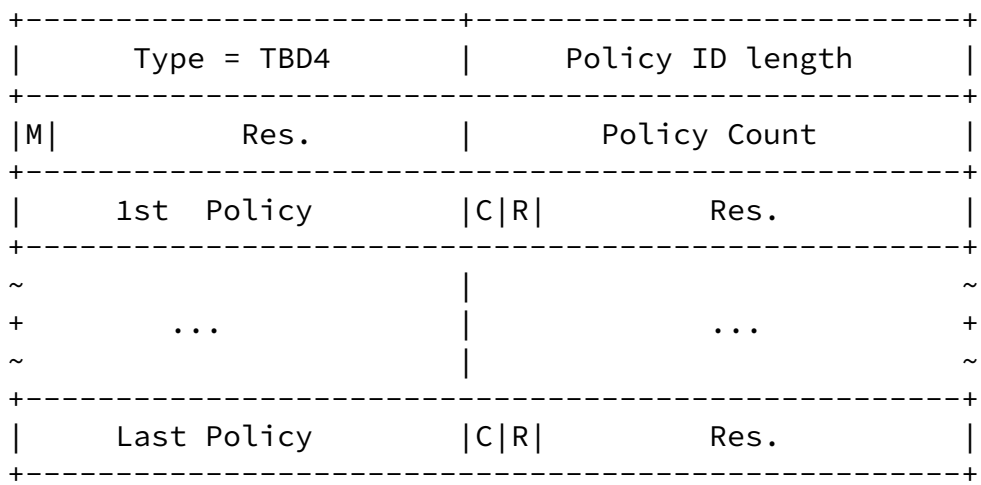


Figure 6: Policy ID TLV

Considering the chaining and recursion of polices and policy items, the Policy ID TLV is defined as follows.

- Type = TBD4 (2 Byte): indicates the type of Policy ID TLV.

- Policy ID length (2 Byte): indicates the length of the Policy ID value field that follows it. The Policy ID value field includes the Reserved bits, the Flag bits, Policy Count field, and Policy field.

- Flag bit M (1 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 elsewise. The remaining 7 bits are reserved for future extension.

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

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

- Flag bit C (1 bit): indicates if the next subsequent policy has chaining relationship to the current policy. "1" means it's

chaining relationship and "0" means elsewise. For the flag byte following the Last Policy field, the C bit SHALL be set to "0".

- Flag bit R (1 bit): indicates if the next subsequent policy has recursioning relationship to the current policy. "1" means it's recursioning relationship and "0" means elsewise. For the flag byte following the Last Policy field, the R bit SHALL be set to "0".

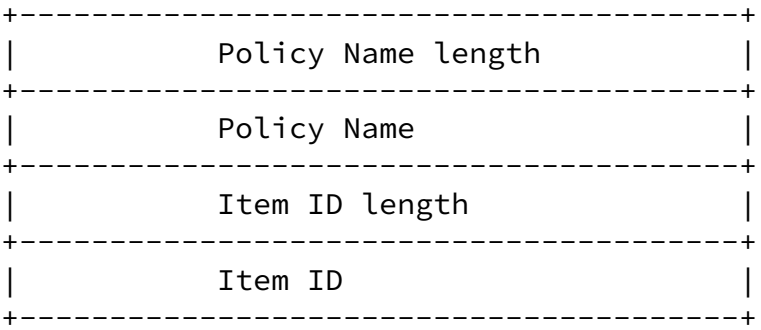
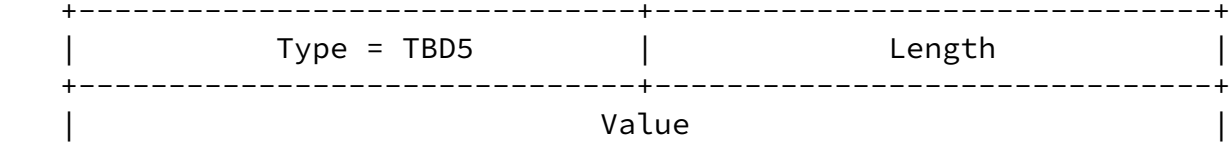


Figure 7: Policy field format

The Policy ID field consists of the Route Policy Name and the Route Policy Item ID. The Policy name and Item ID are in the format of ASCII string, the length of both fields are indicated by the Policy Name length (2 Bytes) and Item length (1 Byte) fields, respectively.

2.3.5. Optional TLV



+--+

Figure 8: Optional TLV

The Optional TLV remains to be defined. One or more Optional TLV types can be defined. One or more Optional TLVs can be used.

One possible way of utilizing the Optional TLV is to define a string Type TLV. The String Type TLV allows flexible textual expression of user-specific information without requiring structural format. Some examples:

- The Policy ID TLV is defined as optional, considering that users may don't want detailed information about the policy but only the result and/or the reasons. Using a string type TLV, one may express "Route rejected due to inbound filtering". However, such expression still requires the tracking of policy processing in realtime, it's just another form of tracking representation to the BMP server and the user.

- Another possible application is for route leak detection. One may express the business relations as "P2C", "P2P" and so on, with the inbound filtering event or the outbound filtering event. Detailed usage is discussed in Section 4.2.

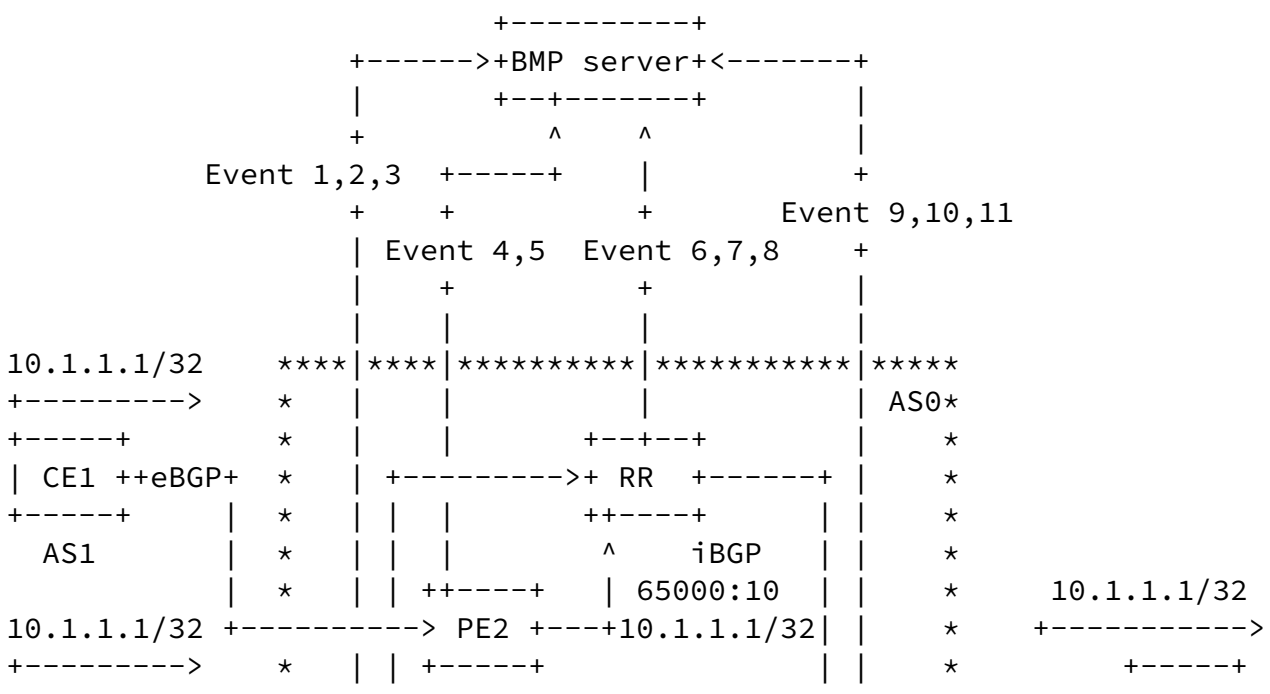
3. Implementation Considerations

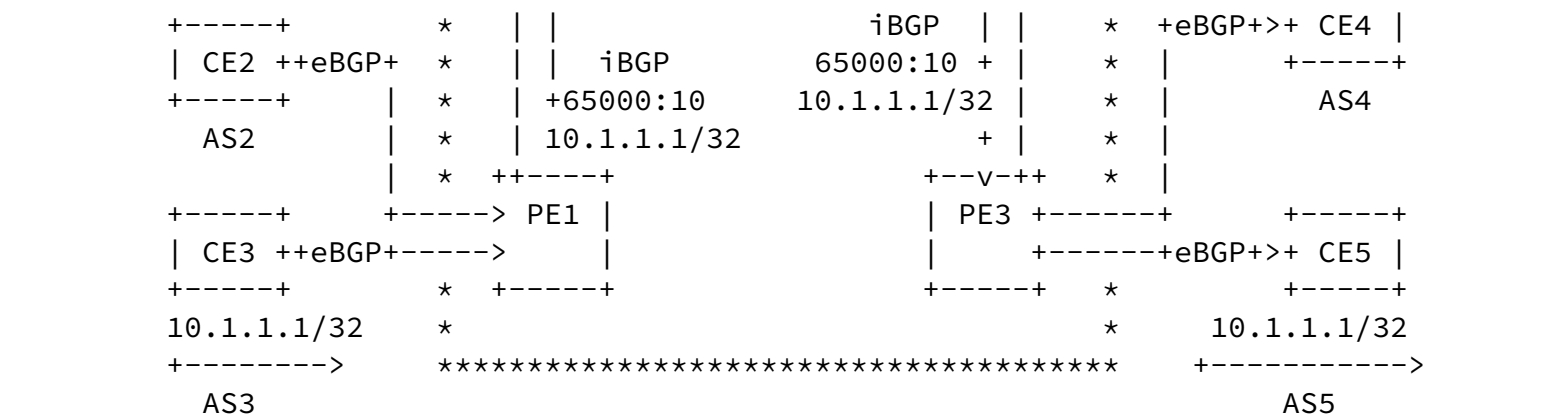
Considering the data amount of monitoring the route and policy trace of all routes from all BMP clients, users MAY trigger the monitoring at any user-specific time. Users MAY configure locally at the BMP client to monitor only user-specific routes or all the routes. In addition, users MAY configure locally 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. Implementation Example

4.1. Route Distribution Tracking





AS3 AS5

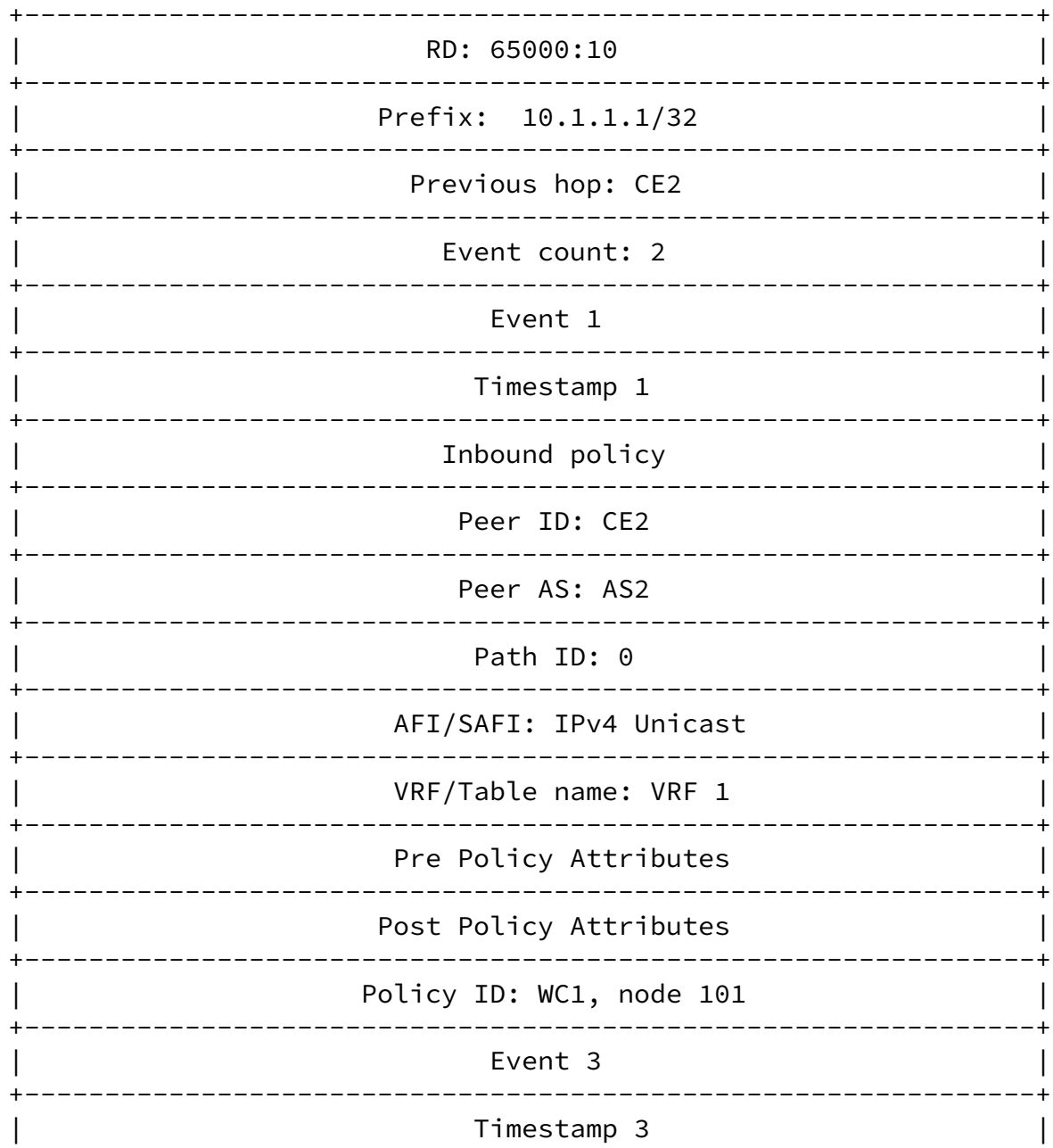
Figure 9: Route Policy and Attribute Trace record implementation example

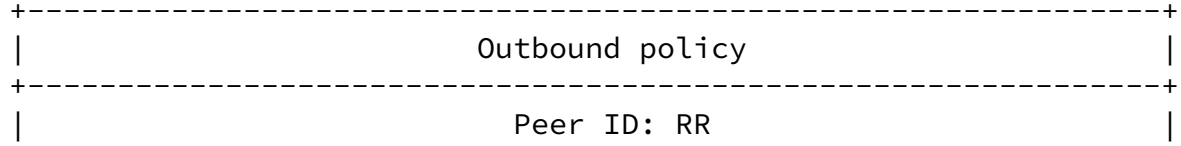
We take the network shown in Figure 9 as an example to show how to use Route Policy and Attribute Trace Messages to recover the footprint of the route propagation. Notice that only basic events required for footprint recovery are illustrated here. Also notice that the event index shown in Figure 9, 10, 11 are for illustration purpose, and may not reflect the actual indexing.

Suppose a prefix 10.1.1.1/32 is sent from both CE2 and CE3 to PE1 through eBGP peering, PE1 processes the two Update messages through 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 10.1.1.1/32 is received from both PE1 and PE2 in the form of VPNv4 route. RR selects the route from PE1 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).





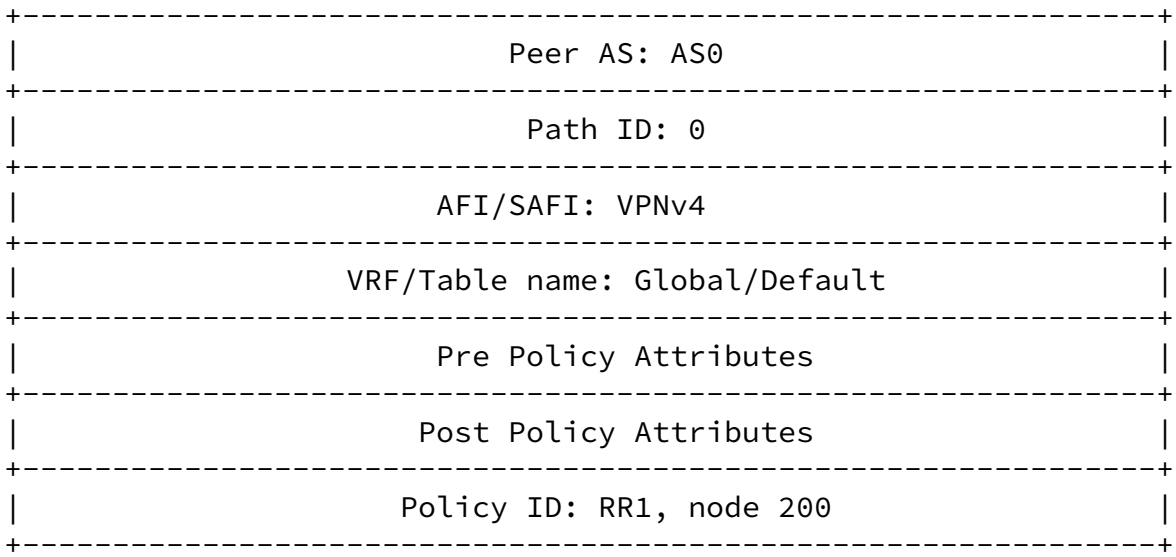


Figure 10: Event 1,3 data view

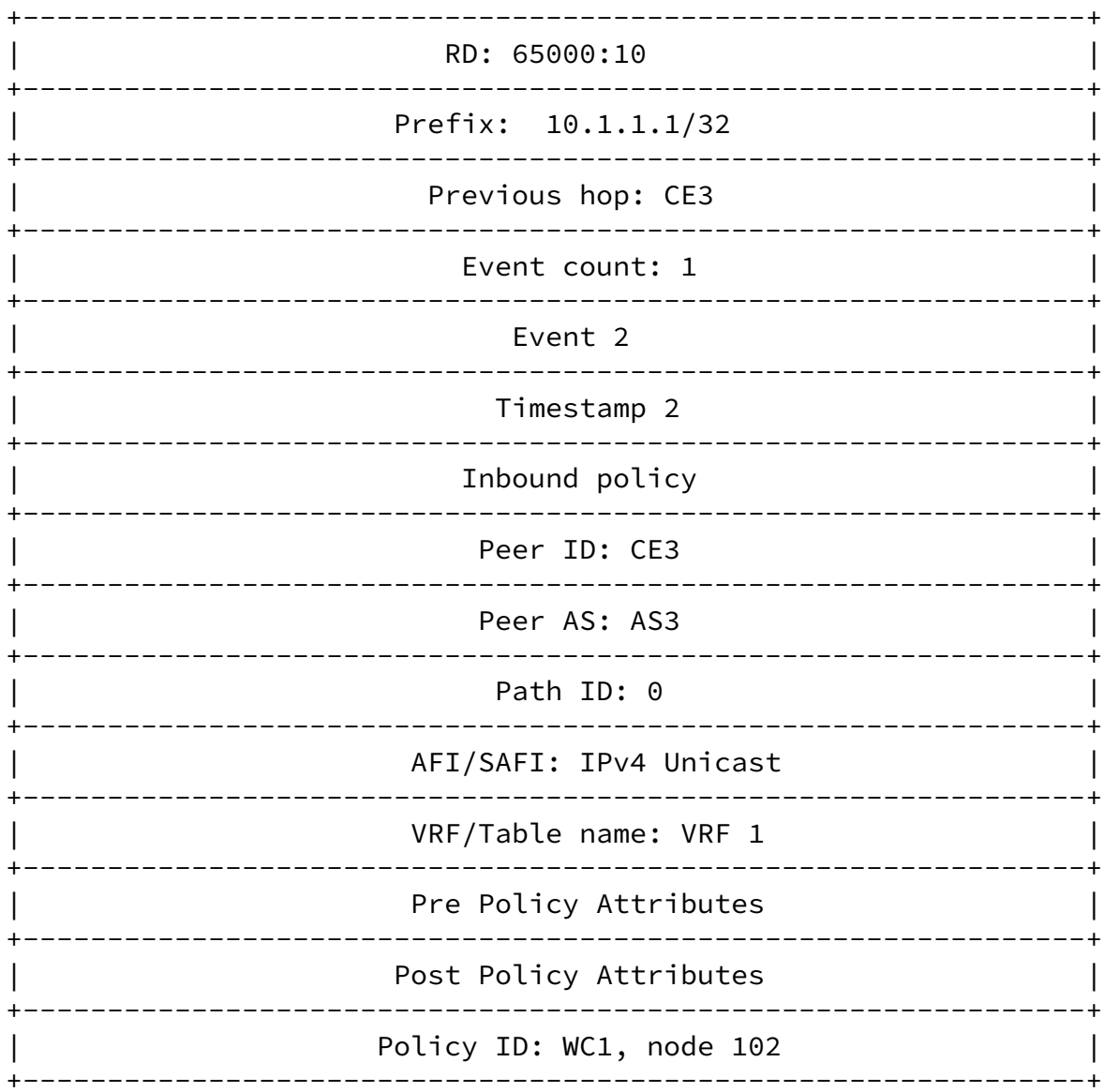


Figure 11: Event 2 data view

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.

4.2. Route Leak Detection

Reusing Figure 9, the Optional TLV of the RoFT Message can be utilized to carry user-specific strings. We present a route leak detection example here.

Suppose, a route leak happens (10.1.1.1/32: AS2 --> AS0 --> AS4). The Bussiness relationships between ASes are shown in Table 2.

Neighbor ASes	Bussiness Relationship
AS 1 : AS 0	P2C
AS 2 : AS 0	P2C
AS 3 : AS 0	P2C
AS 0 : AS 4	C2P
AS 0 : AS 5	P2C

Table 2: Bussiness Relationship

To detect the route leak, the BMP server analyzes the events with bussiness relationship information reported from the ingress device and egress device of AS0 (regarding a specific route)). In this example, regarding 10.1.1.1/32, data from PC1 and PE3 are analized. The bussiness relationship can be expressed in strings, such as "P2C" or "P2P". At PE1, when 10.1.1.1/32 is received from CE2 and going through the inbound policy, PE1 uses the Optional TLV (more specifically the String Type TLV) to carry the text "Bussiness Relationship: P2C" in the Inound Policy event. On the other hand, at PE3, when 10.1.1.1/32 goes through the outbound policy and then sent to CE4, PE3 adds the "Bussiness Relationship: P2C", using the Optional TLV, in the Outbound Policy event. More specifically, the format of the above mentioned two events are listed in Figure 12 (Event 1) and Figure 13 (Event 10), respecitvely.

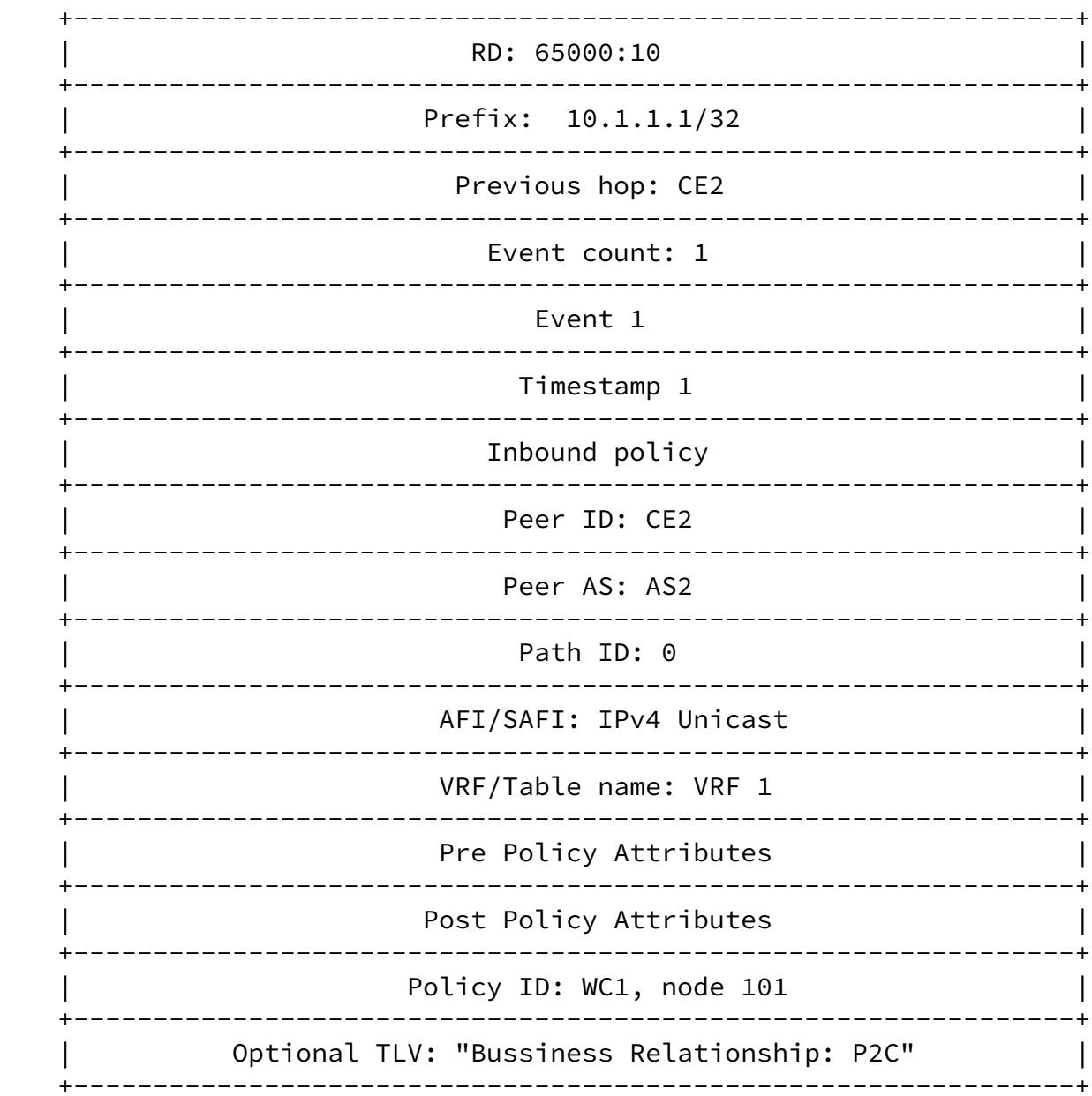
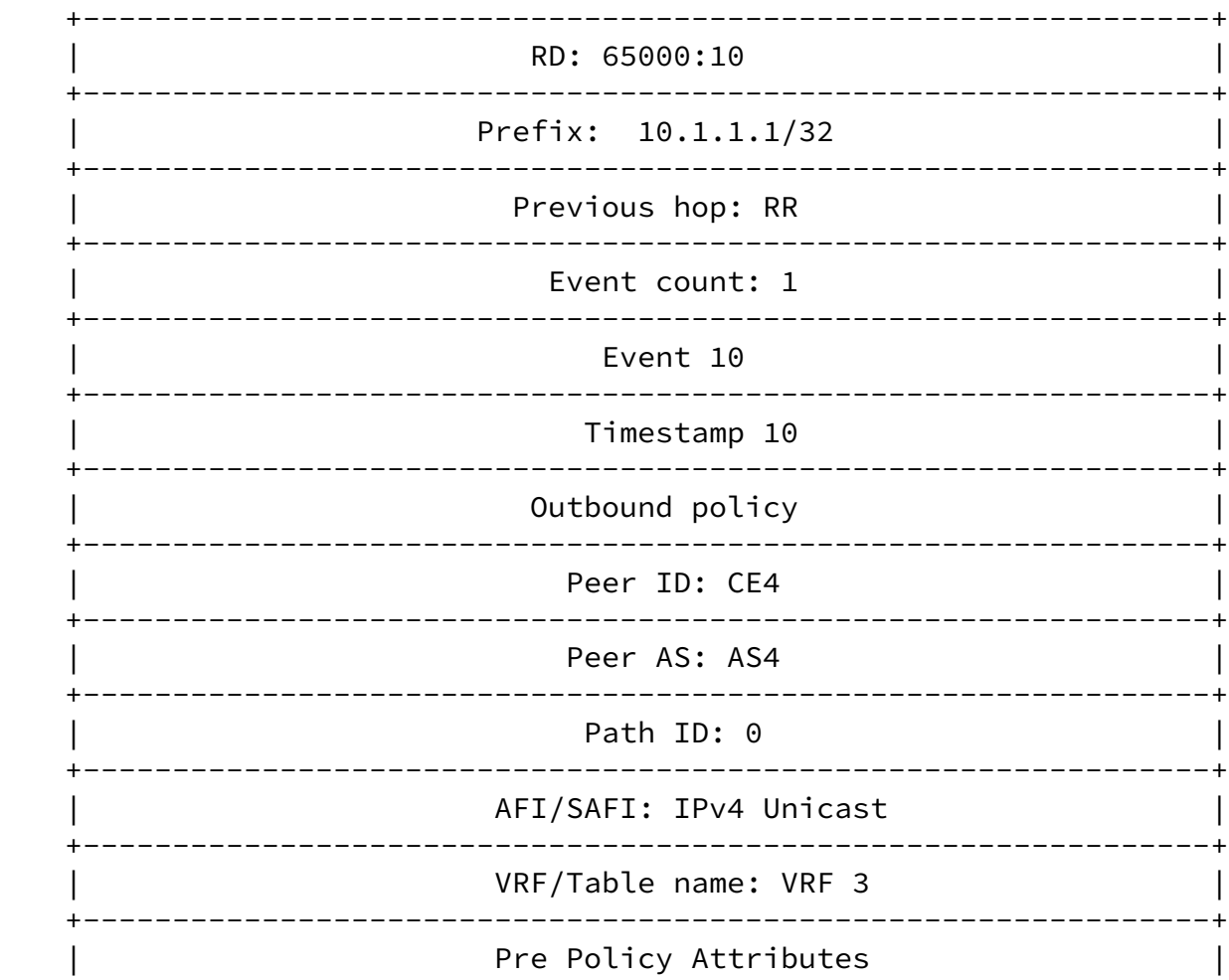

RD: 65000:10
Prefix: 10.1.1.1/32
Previous hop: CE2
Event count: 1
Event 1
Timestamp 1
Inbound policy
Peer ID: CE2
Peer AS: AS2
Path ID: 0
AFI/SAFI: IPv4 Unicast
VRF/Table name: VRF 1
Pre Policy Attributes
Post Policy Attributes
Policy ID: WC1, node 101
Optional TLV: "Bussiness Relationship: P2C"

Figure 12: Event 1 data view



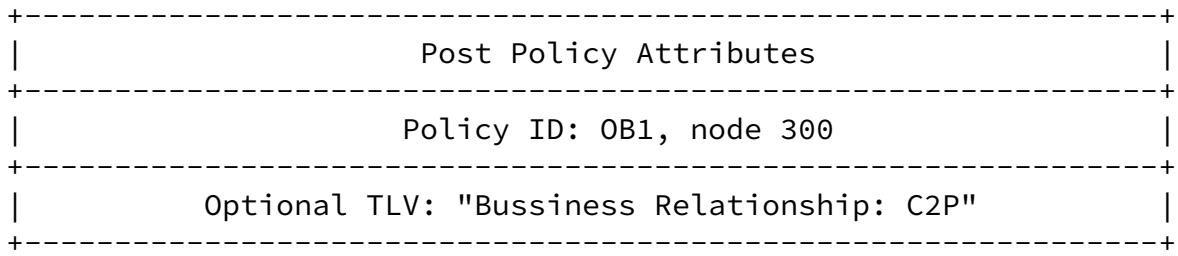


Figure 13: Event 10 data view

The BMP server can use the two Optional TLVs from Event 1 and Event 10 to detect the route leak. What's more, the repsonsible configurations are directly shown in the two events, i.e., the Inbound policy at PE1: "Policy ID: WC1, node 101", the Outbound policy at PE3: "Policy ID: OB1, node 300". No need to correlate with other data sources, the user can detect the leak and figure out the root cause.

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-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.

[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