

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2019

Y. Gu
Huawei
J. Chen
Tencent
P. Mi
S. Zhuang
Z. Li
Huawei
July 02, 2018

VPN Label Monitoring Using BMP
draft-gu-grow-bmp-vpn-label-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 method of collecting the VPN label using BMP, as well as an implementation example.

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 3, 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
2. Extension of BMP Peer Up Message	4
3. Operation	5
4. Acknowledgements	7
5. IANA Considerations	7
6. Security Considerations	7
7. Normative References	7
Authors' Addresses	8

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] 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.

The BGP Monitoring Protocol (BMP) [RFC7854] has been proposed around 2006 to monitor the BGP routing information, which includes the monitoring of BGP peer status, BGP route update, and BGP route statistics. BMP is realized through setting up the TCP session between the monitored BGP device and the BMP monitoring station, and then periodic/event-triggered messages are sent unidirectionally from the monitored device to the BMP monitoring station. Before BMP was introduced, such information could be only obtained through manual query, such as screen scraping. The introduction of BMP greatly improves the BGP routing monitoring efficiency without interrupting or interfering the ongoing services.

Currently, BMP is mainly utilized to monitor the public BGP routes. There are also cases that the VPN (Virtual Private Network) route/label information is needed. For example, for the purpose of Traffic Engineering (TE), the network operator may insert explicit routes, subject to certain constraints or optimization criteria, into related routers with high local preferences so that these routes will be selected and installed into the routing table. Under the VPN environment, the VPN route labels should be collected from the devices, and be distributed back jointly with the explicit routes to the devices, so that the devices can use the VPN labels to correlate the received routes with the appropriate VRFs (VPN Routing and Forwarding tables). The collection and distribution of such labels could be done by an SDN (Software Defined Network) controller, or a route monitoring station equipped with the traffic optimization module.

The VPN routes between CE (Customer Edge) and PE (Provider Edge) can be monitored by BMP using the "RD Instance Peer Type". However, such VPN routes between CE and PE do not include the VPN labels, since labels are allocated at the PE side. On the other hand, the labeled VPN routes are exchanged between PE and PE, which could have been monitored by BMP but are currently not implemented due to the massive data exchange between the monitored devices and the BMP monitoring station. An existing method to collect the VPN route label, considering the L3VPN scenario, is by setting up BGP VPNv4 peering relationship between the monitored device and the monitoring station/controller. The label information is then extracted from the collected VPN-IPv4 routes, carried by the BGP NLRI (Network Layer Reachability Information). However, there are several shortcomings of collecting the VPN label using this method.

- o The VPN labels, instead of the VPNv4 routes, are the necessary information for fulfilling the traffic optimization purpose. Thus, extracting the label from the VPNv4 route requires extra work compared with directly collecting the label information alone.
- o The same VPN label is sometimes used for several VPNv4 routes. 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 interface per label, and so on. For example, in the Multi-AS VPN case, the redistribution of labeled VPN-IPv4 routes from one AS to another can be realized through setting up the EBGp peering between ASBRs (Autonomous System Border Routers) of different ASes. 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 per label allocation is adopted.

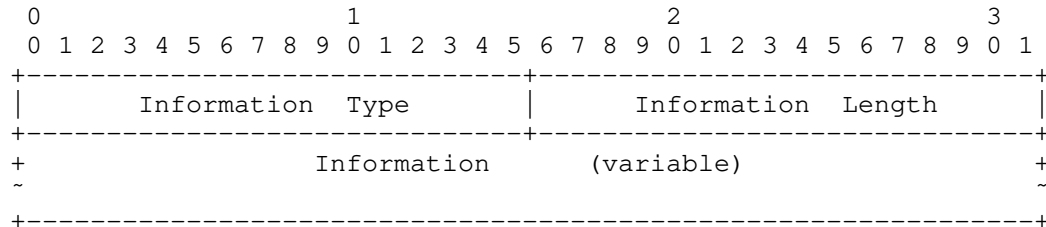
As a result, repeatedly reporting the same label for several routes wastes network resources.

- o The VPN label changes are typically less dynamic compared with the time-varying 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 VPN label independently than extracting it from the collected VPNv4 routes. In this document, we propose a method to utilize BMP to monitor the VPN label. In Section 2, the VPN label is defined to be encapsulated in the BMP Peer Up Notification message, and in Section 3, a specific implementation example is provided to show case the usage of the collected VPN label.

2. Extension of BMP Peer Up Message

The Peer Up message of BMP, defined in [RFC7854], is used to indicate the come-up of a peering session. The VPN route label can be carried in the Peer Up message and reported to the BMP monitoring station in the TLV format. The Information TLV defined in [RFC7854] can be used to encode the label, and new Information Types are defined. Each Information TLV contains at most one label, and one or more Information TLVs can be included in the Peer Up Notification when necessary.



- o Information Type (2 bytes): indicates the type of the Information TLV. Depending on the label allocation method, the following new types are defined:
 - * Type = TBD1: VPN Label, allocated per VRF per label.
 - * Type = TBD2: VPN Label, allocated per interface per label.
 - * Type = TBD3: VPN Label, allocated per route per label.
 - * Type = TBD4: VPN Label, allocated per next hop per label.

- o Information Length (2 bytes): indicates the length of the following Information field, in bytes.
- o Information (variable): specifies the Label information according to the Information Type.
 - * If the Information Type is VPN Label, allocated per VRF per label, the Information field should be the VPN label (20 bits), padded with zeros to 24 bits (3 bytes). The corresponding Length field should be set to 3.
 - * If the Information Type is VPN Label, allocated per interface per label, the Information field should include the VPN label (20 bits label and 4 bits zero padding) and the corresponding interface address, with the total length specified in the Information Length field. One label and one interface address is allowed for each Information TLV.
 - * If the Information Type is VPN Label, allocated per route per label, the Information includes the VPN label (20 bits label and 4 bits zero padding) and the corresponding route prefix, with the total length specified in the Information Length field. The prefix should be in VPNv4 address format. One label and one prefix is allowed for each Information TLV.
 - * If the Information Type is VPN Label, allocated per next hop per label, the Information should include the VPN label (20 bits label and 4 bits zero padding) and the corresponding next hop address, with the total length specified in the Information Length field. One label and one next hop address is allowed for each Information TLV.

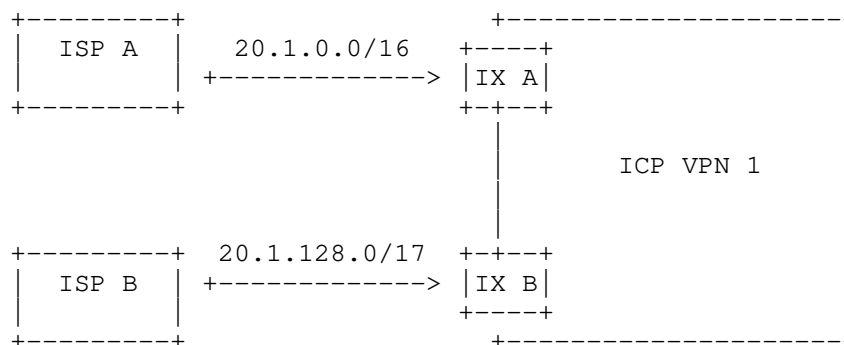
Considering the per VRF per label allocation, instead of extracting this same label information from all the monitored VPNv4 routes, it can be reported only once to save both device and network resources. Similarly, for the per interface per label and per next hop per label, label reporting frequencies can be reduced compared with the VPNv4 routes monitoring. Even for the per route per label case, reporting only the label information can be immune from the update of route changes, and reduce the reported information size.

3. Operation

In this section, we use an example of traffic optimization application to more specifically explain how the BMP VPN label collection functions. An Internet content provider (ICP) may own a Backbone network as the DCI (Data Center Interconnection) and Internet access solutions. Such backbone network may implement

different VPNs as the bearer networks for different services, and the granularity depends on specific service requirement. Each VPN, piggybacking on the backbone network, may connect to the Internet through other ISPs' (Internet Service Providers') networks. Different Internet Exchange (IX) devices are deployed for the Internet traffic exchange between the ICP and different ISPs.

Suppose two ISPs are considered in this example, ISP A and ISP B, as shown in the following figure. The ICP backbone network, implements VPN 1 for a specific service. This VPN exchanges Internet traffic with ISP A and ISP B through IX device A and IX device B, respectively. Prefixes are advertised from ISP A (considered as CE A of VPN 1) and ISP B (CE B) to the IX A (PE A) and IX B (PE B), respectively. Consider the case that ISP B advertises a more specific prefix (20.1.128.0/17) than ISP A (20.1.0.0/16). Both routes would be learnt by the PE devices of VPN 1, and be installed on both PE A's and PE B's routing tables. Now suppose there's a packet with destination 20.1.128.1, then according to the Longest prefix match (LPM) rule, PE B will be used as the ICP's exit for this packet. Similarly, more traffic with such prefixes may choose to exit the ICP to other ISPs through PE B, while PE A is lightly burdened, which leads to unbalanced traffic load and even traffic congestion at PE B.



The above mentioned issue can be solved as follows. Through traffic monitoring, the SDN controller can reoptimize the traffic load through explicit routes installation into PE A and PE B. The Next Hop field is indicated explicitly by the controller for the routes that need to be adjusted. For example, for the destination prefix 20.1.128.1, its next hop in the explicit route sent to PE A is set to the router's address in ISP A, while the next hop in the explicit route sent to PE B is set to PE A. Simultaneously, BMP is used to collect VPN 1's route labels from PE A (Label: 1000) and PE B (Label: 2000). Assume in this example, the labels are allocated per VRF per label, then Label 1000 is the label allocated to PE A for VRF 1, and Label 2000 is the label allocated to PE B for VRF 1. The explicit

routes distributed to PE A and PE B are specified in the following figures, respectively. After receiving the explicit route, PE A/B may use the label information to correlate the route to the correct VRF and then install it into VRF 1. Thus, part of the traffic may exit VPN 1 through PE A to balance the traffic load.

Dest. Addr./Mask	NH	Label	Local Pref.
20.1.128.0/17	ISP A	1000	200

Dest. Addr./Mask	NH	Label	Local Pref.
20.1.128.0/17	PE A	1000	200

4. Acknowledgements

TBD.

5. IANA Considerations

TBD.

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>>.
- [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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 18, 2019

Y. Gu
S. Zhuang
Z. Li
Huawei
July 17, 2018

Network Monitoring Protocol (NMP)
draft-gu-network-monitoring-protocol-00

Abstract

To evolve towards automated network OAM (Operations, administration and management), the monitoring of control plane protocols is a fundamental necessity. In this document, a network monitoring protocol (NMP) is proposed to provision the running status information of control plane protocols, e.g., IGP (Interior Gateway Protocol) and other protocols. By collecting the protocol monitoring data and reporting it to the NMP monitoring server in real-time, NMP can facilitate network troubleshooting. In this document, NMP for IGP troubleshooting are illustrated to showcase the necessity of NMP. IS-IS is used as the demonstration protocol, and the case of OSPF (Open Shortest Path First) and other control protocols will be elaborated in the future versions. The operations of NMP are described, and the NMP message types and message formats are defined in the document.

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 3, 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	3
1.1. Motivation	3
1.2. Overview	4
2. Terminology	5
3. Use Cases	5
3.1. IS-IS Adjacency Issues	5
3.2. Forwarding Path Disconnection	6
3.3. IS-IS LSP Synchronization Failure	6
4. NMP Message Format	7
4.1. Protocol Selection Options	7
4.2. Message Types	7
4.3. Message Format	8
4.3.1. Common Header	8
4.3.2. Per Adjacency Header	9
4.3.3. Initiation Message	10
4.3.4. Adjacency Status Change Notification	11
4.3.5. Statistic Report Message	12
4.3.6. IS-IS PDU Monitoring Message	14
4.3.7. Termination Message	14
5. IANA	15
6. Contributors	15
7. Acknowledgments	15
8. References	15
Authors' Addresses	17

1. Introduction

1.1. Motivation

The requirement for better network OAM approaches has been greatly driven by the network evolvement. The concept of network Telemetry has been proposed to meet the current and future OAM demands w.r.t., massive and real-time data storage, collection, process, exportion, and analysis, and an architectural framework of existing Telemetry approaches is introduced in [I-D.song-ntf]. Network Telemetry provides visibility to the network health conditions, and is beneficial for faster network troubleshooting, network OpEx (operating expenditure) reduction, and network optimization. Telemetry can be applied to the data plane, control plane and management plane. There have been various methods proposed for each plane:

- o Management plane: For example, SNMP (Simple Network Management Protocol) [RFC1157], NETCONF (Network Configuration Protocol) [RFC6241] and gNMI (gRPC Network Management Interface) [I-D.openconfig-rtgwg-gnmi-spec] are three typical widely adopted management plane Telemetry approaches. Various YANG modules are defined for network operational state retrieval and configuration management. Subscription to specific YANG datastore can be realized in combination with gRPC/NETCONF.
- o Data plane: For example, In-situ OAM (iOAM) [I-D.brockners-inband-oam-requirements] embeds an instruction header to the user data packets, and collects the requested data and adds it to the use packet at each network node along the forwarding path. Applications such as path verification, SLA (service-level agreement) assurance can be enabled with iOAM.
- o Control Plane: BGP monitoring protocol (BMP) [RFC7854] is proposed to monitor BGP sessions and intended to provide a convenient interface for obtaining BGP route views. Data collected using BMP can be further analyzed with big data platforms for network health condition visualization, diagnose and prediction applications.

The general idea of most Telemetry approaches is to collect various information from devices and export to the centralized server for further analysis, and thus providing more network insight. It should not be surprising that any future and even current Telemetry applications may require the fusion of data acquired from more than one single approach/one single plane. For example, for network troubleshooting purposes, it requires the collection of comprehensive information from devices, such system ID/router ID, interface status, PDUs (protocol data units), device/protocol statistics and so on.

Information such as system ID/router ID can be reported by management plane Telemetry approaches, while the protocol related data (especially PDUs) are more fit to be monitored using the control plane Telemetry. With rich information collected in real time at the centralized server, network issues can be localized faster and more accurately, and the root cause analysis can be also provided.

The conventional troubleshooting logic is to log in a faulty router, physically or through Telnet, and by using CLI to display related information/logs for fault source localization and further analysis. There are several concerns with the conventional troubleshooting methods:

1. It requires rich OAM experience for the OAM operator to know what information to check on the device, and the operation is complex;
2. In a multi-vendor network, it requires the understanding and familiarity of vendor specific operations and configurations;
3. Locating the fault source device could be non-trivial work, and is often realized through network-wide device-by-device check, which is both time-consuming and labor-consuming; and finally,
4. The acquisition of troubleshooting data can be difficult under some cases, e.g., when auto recovery is used.

This document proposes the Network Monitoring Protocols (NMP) to monitor the running status of control protocols, e.g., PDUs, protocol statistics and peer status, which have not been systematically covered by any other Telemetry approach, to facilitate network troubleshooting.

1.2. Overview

Like BMP, an NMP session is established between each monitored router (NMP client) and the NMP monitoring station (NMP server) through TCP connection. Information are collected directly from each monitored router and reported to the NMP server. The NMP message can be both periodic and event-triggered, depending on the message type.

IS-IS [RFC1195], as one of the most commonly adopted network layer protocols, builds the fundamental network connectivity of an autonomous system (AS). The disfunction of IS-IS, e.g., IS-IS neighbor down, route flapping, MTU mismatch, and so on, could lead to network-wide instability and service interruption. Thus, it is critical to keep track of the health condition of IS-IS, and the availability of information, related to IS-IS running status, is the fundamental requirement. In this document, typical network issues

are illustrated as the use cases of NMP for IS-IS to showcase the necessity of NMP. Then the operations and the message formats of NMP for IS-IS are defined. In this document IS-IS is used as the illustration protocol, and the case of OSPF and other control protocols will be included in the future version.

2. Terminology

IGP: Interior Gateway Protocol

IS-IS: Intermediate System to Intermediate System

NMP: Network Monitoring Protocol

IMP: Network Monitoring Protocol for IGP

BMP: BGP monitoring protocol

IIH: IS-IS Hello Packet

LSP: Link State Packet

CSNP: Complete Sequence Number Packet

NSNP: Partial Sequence Number Packet

3. Use Cases

We have identified several typical network issues due to IS-IS disfunction that are currently difficult to detect or localize. The usage of NMP is not limited to the solve the following listed issues.

3.1. IS-IS Adjacency Issues

IS-IS adjacency issues are identified as top network issues and may take hours to localize. The adjacency issues can be classified into two situations:

1. An existing established adjacency goes down;
2. An adjacency fails to be established.

In Case 1, the adjacency down can be caused by factors such as circuit down, hold timer expiration, device memory low, user configuration change, and so on. Case 2 can be caused by mismatch link MTU, mismatch authentication, mismatch area ID, system ID conflict, and so on. Typically, such adjacency failure events are logged/recorded in the device, but currently there is no real-time

report/alarm of such issue. The conventional troubleshooting process for adjacency issue is to find the faulty devices and then log in to check the logs or the IIH statistics for further analysis.

Using NMP, the IS-IS adjacency status: up, down and initial, is reported to the NMP server in real time, together with the possible recorded reasons. Then the NMP server can solve such issue in about minutes. For example, for an adjacency set up failure due to different authentications, the NMP server can recognize the difference by comparing the IIHs collected from both devices.

3.2. Forwarding Path Disconnection

The PING test can be used to test the reachability of a destination address. However, there are cases of disconnection that cannot be detected by PING. The PING result may return a connected path, but the forwarding of certain-sized packets always fails. This could be caused by factors, such as mismatched MTU values for devices along the path. It can be quite common since vendors have different understanding and configurations of MTU. There are methods proposed to discover the path MTU. For example, router's link MTU is conveyed in the MPLS LDP/RSVP-TE path set up signaling, and the path MTU is decided at the ingress or egress node[RFC3988] [RFC3209]. For IPv4 packets, by setting the DF flag bit of the outgoing packet, any device along the path with smaller MTU will drop the packet, and send back an ICMP Fragmentation Needed message containing its MTU, allowing the source to reduce the MTU. The process is repeated until the MTU is small enough to traverse the entire path without fragmentation[RFC1191]. Apparently, such method is too time-consuming.

Using NMP, each device can report its link MTU to the monitoring station directly. The mismatch can be recognized at the NMP server in seconds.

3.3. IS-IS LSP Synchronization Failure

It happens that two IS-IS neighbors fail to learn the LSPs sent from each other in the following two cases: in Case 1, the LSP fails to be received, and in Case 2, the LSP is received but the LSP information shown in the receiver's LSDB is not the same as the one sent from the transmitter (e.g., one or more prefixes missing, the LSP sequence number modified). Case 1 can be caused by link failure, similar to the adjacency down issue. In Case 2, the received LSP can be processed incorrectly due to hardware/software bugs. In fact, the LSDB synchronization issue is usually hard to localize once happens.

Using NMP, the NMP server can detect the failure by comparing the sent/received LSP statistics from the two neighbors. In the case that the received LSPs are improperly processed within the device, the NMP monitoring station can recognize the LSP synchronization failure by comparing the LSPs sent out from the two neighbors.

4. NMP Message Format

4.1. Protocol Selection Options

Regarding the NMP/IMP monitoring data exportion, BMP has been a good option. First of all, BMP serves similar purposes of NMP that reports routes, route statistics and peer status. In addition, BMP has already been implemented in major vendor devices and utilized by operator. Thus, we propose the following two options for the NMP data exportion.

- o Option 1: Extending BMP with new message types to carry NMP/IMP data: Reusing the BMP framework saves certain implementation cost for both vendors and operators. Besides, the monitoring data exportion of different routing protocols (e.g., BGP, ISIS, OSPF) can be unified.
- o Option 2: Defining NMP to carry NMP/IMP data: This option defines a brand new framework to carry protocol monitoring data, similar to BMP. Defining a new framework provides advantages such as more flexible and customized features for IGP and other protocols, since the monitoring data and troubleshooting of different protocols vary from one another.

In this document, we take Option 2 as the illustration example to define the NMP message types and message formats. The decision of the protocol selection may be further clarified in futures versions.

4.2. Message Types

The variety of IS-IS troubleshooting use cases requires a systematic information report of NMP, so that the NMP server or any third party analyzer could efficiently utilize the reported messages to localize and recover various network issues. We define NMP messages for IS-IS uses the following types:

- o Initiation Message: A message used for the monitored device to inform the NMP monitoring station of its capabilities, vendor, software version and so on. For example, the link MTU can be included within the message. The initiation message is sent once the TCP connection between the monitoring station and monitored

router is set up. During the monitoring session, any change of the initiation message could trigger an Initiation Message update.

- o Adjacency Status Change Notification Message: A message used to inform the monitoring station of the adjacency status change of the monitored device, i.e., from up to down, from down/initiation to up, with possible alarms/logs recorded in the device. This message notifies the NMP server of the ongoing IS-IS adjacency change event and possible reasons. If no reason is provided or the provided reason is not specific enough, the NMP server can further analyze the IS-IS PDU or the IS-IS statistics.
- o Statistic Report Message: A message used to report the statistics of the ongoing IS-IS process at the monitored device. For example, abnormal LSP count of the monitored device can be a sign of route flapping. This message can be sent periodically or event triggered. If sent periodically, the frequency can be configured by the operator depending on the monitoring requirement. If it's event triggered, it could be triggered by a counter/timer exceeding the threshold.
- o IS-IS PDU Monitoring Message: A message used to update the NMP server of any PDU sent from and received at the monitored device. For example, the IIHs collected from two neighbors can be used for analyzing the adjacency set up failure issue. The LSPs collected from two neighbors can be analyzed for the LSP synchronization issue.
- o Termination Message: A message for the monitored router to inform the monitoring station of why it is closing the NMP session. This message is sent when the monitoring session is to be closed.

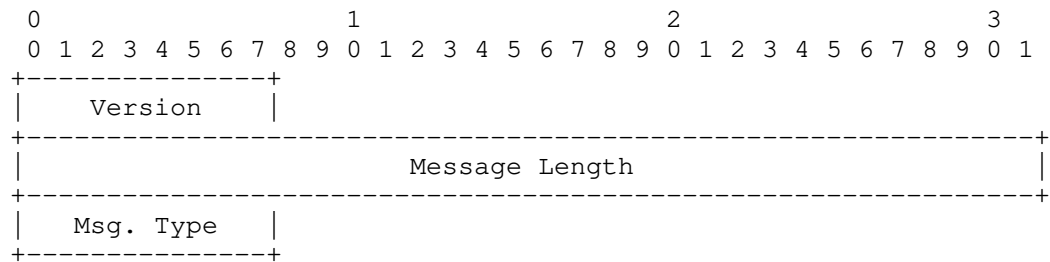
4.3. Message Format

4.3.1. Common Header

The common header is encapsulated in all NMP messages. It includes the Version, Message Length and Message Type fields.

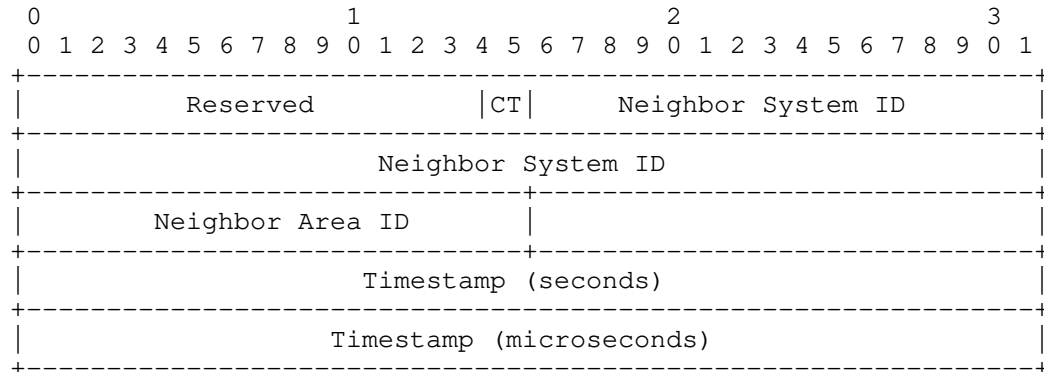
- o Version (1 byte): Indicates the NMP version and is set to '1' for all messages.
- o Message Length (4 bytes): Length of the message in bytes (including headers, data, and encapsulated messages, if any).
- o Message Type (1 byte): This indicates the type of the NMP message, which are listed as follows.

- * Type = 0: Initiation
- * Type = 1: Adjacency Status Change Notification
- * Type = 2: Statistic Report
- * Type = 3: IS-IS PDU Monitoring
- * Type = 4: Termination Message



4.3.2. Per Adjacency Header

Except the Initiation and Termination Message, all the rest messages are per adjacency based. Thus, a per adjacency header is defined as follows.



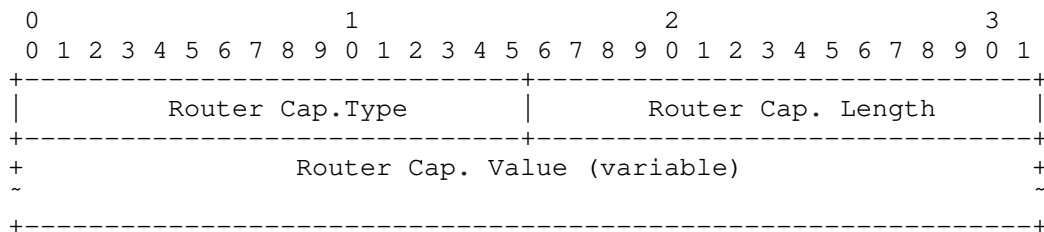
- o Adjacency Flag (2 bytes): The Circuit Type (2 bits) flag specifies if the router is an L1(01), L2(10), or L1/L2(11). If both bits are zeroes (00), the Per Adjacency Header SHALL be ignored. This configuration is used when the statistic is not per-adjacency based, e.g., when reporting the number of adjacencies.

- o Neighbor System ID (6 bytes): identifies the system ID of the remote router.
- o Neighbor Area ID (2 bytes): identifies the area ID of the remote router.
- o Timestamp (4 bytes): records the time when the message is sent/received, expressed in seconds and microseconds since midnight (zero hour), January 1, 1970 (UTC).

4.3.3. Initiation Message

The Initiation Message indicates the monitored router's capabilities, vendor, software version and so on. It consists of the Common Header and the Router Capability TLV. The Common Header can be followed by multiple Router Capability TLVs.

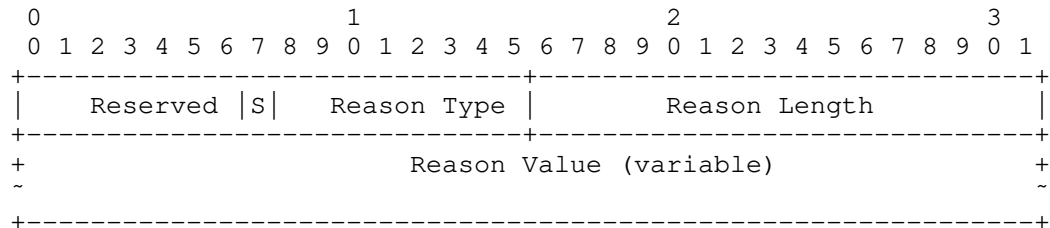
The Router Capability TLV is defined as follows.



- o Router Capability Type: provides the type of the router capability information. Currently defined types are:
 - * Type = 0: sysDescr. The corresponding Router Capability Value field should contain an ASCII string whose value MUST be set to be equal to the value of the sysDescr MIB-II [RFC1213] object.
 - * Type = 1: sysName. The corresponding Router Capability Value field should contain an ASCII string whose value MUST be set to be equal to the value of the sysName MIB-II [RFC1213] object.
 - * Type = 2: Local System ID. The corresponding Router Capability Value field SHALL indicate the router's System ID
 - * Type = 3: Link MTU. The corresponding Router Capability Value field SHALL indicate the router's link MTU.
 - * Type = 4: String. The corresponding Router Capability Value field contains a free-form UTF-8 string whose length is given by the Information Length field.

4.3.4. Adjacency Status Change Notification

The Adjacency Status Change Notification Message indicates an IS-IS adjacency status change: from up to down or from initiation/down to up. It consists of the Common Header, Per Adjacency Header and the Reason TLV. The Notification is triggered whenever the status changes. The Reason TLV is optional, and is defined as follows. More Reason types can be defined if necessary.



- o Reason Flags (1 byte): The S flag (1 bit) indicates if the Adjacency status is from up to down (set to 0) or from down/initial to up (set to 1). The rest bits of the Flag field are reserved. When the S flag is set to 1, the Reason Type SHALL be set to all zeroes (i.e., Type 0), the Reason Length fields SHALL be set to all zeroes, and the Reason Value field SHALL be set empty.
- o Reason Type (1 byte): indicates the possible reason that caused the adjacency status change. Currently defined types are:
 - * Type = 0: Adjacency Up. This type indicates the establishment of an adjacency. For this reason type, the S flag MUST be set to 1, indicating it's a adjacency-up event. There's no further reason to be provided. The reason Length field SHALL be set to all zeroes, and the Reason Value field SHALL be set empty.
 - * Type = 1: Circuit Down. For this data type, the S flag MUST be set to 0, indicating it's a adjacency-down event. The length field is set to all zeroes, and the value field is set empty.
 - * Type = 2: Memory Low. For this data type, the S flag MUST be set to 0, indicating it's a adjacency-down event. The length field is set to all zeroes, and the value field is set empty.
 - * Type = 3: Hold timer expired. For this data type, the S flag MUST be set to 0, indicating it's a adjacency-down event. The length field is set to all zeroes, and the value field is set empty.

- * Type = 4: String. For this data type, the S flag MUST be set to 0, indicating it's a adjacency-down event. The corresponding Reason Value field indicates the reason specified by the monitored router in a free-form UTF-8 string whose length is given by the Reason Length field.
- o Reason Length (2 bytes): indicates the length of the Reason Value field.
- o Reason Value (variable): includes the possible reason why the Adjacency is down.

4.3.5. Statistic Report Message

The Statistic Report Message reports the statistics of the parameters that are of interest to the operator. The message consists of the NMP Common Header, the Per Adjacency Header and the Statistic TLV. The message include both per-adjacency based statistics and non per-adjacency based statistics. For example, the received/sent LSP counts are per-adjacency based statistics, and the local LSP change times count and the number of established adjacencies are non per-adjacency based statistics. For the non per-adjacency based statistics, the CT Flag (2 bits) in the Per Adjacency Header MUST be set to 00. Upon receiving any message with CT flag set to 00, the Per Adjacency Header SHALL be ignored (the total length of the Per Adjacency Header is 18 bytes as defined in Section 3.2.2, and the message reading/analysis SHALL resume from the Statistic TLV part.

The Statistic TLV is defined as follows.

0	1	2	3
0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1
+-----+-----+-----+-----+			
Reserved T Statistic Type Statistic Length			
+-----+-----+-----+-----+			
Statistic Value			
+-----+-----+-----+-----+			

- o Statistic Flags (1 byte): provides information for the reported statistics.
 - * T flag (1 bit): indicates if the statistic is for the received-from direction (set to 1) or sent-to direction the neighbor (set to 0)
- o Statistic Type (1 byte): specifies the statistic type of the counter. Currently defined types are:

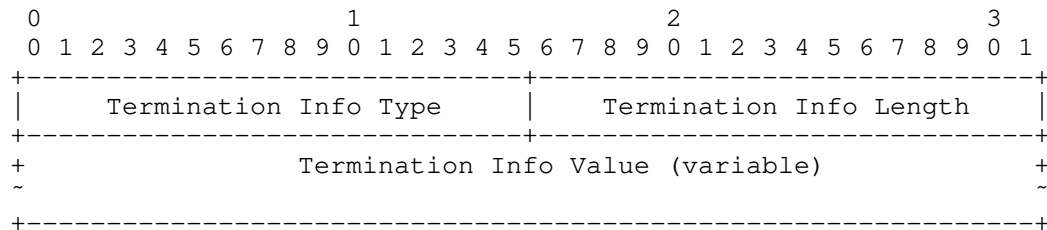
- * Type = 0: IIH count. The T flag indicates if it's a sent or received Hello PDU. It is a per-adjacency based statistic type, and the CT flag in the Per Adjacency Header MUST NOT be set to 00.
 - * Type = 1: Incorrect IIH received count. For this type, the T flag MUST be set to 1. It is a per-adjacency based statistic type, and the CT flag in the Per Adjacency Header MUST NOT be set to 00.
 - * Type = 2: LSP count. The T flag indicates if it's a sent or received LSP. It is a per-adjacency based statistic type, and the CT flag in the Per Adjacency Header MUST NOT be set to 00.
 - * Type = 3: Incorrect LSP received count. For this type, the T flag MUST be set to 1. It is a per-adjacency based statistic type, and the CT flag in the Per Adjacency Header MUST NOT be set to 00.
 - * Type = 4: Retransmitted LSP count. For this type, the T flag MUST be set to 0. It is a per-adjacency based statistic type, and the CT flag in the Per Adjacency Header MUST NOT be set to 00.
 - * Type = 5: CSNP count. The T flag indicates if it's a sent or received CSNP. It is a per-adjacency based statistic type, and the CT flag in the Per Adjacency Header MUST NOT be set to 00.
 - * Type = 6: PSNP count. The T flag indicates if it's a sent or received PSNP. It is a per-adjacency based statistic type, and the CT flag in the Per Adjacency Header MUST NOT be set to 00.
 - * Type = 7: Number of established adjacencies. It's a non per-adjacency based statistic type, and thus for the monitoring station to recognize this type, the CT flag in the Per Adjacency Header MUST be set to 00.
 - * Type = 8: LSP change time count. It's a non per-adjacency based statistic type, and thus for the monitoring station to recognize this type, the CT flag in the Per Adjacency Header MUST be set to 00.
- o Statistic Length (2 bytes): indicates the length of the Statistic Value field.
 - o Statistic Value (4 bytes): specifies the counter value, which is a non-negative integer.

4.3.6. IS-IS PDU Monitoring Message

The IS-IS PDU Monitoring Message is used to update the monitoring station of any PDU sent from and received at the monitored device per neighbor. Following the Common Header and the Per Adjacency Header is the IS-IS PDU. To tell whether it's a sent or received PDU, the monitoring station can analyze the source and destination addresses in the reported PDUs.

4.3.7. Termination Message

The Termination Message is sent when the NMP session is to be closed, and is used to indicate the termination reason to the monitoring station. The TCP session between the monitored router and the monitoring station SHALL be terminated upon receiving this message. It consists of the Common Header and the Termination Info TLVs, defined as follows.



- o Termination Info Type (2 bytes): Provides the termination reason type. Currently defined types are:
 - * Type = 0: Unknown. This reason type specifies that the NMP session is closed for an unknown or unspecified reason. For this data type, the length field is filled with all zeroes, and the value field is set empty.
 - * Type = 1: Memory Low. This reason indicates that the monitored router lacks resources for the NMP session. For this data type, the length field is filled with all zeroes, and the value field is set empty.
 - * Type = 2: Administratively Closed. This reason specifies that the session is closed due to administrative reasons. The corresponding Termination Info Value field may include more details about the reason expressed in a free-form UTF-8 string whose length is given by the Termination Info Length field.
 - * Type = 3: String. The corresponding Termination Info Value field may include details about the reason expressed in a free-

form UTF-8 string whose length is given by the Termination Info Length field.

Termination Info Length (2 bytes): indicates the length of the Termination Info Reason Value field.

- o Termination Info Value (variable): includes more detailed reason for the session termination.

5. IANA

TBD

6. Contributors

TBD

7. Acknowledgments

TBD

8. References

[I-D.brockners-inband-oam-requirements]

Brockners, F., Bhandari, S., Dara, S., Pignataro, C., Gredler, H., Leddy, J., Youell, S., Mozes, D., Mizrahi, T., <>, P., and r. remy@barefootnetworks.com, "Requirements for In-situ OAM", draft-brockners-inband-oam-requirements-03 (work in progress), March 2017.

[I-D.ietf-netconf-yang-push]

Clemm, A., Voit, E., Prieto, A., Tripathy, A., Nilsen-Nygaard, E., Bierman, A., and B. Lengyel, "YANG Datastore Subscription", draft-ietf-netconf-yang-push-17 (work in progress), July 2018.

[I-D.openconfig-rtgwg-gnmi-spec]

Shakir, R., Shaikh, A., Borman, P., Hines, M., Lebsack, C., and C. Morrow, "gRPC Network Management Interface (gNMI)", draft-openconfig-rtgwg-gnmi-spec-01 (work in progress), March 2018.

[I-D.song-ntf]

Song, H., Zhou, T., Li, Z., Fioccola, G., Li, Z., Martinez-Julia, P., Ciavaglia, L., and A. Wang, "Toward a Network Telemetry Framework", draft-song-ntf-02 (work in progress), July 2018.

- [RFC1157] Case, J., Fedor, M., Schoffstall, M., and J. Davin, "Simple Network Management Protocol (SNMP)", RFC 1157, DOI 10.17487/RFC1157, May 1990, <<https://www.rfc-editor.org/info/rfc1157>>.
- [RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <<https://www.rfc-editor.org/info/rfc1191>>.
- [RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <<https://www.rfc-editor.org/info/rfc1195>>.
- [RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets: MIB-II", STD 17, RFC 1213, DOI 10.17487/RFC1213, March 1991, <<https://www.rfc-editor.org/info/rfc1213>>.
- [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>>.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <<https://www.rfc-editor.org/info/rfc3209>>.
- [RFC3988] Black, B. and K. Kompella, "Maximum Transmission Unit Signalling Extensions for the Label Distribution Protocol", RFC 3988, DOI 10.17487/RFC3988, January 2005, <<https://www.rfc-editor.org/info/rfc3988>>.
- [RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011, <<https://www.rfc-editor.org/info/rfc6241>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<https://www.rfc-editor.org/info/rfc7752>>.
- [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
156 Beiqing Road
Beijing, 100095
P.R. China

Email: guyunan@huawei.com

Shunwan Zhuang
Huawei
156 Beiqing Road
Beijing, 100095
P.R. China

Email: zhuangshunwan@huawei.com

Zhenbin Li
Huawei
156 Beiqing Road
Beijing, 100095
P.R. China

Email: lizhenbin@huawei.com

Global Routing Operations
Internet-Draft
Updates: 7854 (if approved)
Intended status: Standards Track
Expires: February 6, 2020

T. Evens
S. Bayraktar
Cisco Systems
P. Lucente
NTT Communications
P. Mi
Tencent
S. Zhuang
Huawei
August 5, 2019

Support for Adj-RIB-Out in BGP Monitoring Protocol (BMP)
draft-ietf-grow-bmp-adj-rib-out-07

Abstract

The BGP Monitoring Protocol (BMP) defines access to only the Adj-RIB-In Routing Information Bases (RIBs). This document updates the BGP Monitoring Protocol (BMP) RFC 7854 by adding access to the Adj-RIB-Out RIBs. It adds a new flag to the peer header to distinguish Adj-RIB-In and Adj-RIB-Out.

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 February 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. Terminology	3
3. Definitions	3
4. Per-Peer Header	4
5. Adj-RIB-Out	4
5.1. Post-Policy	4
5.2. Pre-Policy	5
6. BMP Messages	5
6.1. Route Monitoring and Route Mirroring	5
6.2. Statistics Report	5
6.3. Peer Down and Up Notifications	6
6.3.1. Peer Up Information	6
7. Other Considerations	6
7.1. Peer and Update Groups	7
8. Security Considerations	7
9. IANA Considerations	7
9.1. BMP Peer Flags	8
9.2. BMP Statistics Types	8
9.3. Peer Up Information TLV	8
10. References	9
10.1. Normative References	9
10.2. URIs	9
Acknowledgements	9
Contributors	9
Authors' Addresses	10

1. Introduction

BGP Monitoring Protocol (BMP) defines monitoring of the received (e.g., Adj-RIB-In) Routing Information Bases (RIBs) per peer. The Adj-RIB-In pre-policy conveys to a BMP receiver all RIB data before any policy has been applied. The Adj-RIB-In post-policy conveys to a BMP receiver all RIB data after policy filters and/or modifications have been applied. An example of pre-policy versus post-policy is when an inbound policy applies attribute modification or filters. Pre-policy would contain information prior to the inbound policy changes or filters of data. Post policy would convey the changed data or would not contain the filtered data.

Monitoring the received updates that the router received before any policy has been applied is the primary level of monitoring for most use-cases. Inbound policy validation and auditing is the primary use-case for enabling post-policy monitoring.

In order for a BMP receiver to receive any BGP data, the BMP sender (e.g., router) needs to have an established BGP peering session and actively be receiving updates for an Adj-RIB-In.

Being able to only monitor the Adj-RIB-In puts a restriction on what data is available to BMP receivers via BMP senders (e.g., routers). This is an issue when the receiving end of the BGP peer is not enabled for BMP or when it is not accessible for administrative reasons. For example, a service provider advertises prefixes to a customer, but the service provider cannot see what it advertises via BMP. Asking the customer to enable BMP and monitoring of the Adj-RIB-In is not feasible.

BGP Monitoring Protocol (BMP) RFC 7854 [RFC7854] only defines Adj-RIB-In being sent to BMP receivers. This document updates the peer header in section 4.2 of [RFC7854] by adding a new flag to distinguish Adj-RIB-In versus Adj-RIB-Out. BMP senders use the new flag to send either Adj-RIB-In or Adj-RIB-Out.

Adding Adj-RIB-Out provides the ability for a BMP sender to send to BMP receivers what it advertises to BGP peers, which can be used for outbound policy validation and to monitor routes that were advertised.

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

- o 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."
- o Pre-Policy Adj-RIB-Out: The result before applying the outbound policy to an Adj-RIB-Out. This normally would match what is in the local RIB.

- o **Post-Policy Adj-RIB-Out:** The result of applying outbound policy to an Adj-RIB-Out. This MUST convey to the BMP receiver what is actually transmitted to the peer.

4. Per-Peer Header

The per-peer header has the same structure and flags as defined in section 4.2 of [RFC7854] with the following O flag addition:

```

      0 1 2 3 4 5 6 7
      +---+---+---+---+
      |V|L|A|O| Resv |
      +---+---+---+---+

```

- o The O flag indicates Adj-RIB-In if set to 0 and Adj-RIB-Out if set to 1.

The existing flags are defined in section 4.2 of [RFC7854] and the remaining bits are reserved for future use. They MUST be transmitted as 0 and their values MUST be ignored on receipt.

When the O flag is set to 1, the following fields in the Per-Peer Header are redefined:

- o **Peer Address:** The remote IP address associated with the TCP session over which the encapsulated PDU is sent.
- o **Peer AS:** The Autonomous System number of the peer to which the encapsulated PDU is sent.
- o **Peer BGP ID:** The BGP Identifier of the peer to which the encapsulated PDU is sent.
- o **Timestamp:** The time when the encapsulated routes were advertised (one may also think of this as the time when they were installed in the Adj-RIB-Out), 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. Adj-RIB-Out

5.1. Post-Policy

The primary use-case in monitoring Adj-RIB-Out is to monitor the updates transmitted to a BGP peer after outbound policy has been applied. These updates reflect the result after modifications and filters have been applied (e.g., Adj-RIB-Out Post-Policy). Some

attributes are set when the BGP message is transmitted, such as next-hop. Adj-RIB-Out Post-Policy MUST convey to the BMP receiver what is actually transmitted to the peer.

The L flag MUST be set to 1 to indicate post-policy.

5.2. Pre-Policy

Similarly to Adj-RIB-In policy validation, pre-policy Adj-RIB-Out can be used to validate and audit outbound policies. For example, a comparison between pre-policy and post-policy can be used to validate the outbound policy.

Depending on BGP peering session type (IBGP, IBGP route reflector client, EBGP, BGP confederations, Route Server Client) the candidate routes that make up the Pre-Policy Adj-RIB-Out do not contain all local-rib routes. Pre-Policy Adj-RIB-Out conveys only routes that are available based on the peering type. Post-Policy represents the filtered/changed routes from the available routes.

Some attributes are set only during transmission of the BGP message, i.e., Post-Policy. It is common that next-hop may be null, loopback, or similar during pre-policy phase. All mandatory attributes, such as next-hop, MUST be either ZERO or have an empty length if they are unknown at the Pre-Policy phase completion. The BMP receiver will treat zero or empty mandatory attributes as self-originated.

The L flag MUST be set to 0 to indicate pre-policy.

6. BMP Messages

Many BMP messages have a per-peer header but some are not applicable to Adj-RIB-In or Adj-RIB-Out monitoring, such as peer up and down notifications. Unless otherwise defined, the O flag should be set to 0 in the per-peer header in BMP messages.

6.1. Route Monitoring and Route Mirroring

The O flag MUST be set accordingly to indicate if the route monitor or route mirroring message conveys Adj-RIB-In or Adj-RIB-Out.

6.2. Statistics Report

The Statistics report message has a Stat Type field to indicate the statistic carried in the Stat Data field. Statistics report messages are not specific to Adj-RIB-In or Adj-RIB-Out and MUST have the O flag set to zero. The O flag SHOULD be ignored by the BMP receiver.

The following new statistic types are added:

- o Stat Type = 14: (64-bit Gauge) Number of routes in Adj-RIBs-Out Pre-Policy.
- o Stat Type = 15: (64-bit Gauge) Number of routes in Adj-RIBs-Out Post-Policy.
- o Stat Type = 16: Number of routes in per-AFI/SAFI Adj-RIB-Out Pre-Policy. The value is structured as: 2-byte Address Family Identifier (AFI), 1-byte Subsequent Address Family Identifier (SAFI), followed by a 64-bit Gauge.
- o Stat Type = 17: Number of routes in per-AFI/SAFI Adj-RIB-Out Post-Policy. The value is structured as: 2-byte Address Family Identifier (AFI), 1-byte Subsequent Address Family Identifier (SAFI), followed by a 64-bit Gauge.

6.3. Peer Down and Up Notifications

Peer Up and Down notifications convey BGP peering session state to BMP receivers. The state is independent of whether or not route monitoring or route mirroring messages will be sent for Adj-RIB-In, Adj-RIB-Out, or both. BMP receiver implementations SHOULD ignore the O flag in Peer Up and Down notifications.

6.3.1. Peer Up Information

The following Peer Up message Information TLV type is added:

- o Type = 4: Admin Label. The Information field contains a free-form UTF-8 string whose byte length is given by the Information Length field. The value is administratively assigned. There is no requirement to terminate the string with null or any other character.

Multiple admin labels can be included in the Peer Up notification. When multiple admin labels are included the BMP receiver MUST preserve their order.

The TLV is optional.

7. Other Considerations

7.1. Peer and Update Groups

Peer and update groups are used to group updates shared by many peers. This is a level of efficiency in implementations, not a true representation of what is conveyed to a peer in either Pre-Policy or Post-Policy.

One of the use-cases to monitor Adj-RIB-Out Post-Policy is to validate and continually ensure the egress updates match what is expected. For example, wholesale peers should never have routes with community X:Y sent to them. In this use-case, there may be hundreds of wholesale peers but a single peer could have represented the group.

From a BMP perspective, this should be simple to include a group name in the Peer Up, but it is more complex than that. BGP implementations have evolved to provide comprehensive and structured policy grouping, such as session, AFI/SAFI, and template-based based group policy inheritances.

This level of structure and inheritance of policies does not provide a simple peer group name or ID, such as wholesale peer.

Instead of requiring a group name to be used, a new administrative label informational TLV (Section 6.3.1) is added to the Peer Up message. These labels have administrative scope relevance. For example, labels "type=wholesale" and "region=west" could be used to monitor expected policies.

Configuration and assignment of labels to peers is BGP implementation specific.

8. Security Considerations

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

9. IANA Considerations

This document requests that IANA assign the following new parameters to the BMP parameters name space [1].

9.1. BMP Peer Flags

This document defines the following per-peer header flags (Section 4):

- o Flag 3 as O flag: The O flag indicates Adj-RIB-In if set to 0 and Adj-RIB-Out if set to 1.

9.2. BMP Statistics Types

This document defines four statistic types for statistics reporting (Section 6.2):

- o Stat Type = 14: (64-bit Gauge) Number of routes in Adj-RIBs-Out Pre-Policy.
- o Stat Type = 15: (64-bit Gauge) Number of routes in Adj-RIBs-Out Post-Policy.
- o Stat Type = 16: Number of routes in per-AFI/SAFI Adj-RIB-Out Pre-Policy. The value is structured as: 2-byte Address Family Identifier (AFI), 1-byte Subsequent Address Family Identifier (SAFI), followed by a 64-bit Gauge.
- o Stat Type = 17: Number of routes in per-AFI/SAFI Adj-RIB-Out Post-Policy. The value is structured as: 2-byte Address Family Identifier (AFI), 1-byte Subsequent Address Family Identifier (SAFI), followed by a 64-bit Gauge.

9.3. Peer Up Information TLV

This document defines the following BMP Peer Up Information TLV types (Section 6.3.1):

- o Type = 4: Admin Label. The Information field contains a free-form UTF-8 string whose byte length is given by the Information Length field. The value is administratively assigned. There is no requirement to terminate the string with null or any other character.

Multiple admin labels can be included in the Peer Up notification. When multiple admin labels are included the BMP receiver MUST preserve their order.

The TLV is optional.

10. References

10.1. 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>>.
- [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>>.

10.2. URIs

- [1] <https://www.iana.org/assignments/bmp-parameters/bmp-parameters.xhtml>

Acknowledgements

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

Contributors

Manish Bhardwaj
Cisco Systems
3700 Cisco Way
San Jose, CA 95134
USA

Email: manbhard@cisco.com

Xianyuzheng
Tencent
Tencent Building, Kejizhongyi Avenue,
Hi-techPark, Nanshan District, Shenzhen 518057, P.R.China

Weiguo
Tencent
Tencent Building, Kejizhongyi Avenue,
Hi-techPark, Nanshan District, Shenzhen 518057, P.R.China

Shugang cheng
H3C

Authors' Addresses

Tim Evens
Cisco Systems
2901 Third Avenue, Suite 600
Seattle, WA 98121
USA

Email: tievens@cisco.com

Serpil Bayraktar
Cisco Systems
3700 Cisco Way
San Jose, CA 95134
USA

Email: serpil@cisco.com

Paolo Lucente
NTT Communications
Siriusdreef 70-72
Hoofddorp, WT 2132
NL

Email: paolo@ntt.net

Penghui Mi
Tencent
Tengyun Building, Tower A ,No. 397 Tianlin Road
Shanghai 200233
China

Email: kevinmi@tencent.com

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.

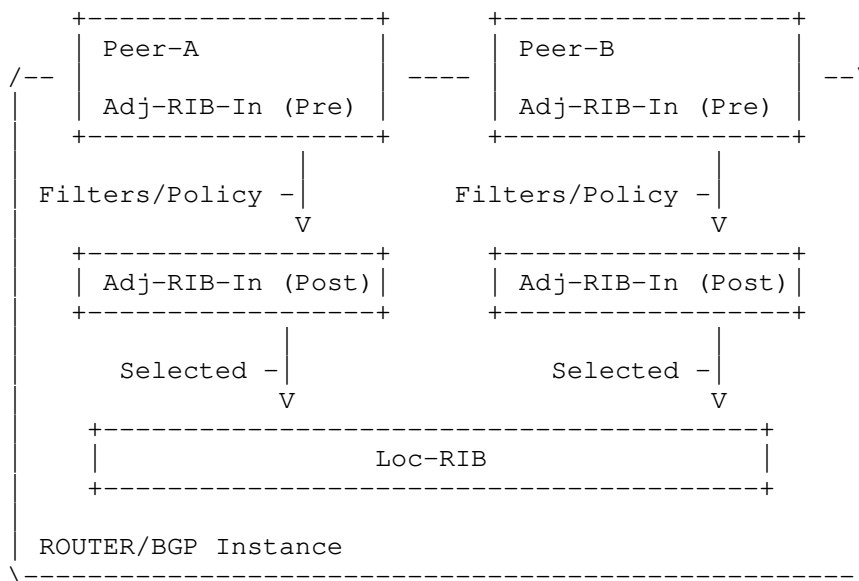


Figure 1: BGP peering Adj-RIBs-In into Loc-RIB

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.

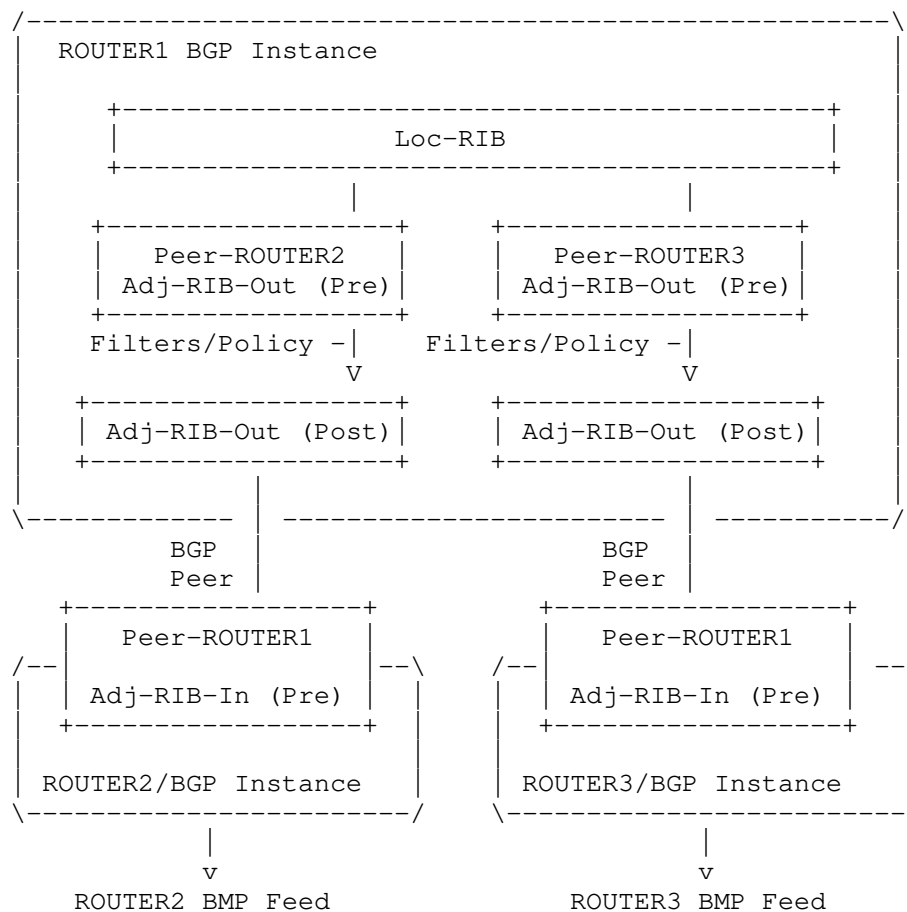


Figure 2: Alternative method to monitor Loc-RIB

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

Global Routing Operations
Internet-Draft
Intended status: Informational
Expires: November 2, 2018

J. Snijders
NTT
M. Stucchi
RIPE NCC
May 1, 2018

RPKI Autonomous Systems Cones: A Profile To Define Sets of Autonomous
Systems Numbers To Facilitate BGP Filtering
draft-ss-grow-rpki-as-cones-00

Abstract

This document describes a way to define groups of Autonomous System numbers in RPKI [RFC6480]. We call them AS-Cones. AS-Cones provide a mechanism to be used by operators for filtering BGP-4 [RFC4271] announcements.

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 November 2, 2018.

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
2. Format of AS-Cone objects	3
2.1. Policy definition object	3
2.1.1. Naming convention for Policy definition objects	3
2.1.2. ASN.1 format of a Policy Definition object	3
2.1.3. Naming convention for neighbour relationships	4
2.2. AS-Cone definition object	4
2.2.1. Naming convention for AS-Cone objects	4
2.2.2. ASN.1 format of an AS-Cone	5
3. Validating an AS-Cone	5
4. Recommendations for use of AS-Cones at Internet Exchange points	6
5. Publication of AS-Cones as IRR objects	7
6. Security Considerations	7
7. IANA Considerations	7
8. Contributors	7
9. Acknowledgments	7
10. References	7
10.1. Normative References	7
10.2. Informative References	8
Authors' Addresses	8

1. Introduction

The main goal of the Resource Public Key Infrastructure (RPKI) system [RFC6480] is to support improved security for the global routing system. This is achieved through the use of information stored in a distributed repository system comprised of signed objects. A commonly used object type is the Route Object Authorisation (ROAs), which describe the prefixes originated by ASNs.

There is however no way for an operator to assert the routes for its customer networks, making it difficult to use the information carried by RPKI to create meaningful BGP-4 filters without relying on RPSL [RFC2622] as-sets.

This memo introduces a new attestation object, called an AS-Cone. An AS-Cone is a digitally signed object with the goal to enable operators to define a set of customers that can be found as "right adjacencies", or transit customer networks, facilitating the construction of prefix filters for a given ASN, thus making routing more secure.

2. Format of AS-Cone objects

AS-Cones are composed of two types of distinct objects:

- o Policy definitions; and
- o The AS-Cones themselves.

These objects are stored in ASN.1 format and are digitally signed according to the same rules and conventions applied for RPKI ROA Objects ([RFC6482]).

2.1. Policy definition object

A policy definition contains a list the upstream and peering relationships for a given Autonomous System that need an AS-Cone to be used for filtering. For each relationship, an AS-Cone is referenced to indicate which BGP networks will be announced to the other end of the relationship.

The default behaviour for a neighbour, if the relationship is not explicitly described in the policy, is to only accept the networks originated by the ASN. This means that a stub ASN does not have to set up any AS-Cone nor any description or policy.

Only one AS-Cone can be supplied for a given relationship. If more than one AS-Cone needs to be announced in the relationship, then it is mandatory to create a third AS-Cone that includes those two.

2.1.1. Naming convention for Policy definition objects

A Policy object is referenced using the Autonomous System number it refers to, preceded by the string "AS".

2.1.2. ASN.1 format of a Policy Definition object

```
ASNPolicy DEFINITIONS ::=
BEGIN
Neighbours ::= SEQUENCE OF Neighbour

Neighbour ::= SEQUENCE
{
    ASN INTEGER (1..4294967296),
    ASCone VisibleString
}

Version ::= INTEGER
LastModified ::= GeneralizedTime
Created ::= GeneralizedTime
END
```

ASN.1 format of a Policy definition object

2.1.3. Naming convention for neighbour relationships

When referring to a neighbour relationship contained in a Policy definition object the following convention should be used:

ASX:ASY

Where X is the number of the AS holder and Y is the number of the ASN intended to use the AS-Cone object to generate a filter.

2.2. AS-Cone definition object

An AS-Cone contains a list of the downstream customers and AS-Cones of a given ASN. The list is used to create filter lists by the networks providing transit or a peering relationship with the ASN.

An AS-Cone can reference another AS-Cone if a customer of the operator also has defined an AS-Cone to be announced upstream.

2.2.1. Naming convention for AS-Cone objects

AS-Cones MUST have a unique name for the ASN they belong to. Names are composed of ASCII strings up to 255 characters long and cannot contain spaces.

In order for AS-Cones to be unique in the global routing system, their string name is preceded by the AS number of the ASN they are part of, followed by ":". For example, AS-Cone "EuropeanCustomers" for ASN 65530 is represented as "AS65530:EuropeanCustomers" when referenced from a third party.

2.2.2. ASN.1 format of an AS-Cone

```
ASConc DEFINITIONS ::=
BEGIN
Entities ::= SEQUENCE OF Entity

Entity CHOICE
{
    ASN INTEGER (1..4294967296),
    OtherASConc VisibleString
}

Version ::= INTEGER
LastModified ::= GeneralizedTime
Created ::= GeneralizedTime
END
```

ASN.1 format of an AS-Cone

3. Validating an AS-Cone

The goal of AS-Cones is to be able to recursively define all the originating ASNs that define the customer base of a given ASN, including all the transit relationships. This means that through AS-Cones, it is possible to create a graph of all the neighbour relationships for the customers of a given ASN.

In order to validate a full AS-Cone, we have to assume that a network operator already runs an RPKI validator software. The software provides access to a validated cache containing all the Policy definitions and AS-Cone objects. Validation occurs following the description in: [RFC6488].

In order to validate a full AS-Cone, an operator should perform the following steps:

1. For Every downstream ASN, the operator takes its policy definition file and collects a list of ASNs for the cone by looking at the following data, in exact order:
 1. A policy for the specific relationship, in the form of ASX:ASY, where ASX is the downstream ASN, and ASY is the ASN of the operator validating the AS-Cone;
 2. If there is no specific definition for the relationship, the ASX:Default policy;

If none of the two objects above exists, then the operator should only consider the ASN of its downstream to be added to the list.

2. These objects can either point to:
 1. An AS-Cone; or
 2. An ASN
3. If the definition points to an AS-Cone, the operator looks for the object referenced, which should be contained in the validated cache;
4. If the validated cache does not contain the referenced object, then the validation moves on to the next downstream network;
5. If the validated cache contains the referenced object, the validation process evaluates every entry in the AS-Cone. For each entry:
 1. If there is a reference to an ASN, then the operator adds the ASN to the list for the given AS-Cone;
 2. If there is a reference to another AS-Cone, the validating process should recursively process all the entries in that AS-Cone first, with the same principles contained in this list.

Since the goal is to build a list of ASNs announcing routes in the AS-Cone, then if an ASN or an AS-Cone are referenced more than once in the process, their contents should only be added once to the list. This is intended to avoid endless loops, and in order to avoid cross-reference of AS-Cones

6. When all the AS-Cones referenced in the policies have been recursively iterated, and all the originating ASNs have been taken into account, the operator can then build a full prefix-list with all the prefixes originated in its AS-Cone. This can be done by querying the RPKI validator software for all the networks originated by every ASN referenced in the AS-Cone.
4. Recommendations for use of AS-Cones at Internet Exchange points

When an operator is a member of an internet exchange point, it is recommended for it to create at least a Default policy.

In case of a peering session with a route server, the operator could publish a policy pointing to the ASN of the route server. A route

server operator, then, could build strict prefix filtering rules for all the participants, and offer it as a service to its members.

5. Publication of AS-Cones as IRR objects

AS-Cones are very similar to AS-Set RPSL Objects, so they could also be published in IRR Databases as AS-Set objects. Every ASN contained in an AS-Cone, and all the AS-Cones referenced should be considered as member: attributes. The naming convention for AS-Cones (ASX:AS-Cone) should be maintained, in order to keep consistency between the two databases.

6. Security Considerations

TBW

7. IANA Considerations

This memo includes no request to IANA.

8. Contributors

The following people contributed significantly to the content of the document:

9. Acknowledgments

The authors would like to thank ...

10. References

10.1. 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>>.

10.2. Informative References

- [RFC2622] Alaettinoglu, C., Villamizar, C., Gerich, E., Kessens, D., Meyer, D., Bates, T., Karrenberg, D., and M. Terpstra, "Routing Policy Specification Language (RPSL)", RFC 2622, DOI 10.17487/RFC2622, June 1999, <<https://www.rfc-editor.org/info/rfc2622>>.
- [RFC6480] Lepinski, M. and S. Kent, "An Infrastructure to Support Secure Internet Routing", RFC 6480, DOI 10.17487/RFC6480, February 2012, <<https://www.rfc-editor.org/info/rfc6480>>.
- [RFC6482] Lepinski, M., Kent, S., and D. Kong, "A Profile for Route Origin Authorizations (ROAs)", RFC 6482, DOI 10.17487/RFC6482, February 2012, <<https://www.rfc-editor.org/info/rfc6482>>.
- [RFC6488] Lepinski, M., Chi, A., and S. Kent, "Signed Object Template for the Resource Public Key Infrastructure (RPKI)", RFC 6488, DOI 10.17487/RFC6488, February 2012, <<https://www.rfc-editor.org/info/rfc6488>>.

Authors' Addresses

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

Email: job@ntt.net

Massimiliano Stucchi
RIPE NCC
Stationsplein, 11
Amsterdam 1012 AB
The Netherlands

Email: mstucchi@ripe.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 23, 2018

J. Borkenhagen
AT&T
R. Bush
Internet Initiative Japan
R. Bonica
Juniper Networks
S. Bayraktar
Cisco Systems
June 21, 2018

Well-Known Community Policy Behavior
draft-ymbk-grow-wkc-behavior-03

Abstract

Well-Known BGP Communities are manipulated inconsistently by current implementations. This results in difficulties for operators. It is recommended that removal policies be applied consistently to Well-Known Communities.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in RFC 2119 [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning.

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 23, 2018.

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
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	4
6. Action Items	5
7. Security Considerations	5
8. IANA Considerations	5
9. Acknowledgements	5
10. Normative References	5
Authors' Addresses	6

1. Introduction

The BGP Communities Attribute was specified in [RFC1997] which introduced the concept of Well-Known Communities. In hindsight, it 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.

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."

A basic operational need is to add or remove one or more communities to the received set. Another common need is to replace all received communities with a new set. To simplify the second case, most BGP policy implementations provide syntax to "set" community that operators use to mean "remove any/all communities present on the update received from the neighbor, 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 received 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.

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 Juniper Networks' JunOS, "community set" removes all received communities, Well-Known or otherwise.

In Cisco Systems' IOS-XR, "set community" removes all received 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

IOS-XR does allow Well-Known communities to be removed one at a time by explicit policy; for example, "delete community accept-own". Operators are advised to consult IOS-XR documentation and/or Cisco Systems support for full details.

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

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

In OpenBSD's OpenBGPD, "set community" does not remove any communities, well-Known or otherwise.

4.1. Note on an Inconsistency

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

IOS-XR's set of well-known communities that "set community" will not overwrite diverges from IANA's list. Quite a few well-known communities from IANA's list do not receive special treatment in IOS-XR, and at least one specific community on IOS-XR's special treatment list (internet == 0:0) is not really on IANA's list -- it's 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

Care should be taken when establishing new [RFC1997]-like attributes (large communities, wide communities, etc) to avoid repeating this mistake.

6. Action Items

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

Vendors SHOULD share the behavior of their implementations for inclusion in this document, especially if their behavior differs from the examples described.

For new well-known communities specified (after this draft), vendors MUST treat "community set" command to mean "remove all other communities, Well-Known or otherwise."

7. Security Considerations

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

8. IANA Considerations

This document has no IANA Considerations other than to be aware that any future Well-Known Communities will be subject to the policy treatment described here.

9. Acknowledgements

The authors thank Martijn Schmidt for his contribution, Qin Wu for the Huawei data point.

10. Normative References

[IANA-WKS]

"IANA Well-Known Communities",
<<https://www.iana.org/assignments/bgp-well-known-communities/bgp-well-known-communities.xhtml>>.

[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>>.

Authors' Addresses

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

Email: jayb@att.com

Randy Bush
Internet Initiative Japan
5147 Crystal Springs
Bainbridge Island, WA 98110
United States of America

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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2019

S. Zhuang
Y. Gu
Z. Li
Huawei
July 02, 2018

Monitoring BGP Parameters Using BMP
draft-zhuang-grow-monitoring-bgp-parameters-00

Abstract

The BGP Monitoring Protocol (BMP) [RFC7854] is designed to monitor BGP [RFC4271] running status, such as BGP peer relationship establishment and termination and route updates. Without BMP, manual query is required if you want to know about BGP running status.

This document provides the use cases that the BMP station can get the optional and default configure parameters of the monitored network device via 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].

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 3, 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. Terminology	2
2. Introduction	2
3. Use cases	3
4. Extension of BMP Initiation Message	4
5. Acknowledgements	6
6. IANA Considerations	6
7. Security Considerations	6
8. Normative References	6
Authors' Addresses	7

1. Terminology

This memo makes use of the terms defined in [RFC7854].

BMP: BGP Monitoring Protocol

BMS: BGP Monitoring Station

Initiation message: Reports to the monitoring server such information as the router vendor and its software version.

2. Introduction

The Border Gateway Protocol (BGP) is a dynamic routing protocol operating on an Autonomous System (AS) and typically configured on a network device. The BGP typically can support a number of optional parameters[RFC5492], e.g., IPv4 Unicast, IPv4 Multicast, IPv6 Unicast, and other Multiple-Protocol Extended Capabilities, Route Refresh Capability, Outbound Route Filtering Capability, Graceful Restart Capability, Support for 4-octet AS number capability etc., and the different BGP may support a different number of different

capabilities. The network device configured with the BGP typically may not enable all optional capabilities supported in the configured BGP, but enable some currently required BGP optional capabilities as required for a current task.

The BGP Monitoring Protocol (BMP) introduces the availability of monitoring BGP running status, such as BGP peer relationship establishment and termination and route updates. Without BMP, manual query is required if you want to know about BGP running status. With BMP, a router can be connected to a monitoring station and configured to report BGP running statistics to the station for monitoring, which improves the network monitoring efficiency. BMP facilitates the monitoring of BGP running status and reports security threats in real time so that preventive measures can be taken promptly.

In order to monitor and manage effectively the operating states of the BGP configured on the respective network devices in the network, the existing practice is that a monitoring station obtains BGP information of the respective network devices in the network to monitor and manage centrally the network devices configured with the BGP in the network. By way of an example of a flow in which the monitoring station obtains the BGP information, after a BGP connection is set up between network devices A and B configured with the BGP (or between peers), taking the network device A as an example, the network devices A and B negotiate about their own enabled BGP optional capabilities in OPEN messages under a BGP rule, and the network device A further includes a BGP Monitoring Protocol (BMP) module connected with the monitoring station, where the BMP module can obtain the enabled BGP optional capabilities of the network device A, and the enabled BGP capabilities of the network device B as a result of negotiation about the enabled BGP capabilities, so that if the BMP module of the network device A sends the configured BGP information of the network device to the monitoring station in a Peer Up Notification message, then the BGP optional capabilities will include only the BGP capabilities enabled on the network device A.

However, sometimes it's not sufficient to report only the capabilities currently enabled at the monitored device to the BMS. In order to better optimize the network, the BMS may want to access all the capabilities that are supported at each monitored devices, as well as the current configuration informations.

3. Use cases

- o BGP Optional Parameters: The Open Message reported to BMS contains only the currently enabled capabilities at the monitored device. If all the supported capabilities of the monitored devices, both

the enabled and not yet enabled ones, are informed to the BMS, the BMS can use the more comprehensive and valid inputs to make decisions about the whole network optimization. For example, if the Graceful Restart Capability is not enabled for a BGP Peer, and thus the BGP Open Message (i.e., the Peer Up Notification in BMP) would not include the GR capability. However, if the BMS or the operator has the knowledge that both devices support the GR capability, and enables it at both devices, it could improve the operational stability of the network.

- o BGP Default Behavior Parameters: As one of the concern from the operators, that in multi-vendor environment, some default configurations or behaviors of devices are vendor-specific, and may cause various issues during the interoperation test or any time after. Take the protocol preferences (distance) of different BGP routes for example: Vendor A assigns value 255 to eBGP, iBGP and BGP local routes by default, while vendor B assigns 20 to eBGP, 200 to iBGP and 200 to BGP local routes by default. In addition, value 255 is not recognized by vendor B, and routes assigned such distance would be ignored.

4. Extension of BMP Initiation Message

As described in Section 4.3 of [RFC7854], the initiation message provides a means for the monitored router to inform the monitoring station of its vendor, software version, and so on.

The initiation message consists of the common BMP header followed by two or more Information TLVs (Section 4.4 of [RFC7854]) containing information about the monitored router. Currently defined types are:

Type = 0: String.

Type = 1: sysDescr.

Type = 2: sysName.

This document defines two new categories of TLV types: the BGP Optional Parameters and the BGP Default Behavior Parameters.

Type = TBD1: BGP Optional Parameters. The Information field is used to specify all the BGP Optional Parameters that have been enabled or not yet enabled at the monitored device. Each optional parameter is encoded as a <Parameter Type, Parameter Length, Parameter Value> triplet, as defined in RFC 4271 [RFC4271] .

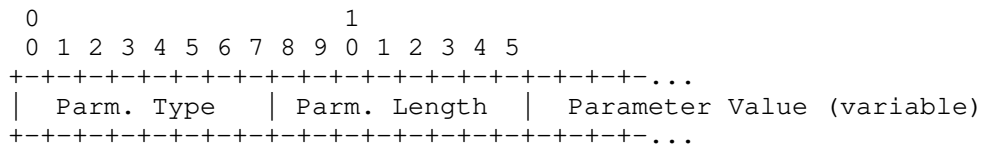


Figure 1 BGP Optional Parameters Information TLV

Parameter Type is a one octet field that unambiguously identifies individual parameters. Parameter Length is a one octet field that contains the length of the Parameter Value field in octets. Parameter Value is a variable length field that is interpreted according to the value of the Parameter Type field. RFC 5492 [RFC5492] defines the Capabilities Optional Parameter.

Type = TBD2: Default Behavior Parameters. The Information field contains a list of default behavior parameters, in which each parameter is encoded as a Default Behavior sub TLV <Default Behavior Type, Default Behavior Length, Default Behavior Value>, which is defined as follows

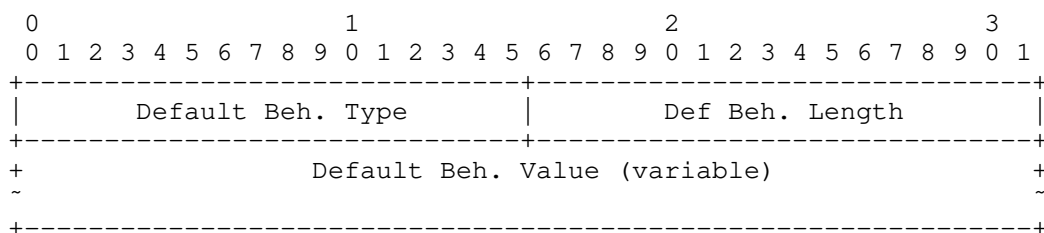


Figure 2 Default Behavior sub TLV

The Default Behavior Type is a one octet field that identifies the default behavior type parameter. Parameter Length is a one octet field that contains the length of the Parameter Value field in octets. Parameter Value is a variable length field that is interpreted according to the value of the Parameter Type field:

- o Type = TBD3, (32-bit integer) Value of default Protocol Preference for Local route
- o Type = TBD4, (32-bit integer) Value of default Protocol Preference for EBGp route
- o Type = TBD5, (32-bit integer) Value of default Protocol Preference for IBGP route

- o Type = TBD6, (32-bit integer) Value of default BGP connect-retry timer time
- o Type = TBD7, (32-bit integer) Value of default BGP Keepalive time
- o Type = TBD8, (32-bit integer) Value of default BGP hold time
- o Type = TBD9, (32-bit integer) Value of EBGp route-update-interval
- o Type = TBD10, (32-bit integer) Value of IBGP route-update-interval
- o Type = TBD11, (32-bit integer) Value of Default local-preference
- o Type = TBD12, (32-bit integer) Value of Default MED

5. Acknowledgements

TBD.

6. IANA Considerations

TBD.

7. Security Considerations

TBD.

8. 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>>.

Authors' Addresses

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

Email: zhuangshunwan@huawei.com

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

Email: guyunan@huawei.com

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

Email: lizhenbin@huawei.com