Global Routing Operations
Internet-Draft
Updates: 7854 (if approved)
Intended Status: Standards Track
Expires: October 1, 2017

T. Evens
S. Bayraktar
Cisco Systems
P. Lucente
NTT Communications
P. Mi
Tencent
S. Zhuang
Huawei

March 30, 2017

Support for Adj-RIB-Out in BGP Monitoring Protocol (BMP)
draft-evens-grow-bmp-adj-rib-out-01

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on October 1, 2017.

Table of Contents

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 verses 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.

This document updates BGP Monitoring Protocol (BMP) RFC 7854 [RFC7854] peer header by adding a new flag to distinguish Adj-RIB-In verses Adj-RIB-Out.

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

1.1. Terminology

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. Definitions

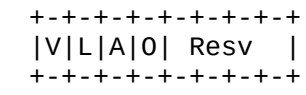
- Adj-RIB-Out: As defined in [RFC4271], "The Adj-RIBs-Out contains the routes for advertisement to specific peers by means of the local speaker's UPDATE messages."

- 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.

- 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.

3. Per-Peer Header

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



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

The remaining bits are reserved for future use. They MUST be transmitted as 0 and their values MUST be ignored on receipt.

4. Adj-RIB-Out

4.1. Post-Policy

The primary use-case in monitoring Adj-RIB-Out is to monitor the updates transmitted to the 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). The L flag MUST be set to 1 in this case to indicate post-policy.

4.2. Pre-Policy

As with Adj-RIB-In policy validation, there are use-cases that pre-policy Adj-RIB-Out is 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. The L flag MUST be set to 0 in

this case to indicate pre-policy.

5. BMP Messages

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

5.1. 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.

5.2. Statistics Report

Statistics report message has 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:

- Stat Type = TBD: (64-bit Gauge) Number of routes in Adj-RIBs-Out Pre-Policy.

- Stat Type = TBD: (64-bit Gauge) Number of routes in Adj-RIBs-Out Post-Policy.

- Stat Type = TBD: 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.

- Stat Type = TBD: 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.

5.3. 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. Security Considerations

It is not believed that this document adds any additional security considerations.

7. IANA Considerations

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

7.1. BMP Peer Flags

This document defines a new flag (Section 3):

- Flag 3 as O flag

7.2. BMP Statistics Types

This document defines four new statistic types for statistics reporting (Section 4.2):

- Stat Type = TBD: (64-bit Gauge) Number of routes in Adj-RIBs-Out Pre-Policy.
- Stat Type = TBD: (64-bit Gauge) Number of routes in Adj-RIBs-Out Post-Policy.
- Stat Type = TBD: 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.
- Stat Type = TBD: 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.

8. References

8.1. URIs

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

8.2. 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, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc7854>.

Acknowledgments

The authors would like to thank John Scudder for his 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 2132 WT
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
Intended Status: Standards Track
Expires: September 11, 2017
March 10, 2017

T. Evens
S. Bayraktar
M. Bhardwaj
Cisco Systems
P. Lucente
NTT Communications

Support for Local RIB in BGP Monitoring Protocol (BMP)
draft-evens-grow-bmp-local-rib-00

Abstract

The BGP Monitoring Protocol (BMP) defines access to the Adj-RIB-In and locally originated routes (e.g. routes distributed into BGP from protocols such as static) but not access to the BGP instance Loc-RIB. This document updates the BGP Monitoring Protocol (BMP) RFC 7854 by adding access to the BGP instance Local-RIB, as defined in RFC 4271 the routes that have been selected by the local BGP speaker’s Decision Process. These are the routes over all peers, locally originated, and after best-path selection.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on September 11, 2017.

Copyright and License Notice

Table of Contents

1. Introduction

The BGP Monitoring Protocol (BMP) suggests that locally originated routes are locally sourced routes, such as redistributed or otherwise added routes to the BGP instance by the local router. It does not specify routes that are in the BGP instance Loc-RIB, such as routes after best-path selection.

Figure 1 shows the flow of received routes from one or more BGP peers into the Loc-RIB.

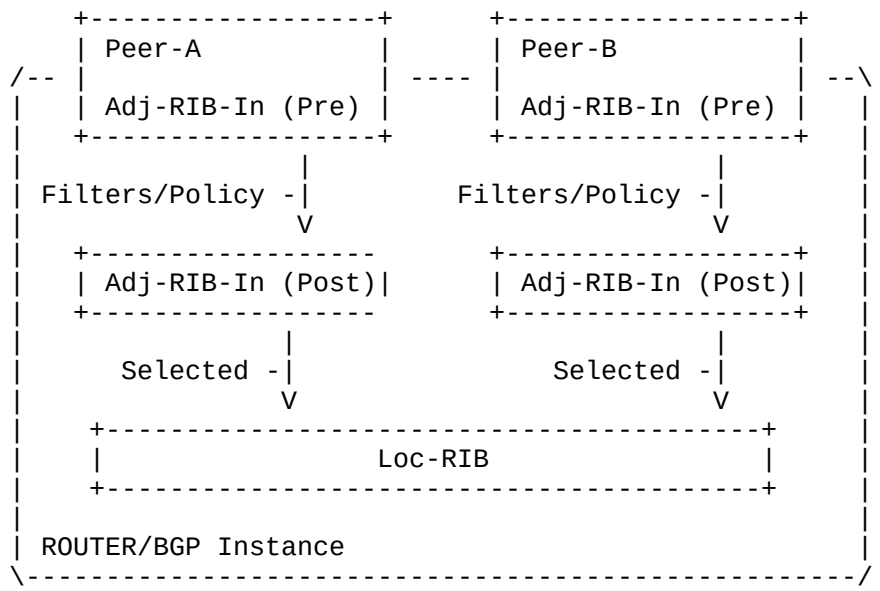


Figure 1: BGP peering Adj-RIBs-In into Loc-RIB

As shown in Figure 2, Locally originated follows a similar flow where the redistributed or otherwise originated routes get installed into the Loc-RIB based on the decision process selection.

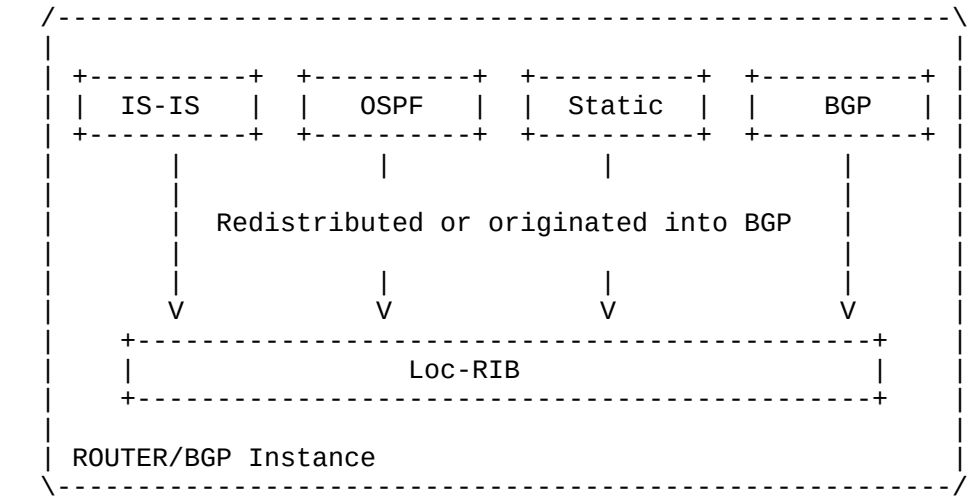


Figure 2: Locally Originated into Loc-RIB

BGP instance Loc-RIB usually provides a similar, if not exact, forwarding information base (FIB) view of the routes from BGP that the router will use. The following are some use-cases for Loc-RIB access:

- Adj-RIBs-In Post-Policy may still contain hundreds of thousands of routes per-peer but only a handful are selected and installed in the Loc-RIB as part of the 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 IGP next-hop metrics. While it is possible to obtain the IGP topology information using BGP-LS, it requires the application to implement SPF and possibly CSPF based on additional policies. This is overly complex for such a simple application that only

needed 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 had the history of Loc-RIB changes. For example, a performance issue might have been seen for only a duration of 5 minutes. Post 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-paths is not enabled or if maximum number of equal paths are different from Loc-RIB to routes advertised.

This document adds Loc-RIB to the BGP Monitoring Protocol and replaces Section 8.2 [RFC7854] Locally Originated Routes.

1.1. Current 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. While it is possible to derive the Loc-RIB, it is also error prone and complex.

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. 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.

Unnecessary resources needed for current method:

- 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, VRF's with no peers, redistributed bgp-ls with no peers, segment routing egress peer engineering where no peers have link-state address family enabled.

Complexities introduced with current method in order to derive (e.g. correlate) peer to 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 specifics, 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 erroneous 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 a Loc-RIB to 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, ASNs, and BGP-ID's). This leads to error prone correlations.

- The BGP-ID's and session addresses to router correlation requires additional data, such as router inventory. This additional data provides the BMP receiver the ability to map and correlate the BGP-ID's 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 effects the integrity of the correlation.

2. Terminology

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].

3. Definitions

- 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 [RFC4271], "The Loc-RIB contains the routes that have been selected by the local BGP speaker's Decision Process." It is further defined that the routes selected include locally originated and routes from all peers.

- 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.

- 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

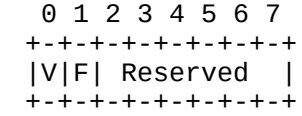
This document defines the following new peer type:

- Peer Type = 3: Loc-RIB Instance Peer

4.2. Peer Flags

In section 4.2 [RFC7854], the "locally sourced routes" comment in the L flag description is removed. Locally sourced routes 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:



- The V flag indicates that the Peer address is an IPv6 address. For IPv4 peers, this is set to 0.

- The F flag indicates that the Loc-RIB is filtered. This indicates that the Loc-RIB does not represent the complete routing table.

The remaining bits are reserved for future use. They MUST be transmitted as 0 and their values MUST be ignored on receipt.

5. Loc-RIB Monitoring

Loc-RIB contains all routes from BGP peers as well as any and all routes redistributed or otherwise locally originated. In this context, only the BGP instance Loc-RIB is included. Routes from other routing protocols that have not been redistributed or received via Adj-RIB-In are not considered.

5.1. Per-Peer Header

All peer messages that include a per-peer header 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 as remote peer address is not applicable.
- Peer AS: Set to the BGP instance global or default ASN value.
- Peer BGP ID: Set to the BGP instance global or RD (e.g. VRF) specific router-id.

5.2. Peer UP Notification

Peer UP notifications follow section 4.10 [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 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-paths is enabled for IPv6 and Loc-RIB contains additional paths, the add-paths capability should be included for IPv6. In the case of add-paths, 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.

- o Received OPEN Message: Repeat of the same Sent Open Message. The duplication allows the BMP receiver to use existing parsing.

5.2.1. Peer UP Information

The following peer UP information TLV Type is added:

- o Type = 3: VRF Name. The Information field contains an ASCII string whose value MUST be equal to the value of the VRF name (e.g. RD instance name) configured. This type is only relevant and used when the Loc-RIB represents a VRF/RD instance.

It is RECOMMENDED that the VRF Name be defined as "global" for the global/default Loc-RIB instance.

5.3. Peer Down Notification

Peer down notification SHOULD follow the section 4.9 [RFC7854] reason 2.

5.4. Route Monitoring

Route Monitoring messages are used for initial synchronization of the Loc-RIB. They are also used for incremental updates upon every change to the RIB. State compression on interval, such as 1 or greater seconds, can mask critical RIB changes. Therefore state compression SHOULD be avoided. If the Loc-RIB changes, a route monitor message should be sent.

As defined in section 4.3 [RFC7854], "Following the common BMP header and per-peer header is a BGP Update PDU."

5.5. Route Mirroring

Route mirroring is not applicable to Loc-RIB.

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 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 local router emulated peer.

6.1.1 Multiple Loc-RIB Peers

There MUST be multiple emulated peers for each Loc-RIB instance, such as with VRF's. The BMP receiver identifies the Loc-RIB's by the peer header distinguisher and BGP ID. The BMP receiver uses the VRF Name from the PEER UP to name 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 it represents 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 maybe be use-cases where BMP receivers should only receive specific routes from Loc-RIB. For example, IPv4 unicast routes may include IBGP, 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 partial.

7. Security Considerations

It is not believed that this document adds any additional security considerations.

8. IANA Considerations

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

- Peer Type = 3: Loc-RIB Instance Peer

9. References

9.1. URIs

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

9.2. 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, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc7854>.

9.3. Informative References

[I-ID.ietf-grow-bmp-adj-rib-out] TBD.

Acknowledgments

TBD.

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

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

Email: manbhard@cisco.com

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

Email: paolo@ntt.net

Global Routing Operations
Internet-Draft
Intended status: Best Current Practice
Expires: September 28, 2017

W. Hargrave
LONAP
M. Griswold
20C
J. Snijders
NTT
N. Hilliard
INEX

March 27, 2017

Mitigating Negative Impact of Maintenance through BGP Session Culling
draft-iops-grow-bgp-session-culling-01

Abstract

This document outlines an approach to mitigate negative impact on networks resulting from maintenance activities. It includes guidance for both IP networks and Internet Exchange Points (IXPs). The approach is to ensure BGP-4 sessions affected by the maintenance are forcefully torn down before the actual maintenance activities commence.

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 http://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 September 28, 2017.

Table of Contents

1. Introduction

BGP Session Culling is the practice of ensuring BGP sessions are forcefully torn down before maintenance activities on a lower layer network commence, which otherwise would affect the flow of data between the BGP speakers.

BGP Session Culling ensures that lower layer network maintenance activities cause the minimum possible amount of disruption, by causing BGP speakers to preemptively gracefully converge onto alternative paths while the lower layer network's forwarding plane remains fully operational.

The grace period required for a successful application of BGP Session Culling is the sum of the time needed to detect the loss of the BGP session, plus the time required for the BGP speaker to converge onto alternative paths. The first value is governed by the BGP Hold Timer (section 6.5 of [RFC4271]), commonly between 90 and 180 seconds, The

second value is implementation specific, but could be as much as 15 minutes when a router with a slow control-plane is receiving a full set of Internet routes.

Throughout this document the "Caretaker" is defined to be the operator of the lower layer network, while "Operators" directly administrate the BGP speakers. Operators and Caretakers implementing BGP Session Culling are encouraged to avoid using a fixed grace period, but instead monitor forwarding plane activity while the culling is taking place and consider it complete once traffic levels have dropped to a minimum (Section 3.3).

2. 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].

3. BGP Session Culling

From the viewpoint of the IP network operator, there are two types of BGP Session Culling:

Voluntary BGP Session Teardown: The operator initiates the tear down of the potentially affected BGP session by issuing an Administrative Shutdown.

Involuntary BGP Session Teardown: The caretaker of the lower layer network disrupts BGP control-plane traffic in the upper layer, causing the BGP Hold Timers of the affected BGP session to expire, subsequently triggering rerouting of end user traffic.

3.1. Voluntary BGP Session Teardown Recommendations

Before an operator commences activities which can cause disruption to the flow of data through the lower layer network, an operator can reduce loss of traffic by issuing an Administratively Shutdown to all BGP sessions running across the lower layer network and wait a few minutes for data-plane traffic to subside.

While architectures exist to facilitate quick network reconvergence (such as BGP PIC [I-D.ietf-rtgwg-bgp-pic]), an operator cannot assume the remote side has such capabilities. As such, a grace period between the Administrative Shutdown and the impacting maintenance activities is warranted.

After the maintenance activities have concluded, the operator is expected to restore the BGP sessions to their original Administrative state.

3.1.1. Maintenance Considerations

Initiators of the Administrative Shutdown could consider to use [Graceful Shutdown] to facilitate smooth drainage of traffic prior to session tear down, and the Shutdown Communication [I-D.ietf-idr-shutdown] to inform the remote side on the nature and duration of the maintenance activities.

3.2. Involuntary BGP Session Teardown Recommendations

In the case where multilateral interconnection between BGP speakers is facilitated through a switched layer-2 fabric, such as commonly seen at Internet Exchange Points (IXPs), different operational considerations can apply.

Operational experience shows many network operators are unable to carry out the Voluntary BGP Session Teardown recommendations, because of the operational cost and risk of co-ordinating the two configuration changes required. This has an adverse affect on Internet performance.

In the absence of notifications from the lower layer (e.g. ethernet link down) consistent with the planned maintenance activities in a densely meshed multi-node layer-2 fabric, the caretaker of the fabric could opt to cull BGP sessions on behalf of the stakeholders connected to the fabric.

Such culling of control-plane traffic will pre-empt the loss of end-user traffic, by causing the expiration of BGP Hold Timers ahead of the moment where the expiration would occur without intervention from the fabric’s caretaker.

In this scenario, BGP Session Culling is accomplished through the application of a combined layer-3 and layer-4 packet filter deployed in the switched fabric itself.

3.2.1. Packet Filter Considerations

The following considerations apply to the packet filter design:

- The packet filter MUST only affect BGP traffic specific to the layer-2 fabric, i.e. forming part of the control plane of the system described, rather than multihop BGP traffic which merely transits

- The packet filter MUST only affect BGP, i.e. TCP/179
- The packet filter SHOULD make provision for the bidirectional nature of BGP, i.e. that sessions may be established in either direction
- The packet filter MUST affect all relevant AFIs

Appendix A contains examples of correct packet filters for various platforms.

3.2.2. Hardware Considerations

Not all hardware is capable of deploying layer 3 / layer 4 filters on layer 2 ports, and even on platforms which support the feature, documented limitations may exist or hardware resource allocation failures may occur during filter deployment which may cause unexpected results. These problems may include:

- Platform inability to apply layer 3/4 filters on ports which already have layer 2 filters applied
- Layer 3/4 filters supported for IPv4 but not for IPv6
- Layer 3/4 filters supported on physical ports, but not on 802.3ad Link Aggregate ports
- Failure of the operator to apply filters to all 802.3ad Link Aggregate ports
- Limitations in ACL hardware mechanisms causing filters not to be applied
- Fragmentation of ACL lookup memory causing transient ACL application problems which are resolved after ACL removal / reapplication
- Temporary service loss during hardware programming
- Reduction in hardware ACL capacity if the platform enables lossless ACL application

It is advisable for the operator to be aware of the limitations of their hardware, and to thoroughly test all complicated configurations in advance to ensure that problems don’t occur during production deployments.

3.3. Procedural Considerations

The caretaker of the lower layer can monitor data-plane traffic (e.g. interface counters) and carry out the maintenance without impact to traffic once session culling is complete.

It is recommended that the packet filters are only deployed for the duration of the maintenance and immediately removed after the maintenance. To prevent unnecessarily troubleshooting, it is RECOMMENDED that caretakers notify the affected operators before the maintenance takes place, and make it explicit that the Involuntary BGP Session Culling methodology will be applied.

4. Acknowledgments

The authors would like to thank the following people for their contributions to this document: Saku Ytti, Greg Hankins, James Bensley, Wolfgang Tremmel, Daniel Roesen, Bruno Decraene, and Tore Anderson.

5. Security Considerations

There are no security considerations.

6. IANA Considerations

This document has no actions for IANA.

7. References

7.1. 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, <http://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, <http://www.rfc-editor.org/info/rfc4271>.

7.2. Informative References

[I-D.ietf-idr-shutdown] Snijders, J., Heitz, J., and J. Scudder, "BGP Administrative Shutdown Communication", draft-ietf-idr-shutdown-07 (work in progress), March 2017.

[I-D.ietf-rtgwg-bgp-pic]
Bashandy, A., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", draft-ietf-rtgwg-bgp-pic-01 (work in progress), June 2016.

7.3. URIs

[1] https://github.com/bgp/bgp-session-culling-config-examples

Appendix A. Example packet filters

Example packet filters for "Involuntary BGP Session Teardown" at an IXP with LAN prefixes 192.0.2.0/24 and 2001:db8:2::/64.

A repository of configuration examples for a number of assorted platforms can be found at github.com/bgp/bgp-session-culling-config-examples [1].

A.1. Cisco IOS, IOS XR & Arista EOS Firewall Example Configuration

```
ipv6 access-list acl-ipv6-permit-all-except-bgp
   10 deny tcp 2001:db8:2::/64 eq bgp 2001:db8:2::/64
   20 deny tcp 2001:db8:2::/64 2001:db8:2::/64 eq bgp
   30 permit ipv6 any any
!
ip access-list acl-ipv4-permit-all-except-bgp
   10 deny tcp 192.0.2.0/24 eq bgp 192.0.2.0/24
   20 deny tcp 192.0.2.0/24 192.0.2.0/24 eq bgp
   30 permit ip any any
!
interface Ethernet33
   description IXP Participant Affected by Maintenance
   ip access-group acl-ipv4-permit-all-except-bgp in
   ipv6 access-group acl-ipv6-permit-all-except-bgp in
!
```

A.2. Nokia SR OS Filter Example Configuration

```
   ip-filter 10 create
       filter-name "ACL IPv4 Permit All Except BGP"
       default-action forward
       entry 10 create
           match protocol tcp
               dst-ip 192.0.2.0/24
               src-ip 192.0.2.0/24
               port eq 179
           exit
           action
               drop
           exit
       exit
   exit

   ipv6-filter 10 create
       filter-name "ACL IPv6 Permit All Except BGP"
       default-action forward
       entry 10 create
           match next-header tcp
               dst-ip 2001:db8:2::/64
               src-ip 2001:db8:2::/64
               port eq 179
           exit
           action
               drop
           exit
       exit
   exit

   interface "port-1/1/1"
       description "IXP Participant Affected by Maintenance"
       ingress
           filter ip 10
           filter ipv6 10
       exit
   exit
```

Authors' Addresses

Will Hargrave
LONAP Ltd
5 Fleet Place
London EC4M 7RD
United Kingdom

Email: will@lonap.net

Matt Griswold
20C
1658 Milwaukee Ave # 100-4506
Chicago, IL 60647
United States of America

Email: grizz@20c.com

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

Email: job@ntt.net

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
Ireland

Email: nick@inex.ie