Abstract

This document proposes to enhance AS-Loop Detection for BGP Inbound/Outbound Route Processing. This could empower networks to quickly and accurately figure out they’re being victimized.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on October 4, 2019.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.
The Border Gateway Protocol (BGP) [RFC4271], as an inter-Autonomous (AS) routing protocol, is used to exchange network reachability information between BGP systems. BGP is widely used by Internet Service Providers (ISPs) and large organizations.

BGP is used to exchange reachable inter-AS routes, establish inter-AS paths, avoid routing loops, and apply routing policies between ASs. BGP loop detection mechanism is defined in section 9.1.2. of RFC4271:

... If the AS_PATH attribute of a BGP route contains an AS loop, the BGP route should be excluded from the Phase 2 decision function. AS loop detection is done by scanning the full AS path (as specified in the AS_PATH attribute), and checking that the autonomous system number of the local system does not appear in the AS path. Operations of a BGP speaker that is configured to accept routes with its own autonomous system number in the AS path are outside the scope of this document. ...

Chen, et al.  Expires October 4, 2019
In ordinary BGP, every AS announces its route information with different prefixes. However, its neighboring ASes cannot validate this route information, but rather directly propagate it across the Internet or simply discard AS-Loop routes directly. Obviously, this weak trust model allows forged route announcement propagations and rarely been found, which is a fundamental security weakness of BGP. Forged routes, which can be generated by configuration errors or malicious attacks, can cause large-scale network connectivity problems.

Some cases can be worse, hackers exploit this property of BGP to achieve their ulterior motives. They can add some providers’ AS number into the forged AS-Path and attempt to make it look like the route had passed through these ASNs, or perhaps they are there to prevent those providers from carrying the route.

For example, the cases shown in Figure 1.

- Forged Case 1: One upstream ISP of AS64596 forged a route with the ASN 64596 as the origin ASN in the AS-Path.
- Forged Case 2: One upstream ISP of AS64596 forged a route with the ASN 64596 as the transit ASN in the AS-Path.

After receiving the above routes, AS64596 treats them as normal loop routes during the loop detecting phase and discards them directly. In most NOSes (Network Operation Systems), such rejected routes are not logged and only visible by putting the router into debugging mode. If the AS64596 is slightly enhanced, it can find that someone has faked himself, which may cause unnecessary trouble for himself.
AS-Loop-Detecting at this point
Discard AS-Loop Routes directly that contains AS64596

v
x.y.z.0/24 Origin AS 64600

AS64595---AS64596---AS64597---AS64598---AS64599----AS64600

Normal Case:
--- x.y.z.0/24, AS-Path: 64597 64598 64599 64600

Forged Case 1:
--- x.y.z.0/24, AS-Path: 64597 64596
(Or: 64597 64598 64596 etc.)

Forged Case 2:
--- x.y.z.0/24, AS-Path: 64597 64596 64600
(Or: 64597 64596 64599 64600 etc.)

Figure 1: BGP Inbound Route Processing

Split-Horizon for EBGP is an optional function that a BGP sender will
not advertise any routes that were previously received from that same
AS. In some current implementation, the BGP outbound route
processing step will simply discard the route if AS-Loop being
detected.

For example, the cases shown in Figure 1.

- Forged Case 1: One upstream ISP of AS64597 forged a route with the
  ASN 64596 as the origin ASN in the AS-Path.

- Forged Case 2: One upstream ISP of AS64597 forged a route with the
  ASN 64596 as the transit ASN in the AS-Path.

When sending the above routes, AS64597 treats them as normal loop
routes and discards them directly. If AS64597 is slightly enhanced,
it can find that someone has faked AS64596, which may cause large-
scale network connectivity problems.
Split-Horizon Enable & AS-Loop-Detecting at this point
Discard AS-Loop Routes directly if sending AS-Path contains AS64596

v x.y.z.0/24 Origin AS 64600
AS64595---AS64596---AS64597---AS64598---AS64599----AS64600
Normal Case:
  --- x.y.z.0/24, AS-Path: 64597 64598 64599 64600
Forged Case 1:
  --- x.y.z.0/24, AS-Path: 64597 64596
      (Or: 64597 64598 64596 etc.)
Forged Case 2:
  --- x.y.z.0/24, AS-Path: 64597 64596 64600
      (Or: 64597 64596 64599 64600 etc.)

Figure 2: BGP Outbound Route Processing

Above cases are also being known As-Path Poisoning Attacks.

This document proposes to enhance AS-Loop Detection for BGP Inbound/Outbound Route Processing. This could empower networks to quickly and accurately figure out they’re being victimized.

2. Terminology

The following terminology is used in this document.

AS: Autonomous System

BGP: Border Gateway Protocol

BGP hijacking : is the illegitimate takeover of groups of IP addresses by corrupting Internet routing tables maintained using the Border Gateway Protocol (BGP). (Sometimes referred to as prefix hijacking, route hijacking or IP hijacking)

EBGP: External BGP

ISP: Internet Service Provider

3. Enhanced AS-Loop Detection for BGP Inbound Route Processing

This section proposes to enhance AS Loop Detection for BGP Inbound Route Processing.
As shown in Figure 3, when receiving the routes from AS64597, AS64596 should check whether its AS number is already in the AS-Path. If yes, it further analyzes the location of the AS64596 in the received AS_Path:

Case 1: AS 64596 is listed as Origin AS in the AS-Path

Lookup the local resource database (Such as ROA Cache) and determine whether the route is originated from the AS 64596.

- Result 1: AS 64596 has no corresponding prefix; it is identified as a purely forged AS_Path prefix hijacking event, which is recorded as incident type 1.

- Result 2: The corresponding prefix is a sub-prefix of a certain prefix of the AS 64596 and the AS 64596 has not advertise it. For example, the prefix being hold by the AS 64596 is 10.10.128.0/17, and the receiving route prefix is 10.10.192.0/24, the latter is a sub-prefix of the former, which indicates that this is a forged AS_Path sub-prefix hijacking event, which is recorded as incident type 2.

- Result 3: The corresponding prefix is a sub-prefix of a certain prefix of the AS 64596 and the AS 64596 has only advertised to some special ASNs, and only wants it to be used internally by those ASNs. The AS 64596 recognizes that At least one special AS violates the route policy. Which is recorded as incident type 3.

- Result 4: The corresponding prefix is originated by the AS 64596, this is the normal case.

Case 2: AS 64596 is listed as transit AS in the AS-Path

For example, AS-Path looks like the following form AS64596’s perspective:

(possible other AS), left AS, local AS(64596), right AS, (possible other AS)

At this point, AS 64596 can lookup the local resource database and check whether there is a real AS relationship between the local AS and the left AS and the right AS. (From the perspective of the local AS, it can manage/hold the AS-relationship database between the local AS and each of its neighboring ASs (such as C2P, P2P, P2C, etc.).)

- Result 1: At least one of the AS (the left AS or the right AS) has no actual AS relationship with the local AS (i.e. A never
before seen AS-AS adjacency). It is a purely forged AS_PATH prefix hijacking event. Which is recorded as incident type 4.

- Result 2: The AS relationships between the local AS and the left AS and the right AS are correct, but the local AS has not previously process this prefix, so it can be recognized that this is a forged route. We classify this incident type as type 5.

- Result 3: The AS relationships between the AS and the left AS and the right AS are correct, and the local AS 64596 has previously processed the prefix, this is the normal case.

Enhanced AS-Loop-Detecting at this point
To identify the attack/forged information

```
  v   x.y.z.0/24 Origin AS 64600
AS64595---AS64596---AS64597---AS64598---AS64599----AS64600
```

Normal Case:
<-- x.y.z.0/24, AS-Path: 64597 64598 64599 64600

Forged Case 1:
<-- x.y.z.0/24, AS-Path: 64596 64595
(Or: 64597 64598 64596 etc.)

Forged Case 2:
<-- x.y.z.0/24, AS-Path: 64597 64596 64600
(Or: 64597 64596 64599 64600 etc.)

Figure 3: Enhance for BGP Inbound Route Processing

The local AS 64596 inputs the detected result to the route hijacking management module, or/and records the log or/and the alarm information, and the maintenance team of the local AS 64596 can notify the maintenance team of the relevant AS to correct the error in their networks.

After the above steps are added, the stability and security of the network can be improved.

4. Enhanced AS-Loop Detection for BGP Outbound Route Processing

This section proposes to enhance AS Loop Detection for BGP Outbound Route Processing.
If Split-Horizon Enable, Enhanced AS-Loop-Detecting at this point
To identify the attack/forged information

v                               x.y.z.0/24 Origin AS 64600
AS64595---AS64596---AS64597---AS64598---AS64599----AS64600
Normal Case:
<-- x.y.z.0/24, AS-Path: 300 64598 64599 64600
Forged Case 1:
<-- x.y.z.0/24, AS-Path: 64597 64596
   (Or: 64597 64598 64596 etc.)
Forged Case 2:
<-- x.y.z.0/24, AS-Path: 64597 64596 64600
   (Or: 64597 64596 64599 64600 etc.)

Figure 4: Enhance for BGP Outbound Route Processing

As shown in Figure 4, when sending the routes from AS64597 to
AS64596, AS64597 will check whether the AS number 64596 is already in
the AS-Path, If yes, it can further analyzes the location of the
AS64596 in the received AS_Path:

The remaining processing steps are the same as the previous section.

5. Benefits

After the enhancements of the AS Loop Detection for BGP Inbound/
Outbound Route Processing are added, the stability and security of
the network can be improved.

6. Acknowledgements

The authors would like to acknowledge the review and inputs from Gang
Yan, Zhenbin Li, Aijun Wang, Jeff Haas, Robert Razzuk, Alexander
Asimov and the working group.

7. IANA Considerations

TBD.
8. Security Considerations

TBD.

9. Normative References


Authors’ Addresses

Huanan Chen
China Telecom
109, West Zhongshan Road, Tianhe District
Guangzhou  510000
China

Email: chenhn8.gd@chinatelecom.cn

Di Ma
ZDNS
4 South 4th St. Zhongguancun
Beijing, Haidian
China

Email: madi@zdns.cn
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 December 25, 2019.

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
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 section 4.2 [RFC7854] per-peer header by adding a new flag to distinguish Adj-RIB-In verses 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

- 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.

4. 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:

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

- 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 [RFC7854] and the remaining bits are reserved for future use. They SHOULD be transmitted as 0 and their values MUST be ignored on receipt.

The following fields in the Per-Peer Header are redefined:

- Peer Address: The remote IP address associated with the TCP session over which the encapsulated PDU is sent.
- Peer AS: The Autonomous System number of the peer from which the encapsulated PDU was sent.
- Peer BGP ID: The BGP Identifier of the peer from which the encapsulated PDU was sent.
- 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 what is actually
transmitted to the peer, next-hop and any attributes set during transmission should also be set and transmitted to the BMP receiver.

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, EBGP 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 this 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:
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. BMP receiver implementations MUST use the per-peer header O flag in route monitoring and mirroring messages to identify if the message is for Adj-RIB-In or Adj-RIB-Out.

6.3.1. Peer Up Information

The following Peer Up message Information TLV type is added:

- Type = 4: Admin Label. The Information field contains a free-form UTF-8 string whose 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

It is not believed that this document adds 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):

- Stat Type = 14: (64-bit Gauge) Number of routes in Adj-RIBs-Out Pre-Policy.
- Stat Type = 15: (64-bit Gauge) Number of routes in Adj-RIBs-Out Post-Policy.
- 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.
- 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):

- Type = 4: Admin Label. The Information field contains a free-form UTF-8 string whose length is given by the Information Length field. The value is administratively 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.

10. References

10.1. Normative References

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
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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 9, 2019.

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.
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.

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.
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. At scale this becomes overly complex and error prone.
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 3, 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.

The current method 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, VRF’s with no peers, redistributed bgp-1s 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 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 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 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 Loc-RIB with or without RD and local instances. Section 4.2 [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

In section 4.2 [RFC7854], the "locally sourced routes" comment under 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:

```
  0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+-+-+
|F|   Reserved   |
+-+-+-+-+-+-+-+-+-+-+
```

- 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, originated by or into BGP, or received via Adj-RIB-In are not considered.

Loc-RIB in this context does not attempt to maintain a pre-policy and post-policy representation. Loc-RIB is the selected and used routes, which is equivalent to post-policy.

For example, VRF "Blue" imports several targets but filters out specific routes. The end result of VRF "Blue" Loc-RIB is conveyed. Even though the import is filtered, the result is complete for VRF "Blue" Loc-RIB. The F flag is not set in this case since the Loc-RIB is complete and not filtered to the BMP receiver.

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. Remote peer address is not applicable. The V flag is not applicable with Local-RIB Instance peer type considering addresses are zero-filed.
- 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.
- 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:

- Type = 3: VRF/Table Name. The Information field contains an ASCII 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. For consistency, it is RECOMMENDED that the VRF/Table Name always 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.

5.3. Peer Down Notification

Peer down notification MUST use reason code TBD3. 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 an ASCII 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 4.3 [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 10.0.0.0/8 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

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/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 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 filtered.

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 parameters to the BMP parameters name space [1].

Evens, et al. Expires December 9, 2019
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 Peer Flags

This document defines a new flag (Section 4.2) and proposes that peer flags are specific to the peer type:

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

8.3. Peer UP Information TLV

This document defines the following new BMP PEER UP informational message TLV types (Section 5.2.1):

- Type = 3: VRF/Table Name. The Information field contains an ASCII 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 = TBD3: Local system closed, TLV data follows.

9. References

9.1. Normative References


9.2. URIs

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

Acknowledgements

The authors would like to thank John Scudder and Jeff Haas for their valuable input.

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: serpis@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, WT 2132
NL

Email: paolo@ntt.net
Revision to Registration Procedures for Multiple BMP Registries
draft-ietf-grow-bmp-registries-change-01.txt

Abstract

This document updates RFC 7854, BGP Monitoring Protocol (BMP) by
making a change to the registration procedures for several
registries. Specifically, any BMP registry with a range of
32768-65530 designated "Specification Required" has that range re-
designated as "First Come First Served".

Status of This Memo

This Internet-Draft is submitted in full conformance with the
provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering
Task Force (IETF). Note that other groups may also distribute
working documents as Internet-Drafts. The list of current Internet-
Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months
and may be updated, replaced, or obsoleted by other documents at any
time. It is inappropriate to use Internet-Drafts as reference
material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 5, 2019.

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
1. Introduction

[RFC7854] creates a number of IANA registries that include a range of 32768-65530 designated "Specification Required". Each such registry also has a large range designated "Standards Action". Subsequent experience has shown two things. First, there is less difference between these two policies in practice than there is in theory (consider that [RFC8126] explains that for Specification Required, "Publication of an RFC is an ideal means of achieving this requirement"). Second, it’s desirable to have a very low bar to registration, to avoid the risk of conflicts introduced by use of unregistered code points (so-called "code point squatting").

Accordingly, this document revises the registration procedures, as given in Section 2.

2. IANA Considerations

IANA is requested to revise the following registries within the BMP group:

- BMP Statistics Types
- BMP Initiation Message TLVs
- BMP Termination Message TLVs
- BMP Termination Message Reason Codes
- BMP Peer Down Reason Codes
- BMP Route Mirroring TLVs
- BMP Route Mirroring Information Codes

For each of these registries, the ranges 32768-65530 whose registration procedures were "Specification Required" are revised to have the registration procedures "First Come First Served".
3. Security Considerations

This revision to registration procedures does not change the underlying security issues inherent in the existing [RFC7854].

4. Acknowledgements

Thanks to Jeff Haas for review and encouragement.

5. Normative References


Author’s Address

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

Email: jgs@juniper.net
RPKI Autonomous Systems Cones: A Profile To Define Sets of Autonomous Systems Numbers To Facilitate BGP Filtering
draft-ietf-grow-rpki-as-cones-01

Abstract

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

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 6, 2019.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.
1. Introduction

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

There is however no way for an operator to assert the routes for its customer networks, making it difficult to use the information carried by RPKI to create meaningful BGP-4 filters without relying on RPSL [RFC2622] as-sets.
This memo introduces a new attestation object, called an AS-Cone. An AS-Cone is a digitally signed object with the goal to enable operators to define a set of customers that can be found as "right adjacencies", or transit customer networks, facilitating the construction of prefix filters for a given ASN, thus making routing more secure.

2. Format of AS-Cone objects

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

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

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

2.1. Policy definition object

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

The default behaviour for a neighbour, if the relationship is not explicitly described in the policy, is to only accept the networks originated by the ASN. This means that a stub ASN neither has to set up any AS-Cone, description, nor policy.

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

2.1.1. Naming convention for Policy definition objects

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

2.1.2. ASN.1 format of a Policy Definition object
ASNPolicy DEFINITIONS ::= BEGIN
  Neighbours ::= SEQUENCE OF Neighbour
  Neighbour ::= SEQUENCE {
    ASN INTEGER (1..42949672965),
    AS Cone VisibleString
  }
  Version ::= INTEGER
  LastModified ::= GeneralizedTime
  Created ::= GeneralizedTime
END

ASN.1 format of a Policy definition object

2.1.3. Naming convention for neighbour relationships

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

ASX:ASY

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

2.2. AS-Cone definition object

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

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

2.2.1. Naming convention for AS-Cone objects

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

In order for AS-Cones to be unique in the global routing system,
their string name is preceded by the AS number of the ASN they are
part of, followed by ":". For example, AS-Cone "EuropeanCustomers"
for ASN 65530 is represented as "AS65530:EuropeanCustomers" when
referenced from a third party.
2.2.2. ASN.1 format of an AS-Cone

AS Cone DEFINITIONS ::= BEGIN
Entities ::= SEQUENCE OF Entity

Entity CHOICE
{ ASN INTEGER (1..4294967295),
  OtherAS Cone VisibleString
}

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

ASN.1 format of an AS-Cone

3. Validating an AS-Cone

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

In order to validate a full AS-Cone, a network operator MUST have access to the validated cache of an RPKI validator software containing all the Policy definition and AS-Cone objects. Validation occurs following the description in: [RFC6488].

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

1. For Every downstream ASN, the operator takes its policy definition file and collects a list of ASNs for the cone by looking at the following data, in exact order:
   1. A policy for the specific relationship, in the form of ASX:ASY, where ASX is the downstream ASN, and ASY is the ASN of the operator validating the AS-Cone;
   2. If there is no specific definition for the relationship, the ASX:Default policy;
If none of the two objects above exists, then the operator should only consider the ASN of its downstream to be added to the list.

2. These objects can either point to:
   1. An AS-Cone; or
   2. An ASN

3. If the definition points to an AS-Cone, the operator looks for the object referenced, which should be contained in the validated cache;

4. If the validated cache does not contain the referenced object, then the validation moves on to the next downstream ASN;

5. If the validated cache contains the referenced object, the validation process evaluates every entry in the AS-Cone. For each entry:
   1. If there is a reference to an ASN, then the operator adds the ASN to the list for the given AS-Cone;
   2. If there is a reference to another AS-Cone, the validating process should recursively process all the entries in that AS-Cone first, with the same principles contained in this list.

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

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

4. Recommendations for use of AS-Cones at Internet Exchange points

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

   In case of a peering session with a route server, the operator could publish a policy pointing to the ASN of the route server. A route
server operator, then, could build strict prefix filtering rules for all the participants, and offer it as a service to its members.

5. Publication of AS-Cones as IRR objects

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

6. Security Considerations

TBW

7. IANA Considerations

This memo includes no request to IANA.

8. Contributors

The following people contributed significantly to the content of the document: Greg Skinner.

9. Acknowledgments

The authors would like to thank ...
10.2. Informative References


Authors’ Addresses

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

Email: job@ntt.net

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

Email: mstucchi@ripe.net
Well-Known Community Policy Behavior
draft-ietf-grow-wkc-behavior-08

Abstract

Well-Known BGP Communities are manipulated differently across various current implementations; resulting in difficulties for operators. Network operators should deploy consistent community handling across their networks while taking the inconsistent behaviors from the various BGP implementations into consideration. This document recommends specific actions to limit future inconsistency, namely BGP implementors must not create further inconsistencies from this point forward.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."
1. Introduction

The BGP Communities Attribute was specified in [RFC1997] which introduced the concept of Well-Known Communities. In hindsight, [RFC1997] did not prescribe as fully as it should have how Well-Known Communities may be manipulated by policies applied by operators. Currently, implementations differ in this regard, and these differences can result in inconsistent behaviors that operators find difficult to identify and resolve.

This document describes the current behavioral differences in order to assist operators in generating consistent community-manipulation policies in a multi-vendor environment, and to prevent the introduction of additional divergence in implementations.
This document recommends specific actions to limit future inconsistency, namely BGP implementors MUST NOT create further inconsistencies from this point forward.

2. Manipulation of Communities by Policy

[RFC1997] says:

"A BGP speaker receiving a route with the COMMUNITIES path attribute may modify this attribute according to the local policy."

One basic operational need is to add or remove one or more communities to the set. The focus of this document is another common operational need, to replace all communities with a new set. To simplify this second case, most BGP policy implementations provide syntax to "set" community that operators use to mean "remove any/all communities present on the route, and apply this set of communities instead."

Some operators prefer to write explicit policy to delete unwanted communities rather than using "set;" i.e. using a "delete community *:*" and then "add community x:y ..." configuration statements in an attempt to replace all communities. The same community manipulation policy differences described in the following section exist in both "set" and "delete community *:*" syntax. For simplicity, the remainder of this document refers only to the "set" behaviors, which we refer to collectively as each implementation’s '"set" directive.'

3. Community Manipulation Policy Differences

Vendor implementations differ in the treatment of certain Well-Known communities when modified using the syntax to "set" the community. Some replace all communities including the Well-Known ones with the new set, while others replace all non-Well-Known Communities but do not modify any Well-Known Communities that are present.

These differences result in what would appear to be identical policy configurations having very different results on different platforms.

4. Documentation of Vendor Implementations

In this section we document the syntax and observed behavior of the "set" directive in several popular BGP implementations to illustrate the severity of the problem operators face.

In Juniper Networks’ Junos OS, "community set" removes all communities, Well-Known or otherwise.
In Cisco IOS XR, "set community" removes all communities except for the following:

<table>
<thead>
<tr>
<th>Numeric</th>
<th>Common Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:0</td>
<td>internet</td>
</tr>
<tr>
<td>65535:0</td>
<td>graceful-shutdown</td>
</tr>
<tr>
<td>65535:1</td>
<td>accept-own rfc7611</td>
</tr>
<tr>
<td>65535:65281</td>
<td>NO_EXPORT</td>
</tr>
<tr>
<td>65535:65282</td>
<td>NO_ADVERTISE</td>
</tr>
<tr>
<td>65535:65283</td>
<td>NO_EXPORT_SUBCONFED (or local-AS)</td>
</tr>
</tbody>
</table>

Communities not removed by Cisco IOS XR

Table 1

Cisco IOS XR does allow Well-Known communities to be removed only by explicitly enumerating one at a time, not in the aggregate; for example, "delete community accept-own". Operators are advised to consult Cisco IOS XR documentation and/or Cisco support for full details.

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

In Huawei’s VRP product, "community set" removes all communities, Well-Known or otherwise.

In OpenBGPD, "set community" does not remove any communities, Well-Known or otherwise.

Nokia’s SR OS has several directives that operate on communities. Its "set" directive is called using the "replace" keyword, replacing all communities, Well-Known or otherwise, with the specified communities.

4.1. Note on an Inconsistency

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

Cisco IOS XR’s set of Well-Known communities that "set community" will not overwrite diverges from the IANA’s list of Well-Known communities. Quite a few Well-Known communities from IANA’s list do not receive special treatment in Cisco IOS XR, and at least one community on Cisco IOS XR’s special treatment list, internet == 0:0,
is not formally a Well-Known Community as it is not in [IANA-WKC];
but taken from the Reserved range [0x00000000-0x0000FFFF].

This merely notes an inconsistency. It is not a plea to ‘protect’
the entire IANA list from "set community."

5. Note for Those Writing RFCs for New Community-Like Attributes

When establishing new [RFC1997]-like attributes (large communities,
wide communities, etc.), RFC authors should state explicitly how the
new attribute is to be handled.

6. Action Items

Network operators are encouraged to limit their use of the "set"
directive (within reason), to improve consistency across platforms.

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

Vendors MUST clearly document the behavior of "set" directive in
their implementations.

Vendors MUST ensure that their implementations’ "set" directive
treatment of any specific community does not change if/when that
community becomes a new Well-Known Community through future
standardization. For most implementations, this means that the "set"
directive MUST continue to remove the community; for those
implementations where the "set" directive removes no communities,
that behavior MUST continue.

Given the implementation inconsistencies described in this document,
network operators are urged never to rely on any implicit
understanding of a neighbor ASN’s BGP community handling. I.e.,
before announcing prefixes with NO_EXPORT or any other community to a
neighbor ASN, the operator should confirm with that neighbor how the
community will be treated.

7. Security Considerations

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

8. IANA Considerations

The IANA is requested to list this document as an additional
reference for the [IANA-WKC] registry.
9. Acknowledgments

The authors thank Martijn Schmidt, Qin Wu for the Huawei data point, Greg Hankins, Job Snijders, David Farmer, John Heasley, and Jakob Heitz.

10. Normative References


Authors’ Addresses

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

Email: jayb@att.com

Randy Bush
IIJ & Arrcus
5147 Crystal Springs
Bainbridge Island, WA  98110
US

Email: randy@psg.com
Abstract

This document describes mechanisms to limit the negative impact of route leaks [RFC7908] and/or resource exhaustion in BGP [RFC4271] implementations.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 9, 2019.

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
1. Introduction

This document describes mechanisms to reduce the negative impact of certain types of misconfigurations and/or resource exhaustions in BGP [RFC4271] operations. While [RFC4271] already described a method to tear down BGP sessions when certain thresholds are exceeded, some nuances in this specification were missing resulting in inconsistencies between BGP implementations. In addition to clarifying "inbound maximum prefix limits", this document also introduces a specification for "outbound maximum prefix limits".

2. Inbound Maximum Prefix Limits

An operator MAY configure a BGP speaker to terminate its BGP session with a neighbor when the number of address prefixes received from that neighbor exceeds a locally configured upper limit. The BGP speaker then MUST send the neighbor a NOTIFICATION message with the Error Code Cease and the Error Subcode "Threshold reached: Maximum Number of Prefixes Received", and MAY support other actions. Reporting when thresholds have been exceeded is an implementation specific consideration, but SHOULD include methods such as Syslog
Inbound Maximum Prefix Limits can be applied in two distinct places in the conceptual model: before or after the application of routing policy.

2.1. Type A: Pre-Policy Inbound Maximum Prefix Limits

The Adj-RIBs-In stores routing information learned from inbound UPDATE messages that were received from another BGP speaker. Section 3.2 [RFC4271]. The Type A pre-policy limit uses the number of NLRIs per Address Family Identifier (AFI) per Subsequent Address Family Identifier (SAFI) as input into its threshold comparisons. For example, when an operator configures the Type A pre-policy limit for IPv4 Unicast to be 50 on a given EBGP session, and the other BGP speaker announces its 51st IPv4 Unicast NLRI, the session MUST be terminated.

Type A pre-policy limits are particularly useful to help dampen the effects of full table route leaks and memory exhaustion when the implementation stores rejected routes.

2.2. Type B: Post-Policy Inbound Maximum Prefix Limits

RFC4271 describes a Policy Information Base (PIB) that contains local policies that can be applied to the information in the Routing Information Base (RIB). The Type B post-policy limit uses the number of NLRIs per Address Family Identifier (AFI) per Subsequent Address Family Identifier (SAFI), after application of the Import Policy as input into its threshold comparisons. For example, when an operator configures the Type B post-policy limit for IPv4 Unicast to be 50 on a given EBGP session, and the other BGP speaker announces a hundred IPv4 Unicast routes of which none are accepted as a result of the local import policy (and thus not considered for the Loc-RIB by the local BGP speaker), the session is not terminated.

Type B post-policy limits are useful to help prevent FIB exhaustion and prevent accidental BGP session teardown due to prefixes not accepted by policy anyway.

3. Outbound Maximum Prefix Limits

An operator MAY configure a BGP speaker to terminate its BGP session with a neighbor when the number of address prefixes to be advertised to that neighbor exceeds a locally configured upper limit. The BGP speaker then MUST send the neighbor a NOTIFICATION message with the Error Code Cease and the Error Subcode "Threshold reached: Maximum Number of Prefixes Send", and MAY support other actions. Reporting when thresholds have been exceeded is an implementation specific...
consideration, but SHOULD include methods such as Syslog [RFC5424]. By definition, Outbound Maximum Prefix Limits are Post-Policy.

The Adj-RIBs-Out stores information selected by the local BGP speaker for advertisement to its neighbors. The routing information stored in the Adj-RIBs-Out will be carried in the local BGP speaker’s UPDATE messages and advertised to its neighbors Section 3.2 [RFC4271]. The Outbound Maximum Prefix Limit uses the number of NLRIIs per Address Family Identifier (AFI) per Subsequent Address Family Identifier (SAFI), after application of the Export Policy, as input into its threshold comparisons. For example, when an operator configures the Outbound Maximum Prefix Limit for IPv4 Unicast to be 50 on a given EBGP session, and were about to announce its 51st IPv4 Unicast NLRI to the other BGP speaker as a result of the local export policy, the session MUST be terminated.

Outbound Maximum Prefix Limits are useful to help dampen the negative effects of a misconfiguration in local policy. In many cases, it would be more desirable to tear down a BGP session rather than causing or propagating a route leak.

4. Considerations for Operations with Multi-Protocol BGP

5. Considerations for soft thresholds

describe soft and hard limits (warning vs teardown)

6. Security Considerations

Maximum Prefix Limits are an essential tool for routing operations and SHOULD be used to increase stability.

7. IANA Considerations

This memo requests that IANA updates the name of subcode "Maximum Number of Prefixes Reached" to "Threshold exceeded: Maximum Number of Prefixes Received" in the "Cease NOTIFICATION message subcodes" registry under the "Border Gateway Protocol (BGP) Parameters" group.

This memo requests that IANA assigns a new subcode named "Threshold exceeded: Maximum Number of Prefixes Send" in the "Cease NOTIFICATION message subcodes" registry under the "Border Gateway Protocol (BGP) Parameters" group.
8. Acknowledgments

The authors would like to thank Saku Ytti and John Heasley (NTT Communications), Jeff Haas, Colby Barth and John Scudder (Juniper Networks), Martijn Schmidt (i3D.net), Teun Vink (BIT), Sabri Berisha (eBay), Martin Pels (Quanza), Steven Bakker (AMS-IX), Aftab Siddiqui (ISOC) and Yu Tianpeng for their support, insightful review, and comments.

9. Implementation status - RFC EDITOR: REMOVE BEFORE PUBLICATION

This section records the status of known implementations of the protocol defined by this specification at the time of posting of this Internet-Draft, and is based on a proposal described in RFC7942. The description of implementations in this section is intended to assist the IETF in its decision processes in progressing drafts to RFCs. Please note that the listing of any individual implementation here does not imply endorsement by the IETF. Furthermore, no effort has been spent to verify the information presented here that was supplied by IETF contributors. This is not intended as, and must not be construed to be, a catalog of available implementations or their features. Readers are advised to note that other implementations may exist.

The below table provides an overview (as of the moment of writing) of which vendors have produced implementation of inbound or outbound maximum prefix limits. Each table cell shows the applicable configuration keywords if the vendor implemented the feature.
### Table 1: Maximum prefix limits capabilities per implementation

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Type A Pre-Policy</th>
<th>Type B Post-Policy</th>
<th>Outbound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cisco IOS XR</td>
<td>maximum-prefix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cisco IOS XE</td>
<td>maximum-prefix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Juniper Junos OS</td>
<td>prefix-limit</td>
<td>accepted-prefix-limit, or prefix-limit combined with 'keep none'</td>
<td></td>
</tr>
<tr>
<td>Nokia SR OS</td>
<td>prefix-limit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIC.CZ BIRD</td>
<td>'import keep filtered' combined with 'receive limit'</td>
<td>'import limit' or 'receive limit'</td>
<td>export limit</td>
</tr>
<tr>
<td>OpenBSD OpenBGPD</td>
<td>max-prefix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arista EOS</td>
<td>maximum-routes</td>
<td>maximum-accepted-routes</td>
<td></td>
</tr>
<tr>
<td>Huawei VRPv5</td>
<td>peer route-limit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huawei VRPv8</td>
<td>peer route-limit</td>
<td>peer route-limit accept-prefix</td>
<td></td>
</tr>
</tbody>
</table>

First presented by Snijders at [RIPE77]

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Type A Pre-Policy</th>
<th>Type B Post-Policy</th>
<th>Outbound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cisco IOS XR</td>
<td>maximum-prefix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cisco IOS XE</td>
<td>maximum-prefix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Juniper Junos OS</td>
<td>prefix-limit</td>
<td>accepted-prefix-limit, or prefix-limit combined with 'keep none'</td>
<td></td>
</tr>
<tr>
<td>Nokia SR OS</td>
<td>prefix-limit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIC.CZ BIRD</td>
<td>'import keep filtered' combined with 'receive limit'</td>
<td>'import limit' or 'receive limit'</td>
<td>export limit</td>
</tr>
<tr>
<td>OpenBSD OpenBGPD</td>
<td>max-prefix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arista EOS</td>
<td>maximum-routes</td>
<td>maximum-accepted-routes</td>
<td></td>
</tr>
<tr>
<td>Huawei VRPv5</td>
<td>peer route-limit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huawei VRPv8</td>
<td>peer route-limit</td>
<td>peer route-limit accept-prefix</td>
<td></td>
</tr>
</tbody>
</table>

10. Appendix: Implementation Guidance

1) make it clear who does what: if A sends too many prefixes to B A should see "ABC" in log B should see "DEF" in log to make it clear which of the two parties does what 2) recommended by default automatically restart after between 15 and 30 minutes
11. References

11.1. Normative References


11.2. Informative References


Authors’ Addresses

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

Email: job@ntt.net
Melchior Aelmans
Juniper Networks
Boeing Avenue 240
Schiphol-Rijk 1119 PZ
The Netherlands

Email: maelmans@juniper.net
Abstract

RFC 7854, BMP, uses different message types for different purposes. Most of these are Type, Length, Value (TLV) structured. One message type, the Peer Up message, lacks a set of TLVs defined for its use, instead sharing a namespace with the Initiation message. Subsequent experience has shown that this namespace sharing was a mistake, as it hampers the extension of the protocol.

This document updates RFC 7854 by creating an independent namespace for the Peer Up message. The changes in this document are formal only, compliant implementations of RFC 7854 also comply with this specification.

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, 2019.

Copyright Notice

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents
## 1. Introduction

[RFC7854] defines a number of different BMP message types. With the exception of the Route Monitoring message type, these messages are TLV-structured. Most message types have distinct namespaces and IANA registries. However, the namespace of the Peer Up message overlaps that of the Initiation message. As the BMP protocol has been extended, this oversight has become problematic. In this document, we create a distinct namespace for the Peer Up message to eliminate this overlap, and create the corresponding missing registry.

The changes in this document are formal only, compliant implementations of [RFC7854] also comply with this specification.

### 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] when, and only when, they appear in all capitals, as shown here.

---

### Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>2</td>
</tr>
<tr>
<td>1.1. Requirements Language</td>
<td>2</td>
</tr>
<tr>
<td>2. String Definition</td>
<td>3</td>
</tr>
<tr>
<td>3. Changes to RFC 7854</td>
<td>3</td>
</tr>
<tr>
<td>3.1. Revision to Information TLV, Renamed as Initiation Information TLV</td>
<td>3</td>
</tr>
<tr>
<td>3.2. Revision to Peer Up Notification</td>
<td>3</td>
</tr>
<tr>
<td>3.3. Definition of Peer Up Information TLV</td>
<td>4</td>
</tr>
<tr>
<td>4. IANA Considerations</td>
<td>4</td>
</tr>
<tr>
<td>5. Security Considerations</td>
<td>5</td>
</tr>
<tr>
<td>6. Acknowledgements</td>
<td>5</td>
</tr>
<tr>
<td>7. Normative References</td>
<td>5</td>
</tr>
<tr>
<td>Author's Address</td>
<td>5</td>
</tr>
</tbody>
</table>

---

[Scudder]: Expires June 17, 2019 [Page 2]
2. String Definition

A string TLV is a free-form sequence of UTF-8 characters whose length is given by the TLV’s Length field. There is no requirement to terminate the string with a null (or any other particular) character -- the Length field gives its termination.

3. Changes to RFC 7854

We update [RFC7854] as follows:

- The "Information TLV" of section 4.4, that was shared between the Initiation and Peer Up message types, is renamed as the "Initiation Information TLV", and is only relevant to the Initiation message type.

- A "Peer Up Information TLV" is defined, and is relevant to the Peer Up message type.

- A "Peer Up TLVs" registry is created, seeded with the Peer Up Information TLV.

Other than as summarized above, and detailed below, there are no other changes.

3.1. Revision to Information TLV, Renamed as Initiation Information TLV

The Information TLV defined in section 4.4 of [RFC7854] is renamed "Initiation Information TLV". It is used only by the Initiation message, not by the Peer Up message.

The definition of Type = 0 is revised to be:

- Type = 0: String. The Information field contains a string (Section 2). The value is administratively assigned. If multiple strings are included, their ordering MUST be preserved when they are reported.

3.2. Revision to Peer Up Notification

The final paragraph of section 4.10 of [RFC7854] references the Information TLV (which is revised above (Section 3.1)). That paragraph is replaced by the following:

- Information: Information about the peer, using the Peer Up Information TLV format defined below (Section 3.3). The String type may be repeated. Inclusion of the Information field is
OPTIONAL. Its presence or absence can be inferred by inspection
of the Message Length in the common header.

3.3. Definition of Peer Up Information TLV

The Peer Up Information TLV is used by the Peer Up message.

```
0 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Information Type     |       Information Length      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                 Information (variable)                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- Information Type (2 bytes): Type of information provided. Defined
types are:
  * Type = 0: String. The Information field contains a string
    (Section 2). The value is administratively assigned. If
    multiple strings are included, their ordering MUST be preserved
    when they are reported.

- Information Length (2 bytes): The length of the following
  Information field, in bytes.

- Information (variable): Information about the monitored router,
  according to the type.

4. IANA Considerations

IANA is requested to create a registry within the BMP group, named
"BMP Peer Up Message TLVs", reference this document.

Registration procedures for this registry are:

```
<table>
<thead>
<tr>
<th>Range</th>
<th>Registration Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-32767</td>
<td>Standards Action</td>
</tr>
<tr>
<td>32768-65530</td>
<td>First Come, First Served</td>
</tr>
<tr>
<td>65531-65534</td>
<td>Experimental</td>
</tr>
<tr>
<td>65535</td>
<td>Reserved</td>
</tr>
</tbody>
</table>
```

Initial values for this registry are:
5. Security Considerations

This rearrangement of deck chairs does not change the underlying security issues inherent in the existing [RFC7854].

6. Acknowledgements

TBD

7. Normative References


Author’s Address

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

Email: jgs@juniper.net
Abstract

Many large-scale service provider networks use some form of scale-out architecture at peering sites. In such an architecture, each participating Autonomous System (AS) deploys multiple independent Autonomous System Border Routers (ASBRs) for peering, and Equal Cost Multi-Path (ECMP) load balancing is used between them. There are numerous benefits to this architecture, including but not limited to N+1 redundancy and the ability to flexibly increase capacity as needed. A cost of this architecture is an increase in the amount of state in both the control and data planes. This has negative consequences for network convergence time and scale.

In this document we describe how to mitigate these negative consequences through configuration of the routing protocols, both BGP and IGP, to utilize what we term the "Abstract Next-Hop" (ANH). Use of ANH allows us to both reduce the number of BGP paths in the control plane and enable rapid path invalidation (hence, network convergence and traffic restoration). We require no new protocol features to achieve these benefits.

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 August 10, 2019.
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 ...................................................... 3
  1.1. Scale-Out peering .......................................... 4
    1.1.1. Low latency ........................................ 4
    1.1.2. All equal cost paths utilization ..................... 4
    1.1.3. Summary ........................................... 5
  1.2. Common BGP Deployment Configurations ...................... 7
    1.2.1. IBGP with Next-Hop Unchanged ......................... 7
      1.2.1.1. Example ....................................... 7
    1.2.2. IBGP with Next-Hop-Self ............................ 8
  2. The BGP Abstract Next-Hop .................................... 8
  3. Use of Abstract Next-Hop in scale-out peering design ...... 9
    3.1. Egress ASBR-Peer AS Abstract Next Hop (AP-ANH) .......... 10
    3.2. The Site-Peer AS Abstract Next Hop (SP-ANH) ........... 11
    3.3. Assignment of Abstract Next Hops ....................... 14
      3.3.1. Native IP Networks ................................ 14
      3.3.2. MPLS ............................................ 14
        3.3.2.1. Identical BGP address space and paths received on all ASBRs .................................. 14
        3.3.2.2. Different address space sets or paths received on different ASBRs .............................. 14
      3.3.3. SPRING ........................................... 15
        3.3.3.1. Identical BGP address space and path received on all ASBRs .................................. 15
        3.3.3.2. Different address space sets or paths received on different ASBRs .............................. 15
  4. Worked Examples ............................................... 16
    4.1. Failure of a proper subset of EBGP sessions with a given peer AS on a single ASBR ............... 16
    4.2. Failure of a proper subset of EBGP sessions with a given peer AS on each ASBR of a given site .... 16
    4.3. Failure of all EBGP sessions with a given peer AS on
1. Introduction

Common to all large Internet networks are the requirements for large aggregate bandwidth and low latency. As network sizes and traffic volumes have increased, it has become common to use scale-out architectures to satisfy these requirements. Use of these techniques within individual networks is well-known. Here, we explore a scale-out architecture for interconnecting different Autonomous Systems (ASes).

Below, we show an example topology. Content is hosted within AS 2, consumers connect via the various ISP Metro ASes.

![Diagram of internet architecture](image)

Figure 1

ASes 1 and 2 are connected at multiple, geographically diverse, sites. Geographic diversity is required for reasons including resiliency, minimization of latency, and minimization of cost associated with long-distance data transmission.
1.1. Scale-Out peering

The same trends that have driven the use of scale-out architectures within ASes drive interest in using them at peering sites. In such an architecture, each AS at the peering site deploys multiple independent Autonomous System Border Routers (ASBRs). Benefits that can be realized include N+1 redundancy and the ability to flexibly increase capacity as needed. The ASBRs are often connected to the rest of their AS in a leaf-spine topology through core routers, and augmented with a per-site pair of BGP route reflectors (RRs). See for example SITE1 in Figure 2, below.

The fundamental requirements in this architecture are:

a. Keep traffic on a path that has low latency.

b. Utilize all peering links that offer low latency.

c. In the event of failure, minimize the time needed to restore service.

1.1.1. Low latency

BGP, the Border Gateway Protocol, does not directly carry delay information. We make the general assumption in this document that paths selected by the BGP best path algorithm [RFC4271] will provide lower latency than those not selected. This assumption is not guaranteed to be true, but lacking special arrangements between peering ASes, it is what the protocol is able to provide.

1.1.2. All equal cost paths utilization

In order to use all links between peering ASes that provide the same BGP path costs to the destination prefix, at a minimum BGP speakers need to be enabled for multi-path operation. Additionally, all AS ingress BGP speakers need to know at least all equal and best paths to the destination via multiple ASBRs. If a full IBGP mesh is used, this happens naturally. However, IBGP full meshes are uncommon in large networks and are even more impractical in scale-out architectures due to the high total number of ASBRs.

The well-known techniques to deal with full-mesh scale challenges – Route Reflection [RFC4456] and Confederations [RFC5065] – hide redundant paths, as they advertise only a single selected path to their clients. While this helps keep path and session scale manageable, it makes BGP multipath unusable. We overcome this by using BGP ADD-PATH [RFC7911] between the RR and its clients (or among sub-ASes).
1.1.3. Summary

In summary, for a scale-out peering architecture:

- BGP multipath needs to be enabled on all IBGP sessions inside the AS.
- BGP multipath needs to be enabled on all EBGP sessions of each ASBR.
- BGP ADD-PATH needs to be enabled on all IBGP sessions.
  
  * RRs need to be able to send multiple paths per prefix. The upper limit depends on:
    
    + The maximum number of ASBRs per site (say N).
    
    + Possibly also on the maximum number of EBGP sessions held by a single ASBR with single peer AS (say M), depending on BGP next-hop attribute (BGP-NH) configuration.
  
  * RR clients/ASBRs may need to be able to send multiple paths per prefix if BGP-NH configuration is "next hop unchanged". The upper limit depends on the maximum number of EBGP sessions held by a single ASBR with single peer AS (say M).

For further consideration the following network diagram will be used for reference:
Figure 2
1.2. Common BGP Deployment Configurations

1.2.1. IBGP with Next-Hop Unchanged

In one standard BGP configuration, an ASBR, when it advertises an externally learned prefix into IBGP, does not modify the BGP-NH. So, the BGP-NH is set to the IP address of an interface on the external peering router. The strength of this technique is the shorter time needed to restore connectivity with all equal cost multi-path (ECMP) in-use and on low latency paths. The drawback is extremely high BGP Routing Information Base (RIB) scale – proportional to the number of inter-AS links.

1.2.1.1. Example

Let’s assume that in the network of Figure 2, all PR2.x of AS2 advertise the same set of prefixes on all sessions to AS1.

If BR1.1-BR1.N and BR2.1-BR2.N’ each advertise only one path per prefix to their respective RRs, then as the result of ADD-PATH among RRs, BRs and CRs, at site 3 the BRs and CRs will learn N+N’ paths per prefix learned from AS2. This is sufficient to equally distribute load among all N ASBRs on site 1 (note the IGP cost between site 2 and site 3).

However, when interfaces over which all BR1.1-BR1.N learned their best path become unavailable (say interfaces to PR2.1 in all cases, as a result of the failure of PR2.1), the route to the BGP BGP-NH – that is, the IP address of the PR2.1 interface – is removed from the IGP. BGP speakers at other sites (BR3.x) will react by temporarily directing traffic to site 2 (BR2.1-BR2.N’). This switchover may happen in sub-second time, in a prefix-scale-independent manner, thanks to techniques commonly known as BGP PIC Edge [I-D.ietf-rtgwg-bgp-pic]. As a result, traffic is on a path other than the lowest cost path, as the connection from site 1 to AS2 is not entirely broken (links to PR2.2-PR2.M are operational).

Subsequently, all BR1.x will update their RRs with a new best path (say for PR2.2) for each prefix (for example, 100,000 of them), triggering global convergence. Such a convergence, for a large number of prefixes, may take many minutes.

In the above example, BRs, RRs, and possibly CRs keep N+N’ paths per prefix (N from site 1, and N’ from site 2). Provided N=N’=4, this makes 8 path per prefix.

The solution for sub-optimal routing right after the failure would be to enable each BR to advertise multiple paths to its RRs, and for
them in turn to propagate it to all other RR and hence BRs. So, each of BR1.x at site 1 will advertise M paths (from PR_2.1-PR_2.M), RR1.x will have N*M ECMP best paths and advertise them to other sites (site 3). As a result, BGP speakers at other sites (BR3.x at site 3) are provided with N*M paths per prefix from site 1 and N’*M’ from site 2. Therefore to achieve optimal routing immediately after failure, a considerably higher scale of BGP paths needs to be handled. If M=N=N’=M’=4 then for each prefix we have 16 best paths and 16 non-best, a total of 32. If AS2 advertises 100,000 prefixes, this becomes 3.2M paths.

Although this solution provides a mean of fast, prefix-scale-independent traffic switchover, it does it only if an ASBR external interface goes down, which triggers an IGP event. In case an EBGP session fails but the underlying interface remains up (misconfiguration, software defect, etc), recovery still requires per-prefix withdrawal/update that could take many minutes at high scale.

1.2.2. IBGP with Next-Hop-Self

The other common technique is to modify BGP-NH to "self" (a local IP address, typically a loopback) when the BR advertises an externally learned path into IBGP. This technique allows the reduction of the number of paths per prefix, while keeping optimal forwarding - least cost and ECMP - in case of failure discussed above (e.g. PR_2.1 node failure). Actually, because IP addresses of BGP-NH as seen by other BGP speakers do not change in response to external failure events, and are resolvable by the IGP, there is no need to reprogram the Forwarding Information Base (FIB) at all. Unfortunately, other failures - loss of all connectivity between a single BR (say BR1.1) and a peer AS (all PRs in AS2) would not be handled quickly. As the BGP-NH advertised by BR_1.1 is not changed and is reachable by the IGP, BGP speakers in AS1 (BRs, CRs) will keep BR_1.1 as a feasible exit point until they receive BGP withdraws on a prefix-by-prefix basis. This is a global convergence process that at high scale can take minutes, during which time packets may be discarded or loop.

2. The BGP Abstract Next-Hop

The Abstract Next Hop (ANH) concept presented below does not require any changes to the BGP protocol itself. It is architectural solution to network configuration, that uses existing protocols’ capabilities while achieving higher scale and faster routing convergence when scale-out peering sites exist.

When a BGP speaker advertises a path to its IBGP peer, it modifies the Protocol Next-Hop to be the ANH value. The ANH is just an IP
address that identifies the BGP session or a set of BGP sessions. The set of BGP sessions is defined by the operator in local configuration, according to network design needs. For example, an ANH might identify:

- a set of BGP sessions with the same peer AS and handled by a given single ASBR
- a set of BGP sessions with same the peer AS and handled by one or more ASBRs at a given site
- a set of BGP sessions with any upstream provider AS
- a set of BGP sessions with a given peer device and handled by one or more of ASBRs of the local AS

A host route to the ANH is installed in the relevant RIB and redistributed into the IGP. BGP maintains the ANH host route based on the state of the associated group of BGP sessions:

- As soon as all BGP sessions in the set go down, the ANH route is removed.
- When at least one BGP session in of the set comes up, the ANH route is created only after initial route convergence is complete for the peer (End-of-RIB (EoR) [RFC4724] is received).

Taken together, these procedures ensure that as soon as the final session in the set goes down, ingress routers will see the associated ANH withdrawn from the IGP. Since the ANH is used to resolve the associated BGP next hops, the ingress routers are triggered to converge to send traffic to their alternate (new best) route. They also ensure that as soon as one session in the set comes up and is synchronized (that is, the EoR is received), ingress routers will see the ANH advertised in the IGP and will be able to reconverge to use routes that are associated with that next hop.

The ANH can be any IP address that the router is eligible to advertise according to the local network’s IP address management scheme. More details are given in Section 3.3.

3. Use of Abstract Next-Hop in scale-out peering design

In traditional configurations as described in Section 1.2 the meaning of the BGP-NH is either:

- An egress interface in the case of next-hop-unchanged configuration, or
An egress ASBR in the case of next-hop-self configuration.

The meaning of Abstract Next Hop is more context-dependent. This document describes network configurations when the BGP-NH identifies:

a. An (egress ASBR, peer AS) pair. The ANH should be advertised into the IGP if, and only if, the given egress ASBR has at least one EBGP session in the ESTABLISHED state with the given peer AS, and the EoR marker has been received on that session. We call this the ASBR-Peer AS Abstract Next Hop (AP-ANH).

b. An (egress site in local AS, peer AS) pair, where a "site" may include multiple ASBRs. The ANH should be advertised into the IGP if, and only if, at least one ASBR of the given site has at least one EBGP session in the ESTABLISHED state with the given peer AS, and the EoR marker has been received on this session. We call this the Site-Peer AS Abstract Next Hop (SP-ANH).

Note that reachability of the ANH address in the IGP depends on EBGP session state and not inter-AS interface state, although of course, interface state may impact session state. How the IP route to the ANH address is instantiated on an ASBR and inserted into the IGP on particular device is a matter of local implementation.

3.1. Egress ASBR-Peer AS Abstract Next Hop (AP-ANH)

The AP-ANH is unique to an ASBR and its peer AS. For example, in the network of Figure 2, BR_1.1 would have two AP-ANH assigned – one for its peering with AS2 and the other for AS3. Similarly, BR_1.2 would have two AP-ANH, one per peer AS, with values different from the AP-ANH of BR_1.1, and so on. All AP-ANH are exported into the IGP by their ASBRs. Each ASBR advertises only one path per prefix to its RR, with the BGP-NH set to the appropriate AP-ANH. The RR will propagate it through the entire AS by means of IBGP ADD-PATH. In consequence, the number of paths learned per prefix is equal to number of ASBRs servicing a given peer AS. In the network as of Figure 2, for AS2 prefixes, this would be N+N' (from site_1 + from site_2) paths per prefix. This sets the scale requirements of this solution to be on par with Next-Hop-Self (Section 1.2.2). However, thanks to the properties of ANH, more failures are covered by prefix-independent techniques, as withdrawal of the ANH from the IGP makes the BGP-NH unresolvable.

Provided that all ASBRs in a given site (site_1 in Figure 2) receive the same routing information from their peer AS (AS2), in non-faulty conditions, one could consider setting the ANH value on all ASBRs the same. However, failure(s) can create situations when multiple ASBRs will have a session in ESTABLISHED state with a given peer AS, but
some prefixes would be learned from EBGP only on a subset of these ASBRs. To prevent problems from arising in this situation, the per-ASBR AP-ANH needs to be advertised into the IGP and ASBRs need to set it as the BGP-NH when advertising routes to the site’s Route Reflectors. However, for IBGP path advertisement being propagated beyond the site (into the RR mesh), the BGP-NH may be replaced by another ANH value, the Site-Peer AS ANH.

3.2. The Site-Peer AS Abstract Next Hop (SP-ANH)

The AP-ANH works on an ASBR level. From a given local AS perspective, the number of ANH is proportional to the number of pairs of ASBRs and ASes each of them peers with. With hundreds of peer ASes, tens of sites and ~10 ASBRs per site, the number of AP-ANH may scale into the thousands. At the same time, it may not be necessary or even desirable for every BGP speaker in the network to have visibility to every path down to individual egress ASBR granularity. With symmetrical multiplane backbone and/or leaf-spine designs, it is sufficient that BGP speakers on other sites have information that a given site (site1 in Figure 2) has at least one ASBR with an ESTABLISHED session to the peer AS (AS2). For example, in the network of Figure 2, even if BR3.1 has only one path with its BGP-NH equal to the ANH of BR1.1, BR3.1 resolves the BGP-NH in the IGP and spreads traffic among all CRs on site 3. Thus, traffic will be delivered to CR1.x at site 1. As long as CR1.x has visibility to all paths, traffic will be distributed equally to all site 1 ASBRs.

At the same time, when multiple paths are available on BGP speakers, every change is propagated, with consequent transmission and processing costs on all BGP speakers across the network. This will be true even if the route change doesn’t impact the forwarding plane. For example, in the network of Figure 2, even if BR3.1 has N paths with BGP-NHs set to the ANHs of BR1.1 through BR1.N, BR3.1 will resolve those BGP-NHs in the IGP and spread traffic among all CRs of site 3. When one of the egress ASBRs (say BR1.2) loses its connectivity to the peer AS, the affected BGP routes (those with BGP-NH equal to AP-ANH of BR1.2) are withdrawn from all BGP speakers (e.g. BR3.1) of the network. All BGP speakers perform path selection and possibly update their forwarding data structures. Since the actual forwarding paths do not change, all this work represents unnecessary churn.

To avoid the above drawbacks, the RR of a given site (site1 in Figure 2), when re-advertising a BGP path learned from its ASBR client, modifies the BGP-NH to another abstract value - the Site-Peer AS Abstract NH (SP-ANH). This value is unique per (site, peer AS) pair, and is shared by all RRs of a given site. With this modification, it is sufficient that inter-site IBGP sessions carry
only one path per prefix (no ADD-PATH needed). Consequently, BGP RIB scale is reduced significantly. This frees up memory, reduces the amount of data RRs need to exchange, and mitigates churn. The BGP speakers in other sites of AS 1 need to resolve SP-ANH in order to build their local FIBs. Therefore SP-ANH have to be present in the IGP - some router(s) in the local site (RR, ASBR or CR) need to inject it into the IGP. While the selection of role that is responsible of SP-ANH injection is discussed below, in any case, the SP-ANH should be reachable in the IGP if, and only if, at least one of AP-ANH (for the same peer AS and ASBR belonging to given site) is reachable. Figure 3 illustrates routing information flow in a network such as that of Figure 2:
3.3. Assignment of Abstract Next Hops

In the following subsections we provide more details of how abstract next hops can be injected in several different common network architectures.

3.3.1. Native IP Networks

In this network every router, including core routers, has full BGP routing information and forwards each packet based on destination IP lookup. Provided that all routers at an egress site receive multiple paths with BGP-NH set to AP-ANH (and not SP-ANH), it is a matter of the operator’s decision which node – RR, ASBR or CR – will inject the SP-ANH route into the IGP. One may argue that injection of SP-ANH by ASBRs may be simpler, as it will be done by the same procedure and policy as injection of AP-ANH. Others may prefer injection at RR, as it limits the number of configuration touch-points.

3.3.2. MPLS

3.3.2.1. Identical BGP address space and paths received on all ASBRs

In the MPLS network, since traffic is carried over LSP tunnels, the SP-ANH needs to be injected into the IGP by a node that has the ability to perform an IP lookup. This eliminates the RR, and possibly CRs (in "BGP-free core" architectures). Instead, all ASBRs are used to insert SP-ANH addresses into the IGP. In case of LDP-based networks, this is sufficient. The CR will create an ECMP forwarding structure for labels of SP-ANH FEC coming from other sites. In RSVP-TE based networks, ECMP needs to happen on the ingress LSR and therefore, every BGP speaker needs to establish an LSP to every ASBR, and the SP-ANH address needs to be part of the FEC for its respective LSP. If SP-ANH is used as an RSVP (signaling) destination, some other means (such as affinity groups) needs to be used to ensure the desired 1:1 LSP to egress ASBR mapping.

3.3.2.2. Different address space sets or paths received on different ASBRs

In the case when the set of prefixes received from a given peer AS by one ASBR is different from the set received by another one, a combination of SP-ANH and MPLS-based load balancing on a CR may lead to a situation where an IP packet will be directed to an ASBR that lacks external routing information and hence can’t forward traffic directly out of the AS. Similarly, if path attributes for a given prefix received by one ASBR are different from those received by another, again packets can be directed to the "wrong" ASBR. In this case the ASBR would use the IBGP route it learned from another ASBR.
of the same site (via RR, with AP-ANH) and forward traffic over an LSP to the "correct" ASBR. This extra hop constitutes a sub-optimal traffic path through the network.

For example in the network of Figure 2, let’s assume that prefix P2 is advertised to BR1.2-BR1.N by AS2 but not to BR1.1. BR3.1 has a BGP best route to P2 with its BGP-NH set to the SP-ANH of (site1, AS2). It resolves it by ECMP over N MPLS LSPs, terminating on BR1.1-BR1.N. So, some packets are forwarded by BR3.1 over an LSP via CR1.x and terminated on BR1.1. BR1.1 has no external route to P2, but it has (N-1) IBGP routes to P2 w/ BGP-NHs equal to the AP-ANHs of BR1.2-BR1.N. Therefore BR1.1 performs an IP lookup and forwards this packet over LSPs via CR1.x and terminated on BR1.2-BR1.N. Traffic is U-turned on BR1.1 and traverses CRs at site 1 twice.

Such asymmetry may be considered acceptable by the provider, as long as it’s a transient condition. However, in the general case such a situation could be persistent, as the result of intentional configuration on the peer AS’s ASBRs. Therefore the better solution would be to insert the SP-ANH into the IGP on CRs. In this case, CRs need to perform forwarding based on destination IP lookup. Therefore CRs would have to be able to learn and handle large IP routing and forwarding tables - at least all prefixes learned from peer ASes by the local ASBRs.

3.3.3. SPRING

3.3.3.1. Identical BGP address space and path received on all ASBRs

For SPRING based networks, we can take advantage of the unique capability of Anycast-SID [RFC8402]. The ASBRs of a single site allocate an Anycast-SID for each SP-ANH address. This SID can be used as the only SID by an ingress BGP speaker or, if a TE routed path is desired, depending on TE constraints, the TE controller can provision a SPRING path with the Anycast-SID at the end, instructing the CR to perform load balancing among connected ASBRs.

3.3.3.2. Different address space sets or paths received on different ASBRs

Similarly to a classic MPLS environment, such a situation may lead to suboptimal routing (redirecting from one ASBR to another), or may require the CR (instead of ASBR) to insert the SP-ANH into the IGP and generate a PREFIX-SID (or Anycast-SID if there is more than one CR) for it.
4. Worked Examples

Below we illustrate the operation of the proposal by working through its operation in the context of several different types of failures. Here, we assume that each ASBR in a given site of the local AS (site 1 of AS1 in Figure 2), that has an EBGP session with the given peer AS (AS2 in Figure 2), receives from its peer routers (PR2.x) routes to exactly same address space on each session.

4.1. Failure of a proper subset of EBGP sessions with a given peer AS on a single ASBR

- The impacted ASBR keeps advertising the AP-ANH into the IGP, as at least one session to the peer AS remains in the ESTABLISHED state.

- The impacted ASBR may send UPDATEs to RRs, however the BGP-NH remains the same and equal to the pre-failure AP-ANH.

- The RRs may send UPDATEs to their clients (CRs, BRs) and to RRs in other sites, however the BGP-NH remains the same as its pre-failure value: AP-ANH and SP-ANH respectively.

- As BGP-NH do not change, there are no changes in forwarding data structures (FIB) on any BGP speaker across the network, except possibly the ASBR that holds the impacted session.

4.2. Failure of a proper subset of EBGP sessions with a given peer AS on each ASBR of a given site

- The impacted ASBRs keep advertising the AP-ANH into the IGP, as at least one session to the peer AS remains in the ESTABLISHED state on each ASBR.

- The impacted ASBRs may send UPDATEs to RRs, however the BGP-NH remains the same and equal to the pre-failure AP-ANH.

- The RRs may send UPDATEs to their clients (CRs, BRs) and to RRs in other sites, however the BGP-NH remains the same and equal to its pre-failure value: AP-ANH and SP-ANH respectively.

- As BGP-NH do not change, there are no changes in forwarding data structures (FIB) on any BGP speaker across the network, except possibly the ASBRs that hold the impacted sessions.
4.3. Failure of all EBGP sessions with a given peer AS on single ASBR; Failure of a single ASBR

- The impacted ASBR stops advertising the AP-ANH into the IGP, as it has lost all sessions with given peer AS.

- The SP-ANH is kept reachable in the IGP.

- All other BGP speakers at the impacted site invalidate all paths with BGP-NH equal to the AP-ANH. This may trigger prefix-independent FIB data-structure patching/temporary fixing for sub-second traffic restoration.

- The impacted ASBR sends WITHDRAWs to its RRs.

- Each RR:
  * Sends WITHDRAWs to its clients at the local site (CRs, BRs) for paths from the impacted ASBR. As these sessions support ADD-PATH, paths from other ASBRs will remain. Other BGP speakers at this site have to modify their FIBs.
  * May send UPDATEs to RRs in other sites, however the BGP-NH remains the same, equal to the pre-failure SP-ANH. As the BGP-NH does not change, there are no changes in forwarding data structure (FIB) on any of BGP speakers across network, except those at the impacted site.

- Routing churn is mitigated in many cases to a single peering site, and does not propagate across the network. FIB changes are limited to a single peering site, and do not propagate across the network.

4.4. All EBGP sessions with a given peer AS on all ASBRs

- Each ASBR stops advertising its AP-ANH into the IGP, as it has lost all sessions with the given peer AS.

- The SP-ANH is no longer reachable in the IGP, as none of AP-ANH are reachable.

- All other BGP speakers across the network invalidate all paths with a BGP-NH equal to the removed AP-ANH or SP-ANH. This may trigger prefix-independent FIB data-structure patching/temporary fixing for sub-second traffic restoration.

- Each impacted ASBR sends WITHDRAWs to its RRs.
The RRs send WITHDRAWs to their clients at the local site (CRs, BRs) and RRs in other sites for paths from the impacted ASBRs. As these sessions support ADD-PATH, paths from ASBRs at other sites will remain. The BGP speakers across the network may need to modify their FIBs.

5. Acknowledgements

Valuable comments and suggestions on solution covered by this document was provided by Mannan Venkatesan, John Scudder and Ron Bonica. Special thanks to John Scudder, who also helped with editorial changes.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

Since this is a deployment architecture and not a protocol modification, it doesn’t introduce any new issues to the BGP protocol itself. General BGP security considerations are discussed in [RFC4271] and [RFC4272], BGP deployment best practices are documented in [RFC7454], and nothing in this proposal impedes their use. Many of the practices recommended in that document are self-evidently still applicable, for example the use of cryptographic session protection methods such as TCP MD5 [RFC2385] or the TCP Authentication Option [RFC5925], and the Generalized TTL Security Mechanism [RFC5082]. Since we propose a novel use of IP addresses to assign ANHs, it’s worth considering if anything new is required to protect them. We conclude there isn’t, they fall into the existing category of "Prefixes Belonging to the Local AS" discussed in section 6.1.4 of [RFC7454].

8. Informative References

[I-D.ietf-rtgwg-bgp-pic]


Authors’ Addresses

Rafal Jan Szarecki (editor)
Juniper Networks Inc.
1133 Innovation Way
Sunnyvale, CA  94089
US
Phone: +1(408)680-9604
Email: rafal@juniper.net

Kaliraj Vairavakkalai
Juniper Networks Inc.
1133 Innovation Way
Sunnyvale, CA  94089
US
Phone: +1(408)936-8872
Email: kaliraj@juniper.net

Natrajan Venkataraman
Juniper Networks Inc.
1133 Innovation Way
Sunnyvale, CA  94089
US
Phone: +1(408)936-6597
Email: natv@juniper.net
BGP Route Policy and Attribute Trace Using BMP
draft-xu-grow-bmp-route-policy-attr-trace-00

Abstract

The generation of BGP adj-rib-in, local-rib or adj-rib-out comes from BGP protocol communication, and route policy processing. BGP Monitoring Protocol (BMP) provides the monitoring of BGP adj-rib-in [RFC7854], BGP local-rib [I-D.ietf-grow-bmp-local-rib] and BGP adj-rib-out [I-D.ietf-grow-bmp-adj-rib-out]. However, there lacks monitoring of how BGP routes are transformed from adj-rib-in into local-rib and then adj-rib-out (i.e., the BGP route policy processing procedures). This document describes a method of using BMP to trace the change of BGP routes in correlation with responsible route policies.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 10, 2019.
1. Introduction

The typical processing procedure after receiving a BGP Update Message at a routing device is as follows: 1. Adding the pre-policy routes into the pre-policy adj-rib-in (if any); 2. Filtering the pre-policy routes through inbound route policies; 3. Selecting the BGP best routes from the post-policy policies; 4. Adding the selected routes into the BGP local-rib; 5-a. Adding the BGP best routes from local-rib to the core routing table manager for selection; 5-b. Filtering the routes from BGP local-rib through outbound route policies w.r.t. per peer or peer groups; 6. Sending the BGP adj-rib-out to the target peer or peer groups. Details may vary by vendors. The BGP Monitoring Protocol (BMP) can be utilized to monitor BGP routes in forms of adj-rib-in, local-rib and adj-rib-out. However, the complete procedure from inbound to outbound policy processing, including other policies, e.g., route redistribution, route selection...
and so on, is currently unobserved. For example, there are 10 policy items (or nodes) configured under one outbound route policy per a specific peer. By collecting the local-rib and adj-rib-out through BMP, the operator finds that the outbound policy didn’t work as expected. However, it’s hard to distinguish which one of the 10 policy items/nodes is responsible for the failure.

1.1. BGP Route Policy and Attribute Trace Overview

This document describes a method that records and reports how each policy item/node processes the routes (e.g., changes the route attribute). Each policy item/node processing is called an event thereafter in this document. Compared with conventional BGP rib entry, which consists of prefix/mask, route attributes, e.g., next hop, MED, local preference, AS path, and so on, the event record discussed in this document includes extra information, such as event index, timestamp, policy information, and so on. For example, if a route is processed by 5 policy items/nodes, there can be 5 event records for the same prefix/mask. Each event is numbered in order of time (e.g., the time of policy execution). The policy information includes the policy name and item/node ID/name so that the server/controller can map to the exact policy either directly from the device or from the configurations collected at the server side.

This document defines a new BMP message type to carry the recorded policy and route data. More detailed message format is defined in Section 2. The message is called the BMP Route Policy and Attribute Trace Message thereafter in this document.

1.2. Use cases

There are cases that a new policy is configured incorrectly, e.g., setting an incorrect community value, or policy placed in incorrect order among other policies. These may result in incorrect route attribute modification, best route selection mistake, or route distribution mistake. With the correlated record of policy and route, the server/controller is able to identify the unexpected route change and its responsible policy. Considering the fact that the BGP route policy impacts not only the route processing within the individual device but also the route distribution to its peers, the route trace data of a single device is always analyzed in correlation with such data collected from its peer devices.

Apart from the policy validation application, the route trace data can also be analyzed to discover the route propagation path within the network. With the route’s inbound and outbound event records collecte from each related device, the server is able to find the propagation path hop by hop. The identified path is helpful for
operators to better understand its network, and thus benefitting both
network troubleshooting and network planning.

2. Extension of BMP for Route Policy and Attribute Trace

2.1. Common Header

This document defines a new BMP message type to carry the Route
Policy and Attribute Trace data.

- Type = TBD: Route Policy and Attribute Trace Message

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

2.2. Per Peer Header

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

2.3. Route Policy and Attribute Trace Message

The Route Policy and Attribute Trace Message format is defined as
follows:

```
+---------------------------------------------------------------+
|                          Prefix length                        |
+---------------------------------------------------------------+
|                             Prefix                            |
+---------------------------------------------------------------+
|                       Route Distinguisher                     |
+---------------------------------------------------------------+
|                          Previous Hop                        |
+---------------------------------------------------------------+
|                           Event count                         |
+---------------------------------------------------------------+
|                       Total event length                      |
+---------------------------------------------------------------+
|          Single event length (1st event)                    |
+---------------------------------------------------------------+
|                           Event index                         |
+---------------------------------------------------------------+
|                        Timestamp(seconds)                     |
+---------------------------------------------------------------+
|                      Timestamp(microseconds)                   |
+---------------------------------------------------------------+
|           Policy ID            |     Policy distinguisher       |
+---------------------------------------------------------------+
```
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Peer ID</td>
<td></td>
</tr>
<tr>
<td>Peer AS</td>
<td></td>
</tr>
<tr>
<td>Peer VRF/Table name</td>
<td></td>
</tr>
<tr>
<td>Peer AFI</td>
<td>Peer SAFI</td>
</tr>
<tr>
<td>Total attribute length</td>
<td></td>
</tr>
<tr>
<td>Attribute TLVs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Single event length (Last event)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attribute TLVs</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Route Policy and Attribute Trace Message format

- Prefix Length (1 Byte): indicates the length of the prefix.
- Prefix (Variable): indicates the monitored prefix, with the length defined by Prefix Length field.
- Route Distinguisher (8 Bytes): If the route is an IPv4 route, this field is zero-filled. If the peer is a VPNv4 route, it is set to the route distinguisher (RD) of the route.
- Previous Hop (4 Bytes): indicates the BGP peer ID where this route is learnt from. If the route is locally generated, then field is set to the local BGP router ID (global or VRF specific).
- Event Count (1 Byte): indicates the total number of policy processing event recorded in this message.
- Total event length (1 Byte): indicates the total length of the following fields including all events, where the total number is indicated by the Event Count field.
Single event length (1 Byte): indicates the total length of a single policy process event, including the following fields that belong to this event.

Event index (1 Byte): indicates the sequence number of this event, starting from 1 and increases by 1 for each event recorded in order.

Timestamp (4 Bytes): indicates the time when the policy of this event starts execution, expressed in seconds and microseconds since midnight (zero hour), January 1, 1970 (UTC).

Policy ID (Variable): indicates the ID of the route policy of this event, which is user specific or vendor specific. It consists of the Route Policy Name and the Route Policy Item/Node ID. The Policy name and Item/Node ID is in the format of ASCII string, the length of both fields are indicated by the Policy length and Item/Node length fields, respectively.

Policy Distinguisher (4 Bits): indicates the category of the policy. Currently 3 policy categories are defined: "0000" indicating the inbound policy, "0001" indicating the outbound policy, "0010" indicating the redistribution policy. More categories to be defined.

Peer ID (4 Bytes): indicates the BGP Peer ID where this policy is configured under. This field is used in combination with the Policy Direction field. If the Policy Direction field is set to "0000", meaning inbound policy, then this field is set to the BGP Peer ID where the route is received from; if the Policy Direction field is set to "0001", meaning outbound policy, then this field is set to the BGP Peer ID where the route is distributed to; If the Policy Direction field is set to "0010", meaning redistribution policy, then this field is set to the local BGP router ID (global or VRF specific).

Peer AS (4 Bytes): indicates the AS number of the BGP Peer that defined the Peer ID field.
- VRF/Table name (Variable): indicates the VRF or table name of this route in the format of ASCII string. The string size MUST be within the range of 1 to 255 bytes. The VRF/Table name information varies for the same route under different policy processing event. For example, an IPv4 route is received from a CE router at the PE router through iGBP, an RD is attached to this IPv4 route (under VRF name A) and making it a VPNv4 route, and then this VPNv4 route (under the Global routing table) is distributed to the RR. During this process, the VRF/Table name information changes from VRF A to the Global routing Table name at the inbound and outbound policy process.

- AFI/SAFI (2 Bytes): indicates the AFI/SAFI of the route. The AFI/SAFI information varies for the same route under different policy processing event. For example, an IPv4 route is received from a CE router at the PE router through iGBP, an RD is attached to this IPv4 route and making it a VPNv4 route, and then this VPNv4 route is distributed to the RR. During this process, the AFI information changes from IPv4 to VPNv4 at the inbound and outbound policy process.

- Total attribute length (2 Bytes): indicates the total length of the following route attribute TLVs.

- Attribute TLVs: include attributes that are currently carried in BGP Update messages (e.g., Community, Ext-community, Next Hop, AS path, MED...) and those that are not (to be defined).

3. Implementation Example
We take the network shown in Figure 2 as an example to show how to use Route Policy and Attribute Trace Messages to recover the footprint of the route propagation. Notice that only basic events required for footprint recovery are listed here.

Suppose a prefix 10.1.1.1/24 is sent from both CE2 and CE3 to PE1 through eBGP peering, PE1 processes the two Update messages with inbound policies. Such procedure is recorded as two events, namely Event 1 and Event 2. Then PE1 selects the route from CE2 as the best route, add it to VRF 1, and then distribute the VPNv4 route to RR. The distribution procedure is recorded by PE1 as Event 3. As an example, the Route Policy and Attribute Trace Message of Event 1, 2, 3 is listed as follows. Only fields related to footprint recovery are listed in the message shown below. Specifically, the Previous Hop information is carried in Event 3 when outbounding the route, indicating that the outbounded route is learnt from CE2. The same
prefix is sent from CE1 to PE2, added to VRF 1 and then distributed to RR in the form of VPNv4 route. Two events, Event 4 (inbound) and Event 5 (outbound) are recorded by PE2. Now for RR, prefix 10.1.1.1/24 is received from both PE1 and PE2 in the form of VPNv4 route. RR selects the route from CE2 as the best route, and distribute it to PE3. Three events, Event 6 (PE2 inbound), Event 7 (PE1 inbound), Event 8 (PE3 outbound) are recorded in this case. PE3 receives the VPNv4 route from RR, adds it to VRF 1 and then distribute the IPv4 route to CE4 and CE5, respectively. Here, three events are recorded, Event 9 (RR inbound), Event 10 (CE4 outbound) and Event 11 (CE5 outbound).

+---------------------------------------------------------------+  
|                         RD: 65000:10                          |  
|---------------------------------------------------------------+  
|                      Prefix:  10.1.1.1/24                     |  
|---------------------------------------------------------------+  
|                            Event 1                           |  
|---------------------------------------------------------------+  
|                           Timestamp 1                        |  
|---------------------------------+-----------------------------+  
|    Policy ID: WC1, node 101    |          Inbound policy    |  
|---------------------------------+-----------------------------+  
|                           Peer ID: CE1                        |  
|-----------------------------+----------------------------++  
|                           Peer AS: AS1                        |  
|---------------------------------------------------------------+  
|                       VRF/Table name: VRF 1                   |  
|---------------------------------------------------------------+  
|                            AFI: IPv4                          |  
|---------------------------------------------------------------+  
|                         Previous Hop: CE1                      |  
|+---------------------------------------------------------------+  
|                            Event 2                           |  
|---------------------------------------------------------------+  
|                           Timestamp 2                        |  
|---------------------------------+-----------------------------+  
|    Policy ID: WC1, node 102    |          Inbound policy    |  
|---------------------------------+-----------------------------+  
|                           Peer ID: CE2                        |  
|-----------------------------+----------------------------++  
|                           Peer AS: AS2                        |  
|---------------------------------------------------------------+  
|                       VRF/Table name: VRF 1                   |  
|---------------------------------------------------------------+  
|                            AFI: IPv4                          |  
|---------------------------------------------------------------+  
|                         Previous Hop: CE2                     |
The BMP server can use the collected events to recover the route footprint. The key information required from recovery is the Timestamp of each event, and the Previous Hop of the route. The Timestamp allows the server to identify the order of each event, while the Previous Hop information, combined with the outbound peer information, allows the server to recover the route propagation hop by hop.

4. Implementation Considerations

Considering the data amount of monitoring the route and policy trace of all routes from all BMP clients, the Route Policy and Attribute Trace monitoring MAY be triggered by user at any user-specific time, and MAY be applied to user-specific routes as well as all routes. Successive recored events from one device MAY be encapsulated in one Route Policy and Attribute Trace Message or multiple Route Policy and Attribute Trace Messages per the user configuration.

5. Acknowledgements

TBD.

6. IANA Considerations

TBD.
7. Security Considerations

TBD.

8. Normative References

[I-D.ietf-grow-bmp-adj-rib-out]

[I-D.ietf-grow-bmp-local-rib]


Authors' Addresses

Feng Xu
Tencent
Guangzhou
China

Email: oliverxu@tencent.com
Yunan Gu
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: guyunan@huawei.com

Shunwan Zhuang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: zhuangshunwan@huawei.com

Zhenbin Li
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com