Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 14, 2010

J. Scudder
R. Fernando
Juniper Networks
S. Stuart
Google

July 13, 2009

BGP Monitoring Protocol
draft-ietf-grow-bmp-02

Status of this Memo

Copyright Notice

Abstract

This document proposes a simple protocol, BMP, which can be used to monitor BGP sessions. BMP is intended to provide a more convenient interface for obtaining route views for research purpose than the screen-scraping approach in common use today. The design goals are to keep BMP simple, useful, easily implemented, and minimally service-affecting. BMP is not suitable for use as a routing protocol.

Table of Contents

1. Introduction

Many researchers wish to have access to the contents of routers' BGP RIBs as well as a view of protocol updates that the router is receiving. This monitoring task cannot be realized by standard protocol mechanisms. At present, this data can only be obtained through screen-scraping.

The BMP protocol provides access to the Adj-RIB-In of a peer on an ongoing basis and a periodic dump of certain statistics that the monitoring station can use for further analysis. The following are the messages provided by BMP.

- Route Monitoring (RM): An initial dump of all routes received from a peer as well as an ongoing mechanism that sends the incremental routes advertised and withdrawn by a peer to the monitoring station.

- Peer Down Notification (PD): A message sent to indicate that a peering session has gone down with information indicating the reason for the session disconnect.

- Stats Reports (SR): This is an ongoing dump of statistics that can be used by the monitoring station as a high level indication of the activity going on in the router.

- Peer Up Notification (PU): A message sent to indicate that a peering session has come up. The message includes information regarding the data exchanged between the peers in their OPEN messages as well as information about the peering TCP session itself.

BMP operates over TCP. All options are controlled by configuration on the monitored router. No message is ever sent from the monitoring station to the monitored router. The monitored router MAY take steps to prevent the monitoring station from sending data (e.g. by half-closing the TCP session or setting its window size to zero) or it MAY silently discard any data erroneously sent by the monitoring station.

The monitoring station is configured to listen on a particular TCP port and the router is configured to establish an active connection to that port and to send messages on that TCP connection. There is no initialization or handshaking phase, messages are simply sent as soon as the connection is established. If the router is unable to connect to the monitoring station, it periodically retries the connection. A suggested default retry period is 30 seconds.

If the monitoring station intends to restart BMP processing, it

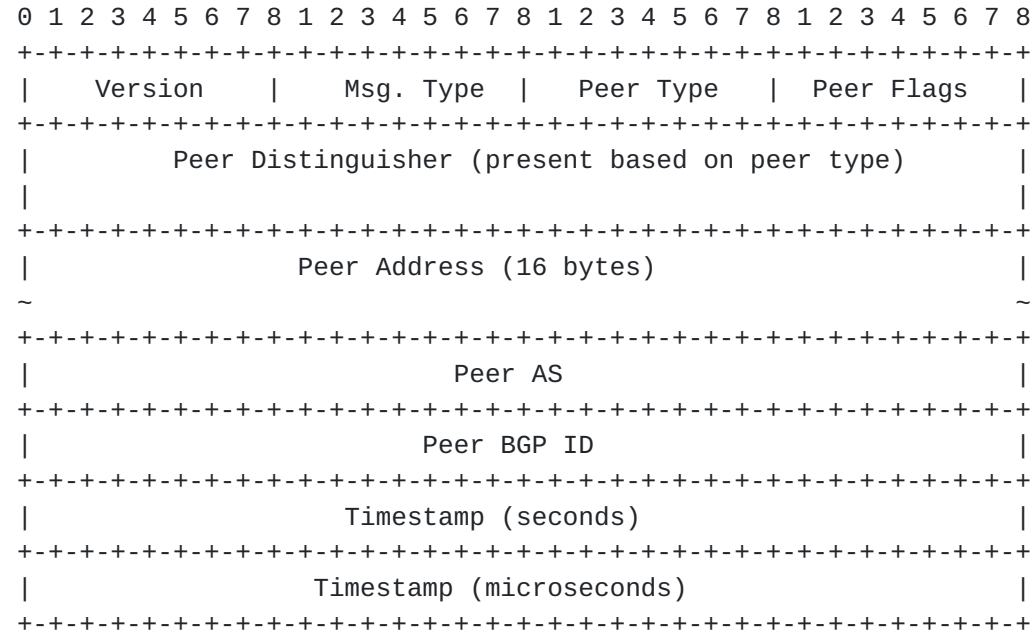
simply drops the connection. The router then re-establishes the connection and resends the messages.

1.1. 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].

2. BMP Message Format

The following common header appears in all BMP messages. The rest of the data in a BMP message is dependent on the "Message Type" field in the common header.



- Version (1 byte): Indicates the BMP version. This is set to '2' for all messages defined in this specification.

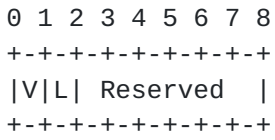
- Message Type (1 byte): This identifies the type of the BMP message,

  * Type = 0: Route Monitoring
  * Type = 1: Statistics Report
  * Type = 2: Peer Down Notification
  * Type = 3: Peer Up Notification

- Peer Type (1 byte): These bits identify the type of the peer. Currently only two types of peers are identified,

* Peer Type = 0: Global Instance Peer
* Peer Type = 1: L3 VPN Instance Peer

o Peer Flags (1 byte): These flags provide more information about the peer. The flags are defined as follows.



* The V flag indicates the the Peer address is an IPv6 address. For IPv4 peers this is set to 0.
* The L flag, if set to 1, indicates that the message reflects the Loc-RIB (i.e., it reflects the application of inbound policy). It is set to 0 if the message reflects the Adj-RIB-In.
* The remaining bits are reserved for future use.

o Peer Distinguisher (8 bytes): Routers today can have multiple instances (example L3VPNs). This field is present to distinguish peers that belong to one address domain from the other.

  If the peer is a "Global Instance Peer", this field is zero filled. If the peer is a "L3VPN Instance Peer", it is set to the route distinguisher of the particular L3VPN instance that the peer belongs to.

o Peer Address: The remote IP address associated with the TCP session over which the encapsulated PDU was received. It is 4 bytes long if an IPv4 address is carried in this field (with most significant bytes zero filled) and 16 bytes long if an IPv6 address is carried in this field.

o Peer AS: The Autonomous System number of the peer from which the encapsulated PDU was received. If a 16 bit AS number is stored in this field [RFC4893], it should be padded with zeroes in the most significant bits.

o Peer BGP ID: The BGP Identifier of the peer from which the encapsulated PDU was received.

o Timestamp: The time when the encapsulated routes were received (one may also think of this as the time when they were installed in the Adj-RIB-In), 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.

2.1. Route Monitoring

Route Monitoring messages are used for initial synchronization of ADJ-RIB-In. They are also used for ongoing monitoring of received advertisements and withdraws. This is discussed in more detail in subsequent sections.

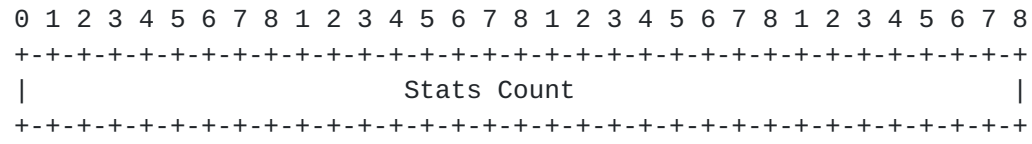
Following the common BMP header is a BGP PDU. The length of the PDU can be determined by parsing it in the normal fashion as specified in [RFC4271].

2.2. Stats Reports

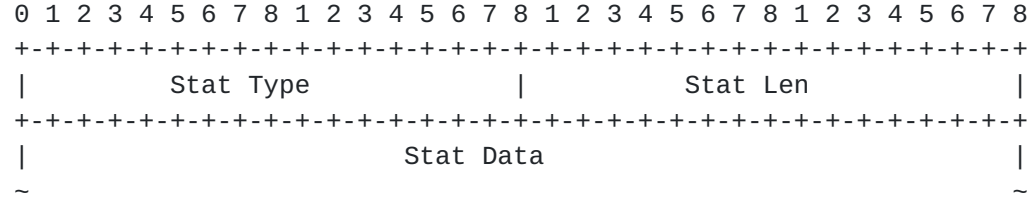
These messages contain information that could be used by the monitoring station to observe interesting events that occur on the router. 'Stats Report' messages have a message type of '3'.

The transmission of the SR messages could be timer triggered or event driven (for example, when a significant event occurs or a threshold is reached). This specification does not impose any timing restrictions on when and on what event these reports have to be transmitted. It is left to the implementation to determine transmission timings -- however, configuration control should be provided of the timer and/or threshold values. This document only specifies the form and content of SR messages.

Following the common BMP header is a 4-byte field that indicates the number of counters in the stats message where each counter is encoded as a TLV.



Each counter is encoded as follows,



- Stat Type (2 bytes): Defines the type of the statistic carried in the "Stat Data" field.

- Stat Len (2 bytes): Defines the length of the "Stat Data" Field.

This specification defines the following statistics. All statistics are 4-byte quantities and the stats data are counters.

- Stat Type = 0: Number of prefixes rejected by inbound policy.

- Stat Type = 1: Number of (known) duplicate prefix advertisements.

- Stat Type = 2: Number of (known) duplicate withdraws.

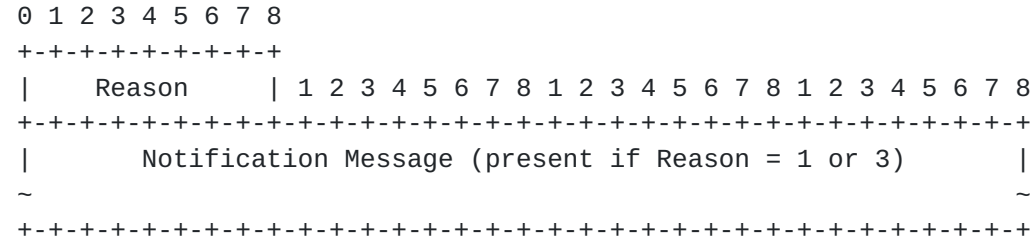
- Stat Type = 3: Number of updates invalidated due to CLUSTER_LIST loop.

- Stat Type = 4: Number of updates invalidated due to AS_PATH loop.

Note that the current specification only specifies 4-byte counters as "Stat Data". This does not preclude future versions from incorporating more complex TLV-type "Stat Data" (for example, one which can carry prefix specific data). SR messages are optional. However if an SR message is transmitted, this specification requires at least one statistic to be carried in it.

2.3. Peer Down Notification

This message is used to indicate that a peering session was terminated. The type of this message is 4.



Reason indicates why the session was closed. Defined values are:

- Reason 1: The local system closed the session. Following the Reason is a BGP PDU containing a BGP NOTIFICATION message that would have been sent to the peer. The length of the PDU can be determined by parsing it in the normal fashion as specified in [RFC4271].

- Reason 2: The local system closed the session. No notification message was sent.

- Reason 3: The remote system closed the session with a notification message. Following the Reason is a BGP PDU containing the BGP NOTIFICATION message as received from the peer. The length of the PDU can be determined by parsing it in the normal fashion as specified in [RFC4271].

- Reason 4: The remote system closed the session without a notification message.

2.4. Peer Up Notification

The Peer Up message is used to indicate that a peering session has come up (i.e., has transitioned into ESTABLISHED state). Following the common BMP header are the full OPEN messages as sent and received by the BGP speaker. The OPEN message transmitted by the monitored router to its peer is first, followed by the OPEN message received by the monitored router from its peer.

The length of the full PU message is the length of the fixed header plus the lengths of the two encapsulated OPEN messages which can be determined by parsing them in the normal fashion as specified in [RFC4271].

3. Route Monitoring

After the BMP session is up, Route Monitoring messages are used to provide a snapshot of the Adj-RIB-In of a particular peer. It does so by sending all routes stored in the Adj-RIB-In of that peer using standard BGP Update messages. There is no requirement on the ordering of messages in the peer dump.

Depending on the implementation or configuration, it may only be possible to send the Loc-RIB (post-policy routes) instead of the Adj-RIB-In. This is because it is possible that a BGP implementation may not store, for example, routes which have been filtered out by policy. If this is the case, the implementation may send the Loc-RIB path that pertains to a particular peer in the route monitor message.

If the implementation is able to provide information about when routes were received, it MAY provide such information in the BMP timestamp field. Otherwise, the BMP timestamp field MUST be set to zero, indicating that time is not available.

Ongoing monitoring is accomplished by propagating route changes in

BGP UPDATE PDUs and forwarding those PDUs to the monitoring station, again using RM messages. When a change occurs to a route, such as an attribute change, the router must update the monitor with the new attribute. When a route is withdrawn by a peer, a corresponding withdraw is sent to the monitor. Multiple changed routes MAY be grouped into a single BGP UPDATE PDU when feasible, exactly as in the standard BGP protocol.

It's important to note that RM messages are not real time replicated messages received from a peer. While the router should attempt to generate updates as soon as they are received there is a finite time that could elapse between reception of an update and the generation an RM message and its transmission to the monitoring station. If there are state changes in the interim for that prefix, it is acceptable that the router generate the final state of that prefix to the monitoring station. The actual PDU generated and transmitted to the station might also differ from the exact PDU received from the peer, for example due to differences between how different implementations format path attributes.

4. Stat Reports

As outlined above, SR messages are used to monitor specific events and counters on the monitored router. One type of monitoring could be to find out if there are an undue number of route advertisements and withdraws happening (churn) on the monitored router. Another metric is to evaluate the number of looped AS-Paths on the router.

While this document proposes a small set of counters to begin with, the authors envision this list may grow in the future with new applications that require BMP style monitoring.

5. Other Considerations

Some routers may support multiple instances of the BGP protocol, for example as "logical routers" or through some other facility. The BMP protocol relates to a single instance of BGP; thus, if a router supports multiple BGP instances it should also support multiple BMP instances (one per BMP instance).

6. Using BMP

Once the BMP session is established route monitoring starts dumping the current snapshot as well as incremental changes simultaneously.

It is fine to have these operations occur concurrently. If the initial dump visits a route and subsequently a withdraw is received, this will be forwarded to the monitoring station which would have to correlate and reflect the deletion of that route in its internal state. This is an operation a monitoring station would need to support regardless.

If the router receives a withdraw for a prefix even before the peer dump procedure visits that prefix, then the router would clean up that route from its internal state and will not forward it to the monitoring station. In this case, the monitoring station may receive a bogus withdraw which it can safely ignore.

7. IANA Considerations

This document defines four message types for transferring BGP messages between cooperating systems (Section 2):

- Type 0: Route Monitor
- Type 1: Statistics Report
- Type 2: Peer Down Notification
- Type 3: Peer Up Notification

Type values 4 through 255 MUST be assigned using the "IETF Consensus" policy defined in [RFC5226].

This document defines five statistics types for statistics reporting (Section 2.2):

- Stat Type = 0: Number of prefixes rejected by inbound policy.
- Stat Type = 1: Number of (known) duplicate prefix advertisements.
- Stat Type = 2: Number of (known) duplicate withdraws.
- Stat Type = 3: Number of updates invalidated due to CLUSTER_LIST loop.
- Stat Type = 4: Number of updates invalidated due to AS_PATH loop.

Stat Type values 5 through 32767 MUST be assigned using the "IETF Consensus" policy, and values 32768 through 65535 using the "First Come First Served" policy, defined in [RFC5226].

8. Security Considerations

This document defines a mechanism to obtain a full dump or provide continuous monitoring of a BGP speaker's local BGP table, including received BGP messages. This capability could allow an outside party to obtain information not otherwise obtainable.

Implementations of this protocol MUST require manual configuration of the monitored and monitoring devices.

Users of this protocol MAY use some type of secure transmission mechanism, such as IPSec [RFC4303], to transmit this data.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

9.2. Informative References

[RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, December 2005.

Appendix A. Changes Between BMP Versions 1 and 2

- Added Peer Up Message
- Added L flag
- Editorial changes

Authors' Addresses

John Scudder
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: jgs@juniper.net

Rex Fernando
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: rex@juniper.net

Stephen Stuart
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
USA

Email: sstuart@google.com