

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: June 17, 2018

P. Francois, Ed.
Individual Contributor
B. Decraene, Ed.
Orange
C. Pelsser
Strasbourg University
K. Patel
Arrcus, Inc.
C. Filsfils
Cisco Systems
December 14, 2017

Graceful BGP session shutdown
draft-ietf-grow-bgp-gshut-13

Abstract

This draft standardizes a new well-known BGP community `GRACEFUL_SHUTDOWN` to signal the graceful shutdown of paths. This draft also describes operational procedures which use this community to reduce the amount of traffic lost when BGP peering sessions are about to be shut down deliberately, e.g. for planned maintenance.

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 June 17, 2018.

Copyright Notice

Copyright (c) 2017 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. Packet loss upon manual EBGp session shutdown	3
4. EBGp graceful shutdown procedure	4
4.1. Pre-configuration	4
4.2. Operations at maintenance time	4
4.3. BGP implementation support for graceful shutdown	5
5. IANA Considerations	5
6. Security Considerations	5
7. Acknowledgments	6
8. References	6
8.1. Normative References	6
8.2. Informative References	6
Appendix A. Alternative techniques with limited applicability	7
A.1. Multi Exit Discriminator tweaking	7
A.2. IGP distance Poisoning	7
Appendix B. Configuration Examples	7
B.1. Cisco IOS XR	7
B.2. BIRD	8
B.3. OpenBGPD	8
Appendix C. Beyond EBGp graceful shutdown	9
C.1. IBGP graceful shutdown	9
C.2. EBGp session establishment	9
Authors' Addresses	10

1. Introduction

Routing changes in BGP can be caused by planned maintenance operations. This document defines a well-known community [RFC1997], called `GRACEFUL_SHUTDOWN`, for the purpose of reducing the management overhead of gracefully shutting down BGP sessions. The well-known community allows implementers to provide an automated graceful shutdown mechanism that does not require any router reconfiguration at maintenance time.

This document discusses operational procedures to be applied in order to reduce or eliminate loss of packets during a maintenance operation. Loss comes from transient lack of reachability during BGP

convergence which follows the shutdown of an EBGp peering session between two Autonomous System Border Routers (ASBR).

This document presents procedures for the cases where the forwarding plane is impacted by the maintenance, hence when the use of Graceful Restart does not apply.

The procedures described in this document can be applied to reduce or avoid packet loss for outbound and inbound traffic flows initially forwarded along the peering link to be shut down. These procedures trigger, in both Autonomous Systems (AS), rerouting to alternate paths if they exist within the AS, while allowing the use of the old path until alternate ones are learned. This ensures that routers always have a valid route available during the convergence process.

The goal of the document is to meet the requirements described in [RFC6198] at best, without changing the BGP protocol.

Other maintenance cases, such as the shutdown of an IBGP session or the establishment of an EBGp session, are out of scope of this document. For information, they are briefly discussed in Appendix C.

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 [BCP14] [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Terminology

graceful shutdown initiator: a router on which the session shutdown is performed for the maintenance.

graceful shutdown receiver: a router that has a BGP session, to be shutdown, with the graceful shutdown initiator.

3. Packet loss upon manual EBGp session shutdown

Packets can be lost during the BGP convergence following a manual shutdown of an EBGp session for two reasons.

First, some routers can have no path toward an affected prefix, and drop traffic destined to this prefix. This is because alternate paths can be hidden by nodes of an AS. This happens when [RFC7911] is not used and the paths are not selected as best by the ASBR that receive them on an EBGp session, or by Route Reflectors that do not propagate them further in the IBGP topology because they do not select them as best.

Second, the FIB can be inconsistent between routers within the AS, and packets toward affected prefixes can loop and be dropped unless encapsulation is used within the AS.

This document only addresses the first reason.

4. EBGP graceful shutdown procedure

This section describes configurations and actions to be performed for the graceful shutdown of EBGP peering links.

The goal of this procedure is to retain the paths to be shutdown between the peers, but with a lower LOCAL_PREF value, allowing the paths to remain in use while alternate paths are selected and propagated, rather than simply withdrawing the paths. The LOCAL_PREF value SHOULD be lower than any of the alternative paths. The RECOMMENDED value is 0.

Note that some alternative techniques with limited applicability are discussed for information in Appendix A.

4.1. Pre-configuration

On each ASBR supporting the graceful shutdown receiver procedure, an inbound BGP route policy is applied on all EBGP sessions of the ASBR, that:

- o matches the GRACEFUL_SHUTDOWN community.
- o sets the LOCAL_PREF attribute of the paths tagged with the GRACEFUL_SHUTDOWN community to a low value.

For information purpose, example of configurations are provided in Appendix B.

4.2. Operations at maintenance time

On the graceful shutdown initiator, at maintenance time, the operator:

- o applies an outbound BGP route policy on the EBGP session to be shutdown. This policy tags the paths propagated over the session with the GRACEFUL_SHUTDOWN community. This will trigger the BGP implementation to re-advertise all active routes previously advertised, and tag them with the GRACEFUL_SHUTDOWN community.
- o applies an inbound BGP route policy on the EBGP session to be shutdown. This policy tags the paths received over the session

with the GRACEFUL_SHUTDOWN community and sets LOCAL_PREF to a low value.

- o wait for route readvertisement over the EBGp session, and BGP routing convergence on both ASBRs.
- o shutdown the EBGp session, optionally using [RFC8203] to communicate the reason of the shutdown.

In the case of a shutdown of the whole router, in addition to the graceful shutdown of all EBGp sessions, there is a need to gracefully shutdown the routes originated by this router (e.g, BGP aggregates redistributed from other protocols, including static routes). This can be performed by tagging these routes with the GRACEFUL_SHUTDOWN community and setting LOCAL_PREF to a low value.

4.3. BGP implementation support for graceful shutdown

BGP Implementers SHOULD provide configuration knobs that utilize the GRACEFUL_SHUTDOWN community to drain BGP neighbors in preparation of impending neighbor shutdown. Implementation details are outside the scope of this document.

5. IANA Considerations

The IANA has assigned the community value 0xFFFF0000 to the planned-shut community in the "BGP Well-known Communities" registry. IANA is requested to change the name planned-shut to GRACEFUL_SHUTDOWN and set this document as the reference.

6. Security Considerations

By providing the graceful shutdown service to a neighboring AS, an ISP provides means to this neighbor and possibly its downstream ASes to lower the LOCAL_PREF value assigned to the paths received from this neighbor.

The neighbor could abuse the technique and do inbound traffic engineering by declaring some prefixes as undergoing a maintenance so as to switch traffic to another peering link.

If this behavior is not tolerated by the ISP, it SHOULD monitor the use of the graceful shutdown community.

7. Acknowledgments

The authors wish to thank Olivier Bonaventure, Pradosh Mohapatra, Job Snijders, John Heasley, and Christopher Morrow for their useful comments.

8. References

8.1. Normative References

- [RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <<https://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, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC6198] Decraene, B., Francois, P., Pelsser, C., Ahmad, Z., Elizondo Armengol, A., and T. Takeda, "Requirements for the Graceful Shutdown of BGP Sessions", RFC 6198, DOI 10.17487/RFC6198, April 2011, <<https://www.rfc-editor.org/info/rfc6198>>.
- [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>>.

8.2. Informative References

- [I-D.ietf-idr-best-external] Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H. Gredler, "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-05 (work in progress), January 2012.
- [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>>.
- [RFC8203] Snijders, J., Heitz, J., and J. Scudder, "BGP Administrative Shutdown Communication", RFC 8203, DOI 10.17487/RFC8203, July 2017, <<https://www.rfc-editor.org/info/rfc8203>>.

Appendix A. Alternative techniques with limited applicability

A few alternative techniques have been considered to provide graceful shutdown capabilities but have been rejected due to their limited applicability. This section describes them for possible reference.

A.1. Multi Exit Discriminator tweaking

The MED attribute of the paths to be avoided can be increased so as to force the routers in the neighboring AS to select other paths.

The solution only works if the alternate paths are as good as the initial ones with respect to the LOCAL_PREF value and the AS Path Length value. In the other cases, increasing the MED value will not have an impact on the decision process of the routers in the neighboring AS.

A.2. IGP distance Poisoning

The distance to the BGP NEXT_HOP corresponding to the maintained session can be increased in the IGP so that the old paths will be less preferred during the application of the IGP distance tie-break rule. However, this solution only works for the paths whose alternates are as good as the old paths with respect to their LOCAL_PREF value, their AS Path length, and their MED value.

Also, this poisoning cannot be applied when BGP NEXT_HOP self is used as there is no BGP NEXT_HOP specific to the maintained session to poison in the IGP.

Appendix B. Configuration Examples

This appendix is non-normative.

Example routing policy configurations to honor the GRACEFUL_SHUTDOWN well-known BGP community.

B.1. Cisco IOS XR

```
community-set comm-graceful-shutdown
  65535:0
end-set
!
route-policy AS64497-ebgp-inbound
  ! normally this policy would contain much more
  if community matches-any comm-graceful-shutdown then
    set local-preference 0
  endif
end-policy
!
router bgp 64496
  neighbor 2001:db8:1:2::1
    remote-as 64497
    address-family ipv6 unicast
      send-community-ebgp
      route-policy AS64497-ebgp-inbound in
    !
  !
!
```

B.2. BIRD

```
function honor_graceful_shutdown() {
  if (65535, 0) ~ bgp_community then {
    bgp_local_pref = 0;
  }
}
filter AS64497_ebgp_inbound
{
  # normally this policy would contain much more
  honor_graceful_shutdown();
}
protocol bgp peer_64497_1 {
  neighbor 2001:db8:1:2::1 as 64497;
  local as 64496;
  import keep filtered;
  import filter AS64497_ebgp_inbound;
}
```

B.3. OpenBGPD


```
AS 64496
router-id 192.0.2.1
neighbor 2001:db8:1:2::1 {
    remote-as 64497
}
# normally this policy would contain much more
match from any community GRACEFUL_SHUTDOWN set { localpref 0 }
```

Appendix C. Beyond EBGp graceful shutdown

C.1. IBGP graceful shutdown

For the shutdown of an IBGP session, provided the IBGP topology is viable after the maintenance of the session, i.e, if all BGP speakers of the AS have an IBGP signaling path for all prefixes advertised on this graceful shutdown IBGP session, then the shutdown of an IBGP session does not lead to transient unreachability. As a consequence, no specific graceful shutdown action is required.

C.2. EBGp session establishment

We identify two potential causes for transient packet losses upon the establishment of an EBGp session. The first one is local to the startup initiator, the second one is due to the BGP convergence following the injection of new best paths within the IBGP topology.

C.2.1. Unreachability local to the ASBR

An ASBR that selects as best a path received over a newly established EBGp session may transiently drop traffic. This can typically happen when the NEXT_HOP attribute differs from the IP address of the EBGp peer, and the receiving ASBR has not yet resolved the MAC address associated with the IP address of that "third party" NEXT_HOP.

A BGP speaker implementation MAY avoid such losses by ensuring that "third party" NEXT_HOPs are resolved before installing paths using these in the RIB.

Alternatively, the operator (script) MAY ping third party NEXT_HOPs that are expected to be used before establishing the session. By proceeding like this, the MAC addresses associated with these third party NEXT_HOPs are resolved by the startup initiator.

C.2.2. IBGP convergence

During the establishment of an EBGp session, in some corner cases a router may have no path toward an affected prefix, leading to loss of connectivity.

A typical example for such transient unreachability for a given prefix is the following:

Let's consider three Route Reflectors (RR): RR1, RR2, RR3. There is a full mesh of IBGP sessions between them.

1. RR1 is initially advertising the current best path to the members of its IBGP RR full-mesh. It propagated that path within its RR full-mesh. RR2 knows only that path toward the prefix.
2. RR3 receives a new best path originated by the startup initiator, being one of its RR clients. RR3 selects it as best, and propagates an UPDATE within its RR full-mesh, i.e., to RR1 and RR2.
3. RR1 receives that path, reruns its decision process, and picks this new path as best. As a result, RR1 withdraws its previously announced best-path on the IBGP sessions of its RR full-mesh.
4. If, for any reason, RR3 processes the withdraw generated in step 3, before processing the update generated in step 2, RR3 transiently suffers from unreachability for the affected prefix.

The use of [RFC7911] or [I-D.ietf-idr-best-external] among the RR of the IBGP full-mesh can solve these corner cases by ensuring that within an AS, the advertisement of a new route is not translated into the withdraw of a former route.

Indeed, "best-external" ensures that an ASBR does not withdraw a previously advertised (EBGP) path when it receives an additional, preferred path over an IBGP session. Also, "best-intra-cluster" ensures that a RR does not withdraw a previously advertised (IBGP) path to its non clients (e.g. other RRs in a mesh of RR) when it receives a new, preferred path over an IBGP session.

Authors' Addresses

Pierre Francois (editor)
Individual Contributor

Email: pfrpfr@gmail.com

Bruno Decraene (editor)
Orange

Email: bruno.decraene@orange.com

Cristel Pelsser
Strasbourg University

Email: pelsser@unistra.fr

Keyur Patel
Arrcus, Inc.

Email: keyur@arrcus.com

Clarence Filsfils
Cisco Systems

Email: cfilsfil@cisco.com

Global Routing Operations
Internet-Draft
Intended status: Best Current Practice
Expires: April 1, 2018

W. Hargrave
LONAP
M. Griswold
20C
J. Snijders
NTT
N. Hilliard
INEX
September 28, 2017

Mitigating Negative Impact of Maintenance through BGP Session Culling
draft-ietf-grow-bgp-session-culling-05

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 <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 April 1, 2018.

Copyright Notice

Copyright (c) 2017 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. Requirements Language	3
3. BGP Session Culling	3
3.1. Voluntary BGP Session Teardown Recommendations	3
3.1.1. Maintenance Considerations	4
3.2. Involuntary BGP Session Teardown Recommendations	4
3.2.1. Packet Filter Considerations	4
3.2.2. Hardware Considerations	5
3.3. Procedural Considerations	6
4. Acknowledgments	6
5. Security Considerations	6
6. IANA Considerations	6
7. References	6
7.1. Normative References	6
7.2. Informative References	7
Appendix A. Example packet filters	7
A.1. Cisco IOS, IOS XR & Arista EOS Firewall Example Configuration	7
A.2. Nokia SR OS Filter Example Configuration	8
Authors' Addresses	8

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 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 often 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 in control 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 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 (higher layer) BGP control-plane traffic, 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 administrative 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 MAY consider using Graceful Shutdown [I-D.ietf-grow-bgp-gshut] to facilitate smooth drainage of traffic prior to session tear down, and the Shutdown Communication [RFC8203] 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 Operators are unable to carry out the Voluntary BGP Session Teardown recommendations, because of the operational cost and risk of coordinating 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 switched layer-2 fabric, the Caretaker of the fabric could choose to cull BGP sessions on behalf of the Operators connected to the fabric.

Such culling of control-plane traffic will preempt 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 as described in the next sub-section, through the application of a combined layer-3 and layer-4 packet filter deployed in the Caretaker's switched fabric.

3.2.1. Packet Filter Considerations

The peering LAN prefixes used by the IXP form the control plane, and following considerations apply to the packet filter design:

- o 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.

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

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 claim support for such a feature, limitations may exist or hardware resource allocation failures may occur during filter deployment which may cause unexpected results. These problems may include:

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

It is advisable for the Caretaker 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 network 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, Tore Anderson, John Heasley, Warren Kumari, Stig Venaas, and Brian Carpenter.

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, <<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>>.

7.2. Informative References

- [I-D.ietf-grow-bgp-gshut]
Francois, P., Decraene, B., Pelsser, C., Patel, K., and C. Filsfils, "Graceful BGP session shutdown", draft-ietf-grow-bgp-gshut-11 (work in progress), September 2017.
- [I-D.ietf-rtgwg-bgp-pic]
Bashandy, A., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", draft-ietf-rtgwg-bgp-pic-05 (work in progress), May 2017.
- [RFC8203] Snijders, J., Heitz, J., and J. Scudder, "BGP Administrative Shutdown Communication", RFC 8203, DOI 10.17487/RFC8203, July 2017, <<https://www.rfc-editor.org/info/rfc8203>>.

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 using peering LAN prefixes 192.0.2.0/24 and 2001:db8:2::/64 as its control plane.

A repository of configuration examples for a number of assorted platforms can be found at <https://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

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.

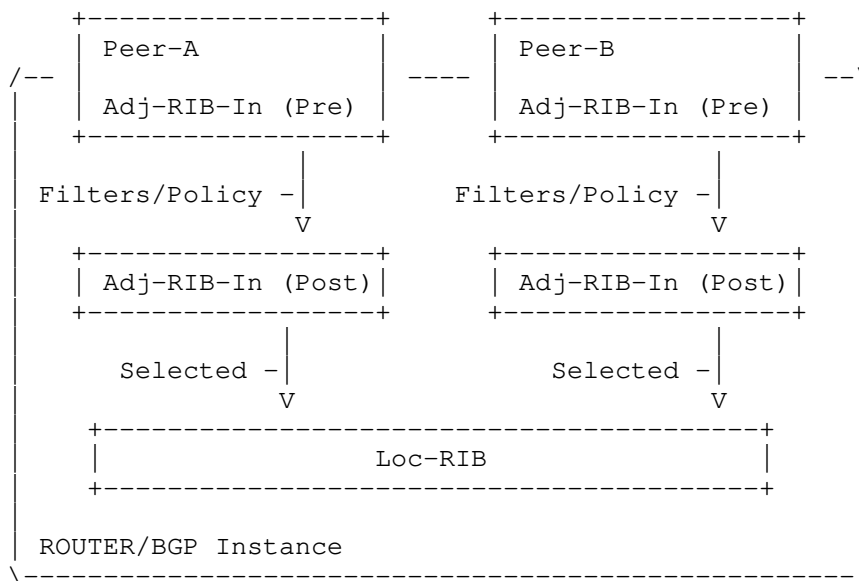


Figure 1: BGP peering Adj-RIBs-In into Loc-RIB

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.

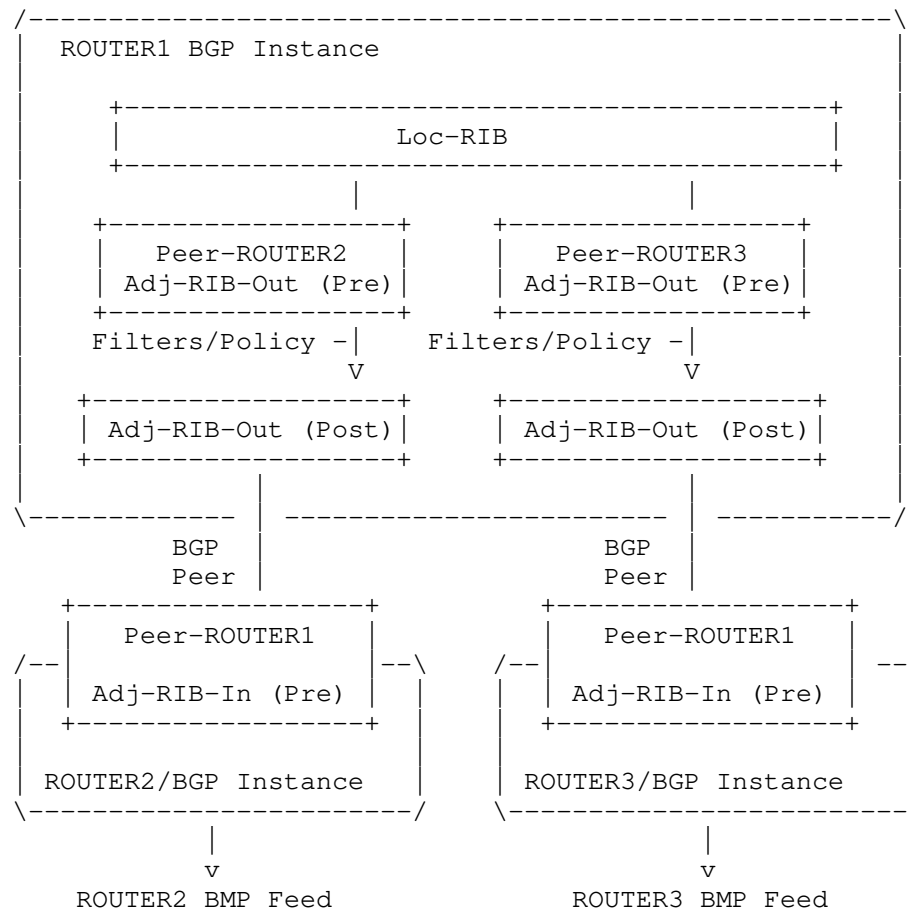


Figure 2: Alternative method to monitor Loc-RIB

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