

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 25, 2013

Pierre Francois
Institute IMDEA Networks
Bruno Decraene
France Telecom
Cristel Pelsser
Internet Initiative Japan
Keyur Patel
Clarence Filsfils
Cisco Systems
October 22, 2012

Graceful BGP session shutdown
draft-ietf-grow-bgp-gshut-04

Abstract

This draft describes operational procedures aimed at reducing the amount of traffic lost during planned maintenances of routers or links, involving the shutdown of BGP peering sessions.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 25, 2013.

Copyright Notice

Copyright (c) 2012 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 (<http://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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	4
2. Terminology	4
3. Packet loss upon manual eBGP session shutdown	5
4. Practices to avoid packet losses	5
4.1. Improving availability of alternate paths	5
4.2. Make before break convergence: g-shut	6
4.2.1. eBGP g-shut	6
4.2.2. iBGP g-shut	7
4.2.3. Router g-shut	7
5. Forwarding modes and transient forwarding loops during convergence	8
6. Link Up cases	8
6.1. Unreachability local to the ASBR	8
6.2. iBGP convergence	9
7. IANA assigned g-shut BGP community	9
8. Security Considerations	10
9. Acknowledgments	10
10. References	10
Appendix A. Alternative techniques with limited applicability . .	11
A.1. Multi Exit Discriminator tweaking	11
A.2. IGP distance Poisoning	11
Authors' Addresses	11

1. Introduction

Routing changes in BGP can be caused by planned, maintenance operations. This document discusses operational procedures to be applied in order to reduce or eliminate losses of packets during the maintenance. These losses come from the transient lack of reachability during the BGP convergence following the shutdown of an eBGP peering session between two Autonomous System Border Routers (ASBR).

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

The procedures described in this document can be applied to reduce or avoid packet loss for outbound and inbound traffic flows initially forwarded along the peering link to be shut down. These procedures trigger, in both involved ASes, rerouting to the alternate path, while allowing routers to keep using old paths until alternate ones are learned, installed in the RIB and in the FIB. This ensures that routers always have a valid route available during the convergence process.

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

Still, it explains why reserving a community value for the purpose of BGP session graceful shutdown would reduce the management overhead bound with the solution. It would also allow vendors to provide an automatic graceful shutdown mechanism that does not require any router reconfiguration at maintenance time.

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

2. Terminology

g-shut initiator: a router on which the session shutdown is performed for the maintenance.

g-shut neighbor: a router that peers with the g-shut initiator via (one of) the session(s) to be shut down.

Initiator AS: the Autonomous System of the g-shut initiator.

Neighbor AS: the Autonomous System of the g-shut neighbor.

Loss of Connectivity (LoC: the state when a router has no path towards an affected prefix.

3. Packet loss upon manual eBGP session shutdown

Packets can be lost during a manual shutdown of an eBGP session for two reasons.

First, routers involved in the convergence process can transiently lack of paths towards an affected prefix, and drop traffic destined to this prefix. This is because alternate paths can be hidden by nodes of an AS. This happens when the paths are not selected as best by the ASBR that receive them on an eBGP session, or by Route Reflectors that do not propagate them further in the iBGP topology because they do not select them as best.

Second, within the AS, the FIB of routers can be transiently inconsistent during the BGP convergence and packets towards affected prefixes can loop and be dropped. Note that these loops only happen when ASBR-to-ASBR encapsulation is not used within the AS.

This document only addresses the first reason.

4. Practices to avoid packet losses

This section describes means for an ISP to reduce the transient loss of packets upon a manual shutdown of a BGP session.

4.1. Improving availability of alternate paths

All solutions that increase the availability of alternate BGP paths at routers performing packet lookups in BGP tables such as [BestExternal] and [AddPath] help in reducing the LoC bound with manual shutdown of eBGP sessions.

One of such solutions increasing diversity in such a way that, at any single step of the convergence process following the eBGP session shutdown, a BGP router does not receive a message withdrawing the only path it currently knows for a given NLRI, allows for a simplified g-shut procedure.

Note that the LoC for the inbound traffic of the maintained router, induced by a lack of alternate path propagation within the iBGP topology of a neighboring AS is not under the control of the operator performing the maintenance. The part of the procedure aimed at avoiding LoC for incoming paths can thus be applied even if no LoC

are expected for the outgoing paths.

4.2. Make before break convergence: g-shut

This section describes configurations and actions to be performed to perform a graceful shutdown procedure for eBGP peering links.

The goal of this procedure is to let the paths being shutdown visible, but with a lower LOCAL_PREF value, while alternate paths spread through the iBGP topology. Instead of withdrawing the path, routers of an AS will keep on using it until they become aware of alternate paths.

4.2.1. eBGP g-shut

4.2.1.1. Pre-configuration

On each ASBR supporting the g-shut procedure, an outbound BGP route policy is applied on all iBGP sessions of the ASBR, that:

- o matches the g-shut community
- o sets the LOCAL_PREF attribute of the paths tagged with the g-shut community to a low value
- o removes the g-shut community from the paths.
- o optionally, adds an AS specific g-shut community on these paths to indicate that these are to be withdrawn soon. If some ingress ASBRs reset the LOCAL_PREF attribute, this AS specific g-shut community will be used to override other LOCAL_PREF preference changes.

Note that in the case where an AS is aggregating multiple routes under a covering prefix, it is recommended to filter out the g-shut community from the resulting aggregate BGP route. By doing so, the setting of the g-shut community on one of the aggregated routes will not let the entire aggregate inherit the community. Not doing so would let the entire aggregate undergo the g-shut behavior.

4.2.1.2. Operations at maintenance time

On the g-shut initiator, upon maintenance time, it is required to:

- o apply an outbound BGP route policy on the maintained eBGP session to tag the paths propagated over the session with the g-shut community. This will trigger the BGP implementation to re-advertise all active routes previously advertised, and tag them with the g-shut community.
- o apply an inbound BGP route policy on the maintained eBGP session to tag the paths received over the session with the g-shut community.

- o wait for convergence to happen.
- o perform a BGP session shutdown.

4.2.1.3. BGP implementation support for G-Shut

A BGP router implementation MAY provide features aimed at automating the application of the graceful shutdown procedures described above.

Upon a session shutdown specified as graceful by the operator, a BGP implementation supporting a g-shut feature SHOULD:

1. On the eBGP side, update all the paths propagated over the corresponding eBGP session, tagging the GSHUT community to them. Any subsequent update sent to the session being gracefully shut down would be tagged with the GSHUT community.
2. On the iBGP side, lower the LOCAL_PREF value of the paths received over the eBGP session being shut down, upon their propagation over iBGP sessions. Optionally, also tag these paths with an AS specific g-shut community. Note that alternatively, the LOCAL_PREF of the paths received over the eBGP session can be lowered on the g-shut initiator itself, instead of only when propagating over its iBGP sessions.
3. Optionally shut down the session after a configured time.
4. Prevent the GSHUT community from being inherited by a path that would aggregate some paths tagged with the GSHUT community. This behavior avoids the GSHUT procedure to be applied to the aggregate upon the graceful shutdown of one of its covered prefixes.

A BGP implementation supporting a g-shut feature SHOULD also automatically install the BGP policies that are supposed to be configured, as described in Section 4.2.1.1 for sessions over which g-shut is to be supported.

4.2.2. iBGP g-shut

If the iBGP topology is viable after the maintenance of the session, i.e, if all BGP speakers of the AS have an iBGP signaling path for all prefixes advertised on this g-shut iBGP session, then the shutdown of an iBGP session does not lead to transient unreachability.

4.2.3. Router g-shut

In the case of a shutdown of a router, a reconfiguration of the outbound BGP route policies of the g-shut initiator SHOULD be performed to set a low LOCAL_PREF value for the paths originated by the g-shut initiator (e.g, BGP aggregates redistributed from other

protocols, including static routes).

This behavior is equivalent to the recommended behavior for paths "redistributed" from eBGP sessions to iBGP sessions in the case of the shutdown of an ASBR.

5. Forwarding modes and transient forwarding loops during convergence

The g-shut procedure or the solutions improving the availability of alternate paths, do not change the fact that BGP convergence and the subsequent FIB updates are runned independently on each router of the ASes. If the AS applying the solution does not rely on encapsulation to forward packets from the Ingress Border Router to the Egress Border Router, then transient forwarding loops and consequent packet losses can occur during the convergence process. If zero LoC is required, encapsulation is required between ASBRs of the AS.

6. Link Up cases

We identify two potential causes for transient packet losses upon an eBGP link up event. The first one is local to the g-no-shut initiator, the second one is due to the BGP convergence following the injection of new best paths within the iBGP topology.

6.1. Unreachability local to the ASBR

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

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

If the link up event corresponds to an eBGP session that is being manually brought up, over an already up multi-access link, then the operator can ping third party nexthops that are expected to be used before actually bringing the session up, or ping directed broadcast the subnet IP address of the link. By proceeding like this, the MAC addresses associated with these third party nexthops will be resolved by the g-no-shut initiator.

6.2. iBGP convergence

Corner cases leading to LoC can occur during an eBGP link up event.

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

Let's consider 3 route reflectors RR1, RR2, RR3. There is a full mesh of iBGP session between them.

1. RR1 is initially advertising the current best path to the members of its iBGP RR full-mesh. It propagated that path within its RR full-mesh. RR2 knows only that path towards the prefix.
2. RR3 receives a new best path originated by the "g-no-shut" initiator, being one of its RR clients. RR3 selects it as best, and propagates an UPDATE within its RR full-mesh, i.e., to RR1 and RR2.
3. RR1 receives that path, reruns its decision process, and picks this new path as best. As a result, RR1 withdraws its previously announced best-path on the iBGP sessions of its RR full-mesh.
4. If, for any reason, RR3 processes the withdraw generated in step 3, before processing the update generated in step 2, RR3 transiently suffers from unreachability for the affected prefix.

The use of [BestExternal] among the RR of the iBGP full-mesh can solve these corner cases by ensuring that within an AS, the advertisement of a new route is not translated into the withdraw of a former route.

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

7. IANA assigned g-shut BGP community

Applying the g-shut procedure is rendered much easier with the use of a single g-shut community value which could be used on all eBGP sessions, for both inbound and outbound signaling. The community value 0xFFFF0000 has been assigned by IANA for this purpose.

For Internet routes, a non transitive extended community will be reserved from the pool defined in [EXT_POOL]. Using such a community

type allows for not leaking graceful signaling out of the AS boundaries, without the need to explicitly configure filters to strip the community off upon path propagation.

8. Security Considerations

By providing the g-shut service to a neighboring AS, an ISP provides means to this neighbor to lower the LOCAL_PREF value assigned to the paths received from this neighbor.

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

If this behavior is not tolerated by the ISP, it SHOULD monitor the use of the g-shut community by this neighbor.

ASes using the regular (transitive) g-shut community SHOULD remove the community from neighboring ASes that do not support the g-shut procedure. Doing so prevents malignant remote ASes from using the community through intermediate ASes that do not support the feature, in order to perform inbound traffic engineering. ASes using the non-transitive extended community do not need to do this as the community is non transitive and hence cannot be used by remote ASes.

9. Acknowledgments

The authors wish to thank Olivier Bonaventure and Pradosh Mohapatra for their useful comments on this work.

10. References

- [AddPath] D. Walton, E. Chen, A. Retana, and J. Scudder,
"Advertisement of Multiple Paths in BGP",
draft-ietf-idr-add-paths-07.txt (work in progress).
- [BestExternal]
Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H.
Gredler, "Advertisement of the best-external route to
IBGP", draft-ietf-idr-best-external-05.txt.
- [REQS] Decraene, B., Francois, P., Pelsser, C., Ahmad, Z.,
Armengol, A., and T. Takeda, "Requirements for the
graceful shutdown of BGP sessions", RFC 6198.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[EXT_POOL] Decraene, B. and P. Francois, "Assigned BGP extended communities", draft-ietf-idr-reserved-extended-communities-03.

[BGPWKC] "<http://www.iana.org/assignments/bgp-well-known-communities>".

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

Appendix A. Alternative techniques with limited applicability

A few alternative techniques have been considered to provide g-shut capabilities but have been rejected due to their limited applicability. This section describe them for possible reference.

A.1. Multi Exit Discriminator tweaking

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

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

A.2. IGP distance Poisoning

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

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

Authors' Addresses

Pierre Francois
Institute IMDEA Networks
Avda. del Mar Mediterraneo, 22
Leganese 28918
ES

Email: pierre.francois@imdea.org

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
92794 Issy Moulineaux cedex 9
FR

Email: bruno.decraene@orange.com

Cristel Pelsser
Internet Initiative Japan
Jinbocho Mitsui Bldg.
1-105 Kanda Jinbo-cho
Tokyo 101-0051
JP

Email: cristel@iij.ad.jp

Keyur Patel
Cisco Systems
170 West Tasman Dr
San Jose, CA 95134
US

Email: keyupate@cisco.com

Clarence Filsfils
Cisco Systems
De kleetlaan 6a
Diegem 1831
BE

Email: cfilsfil@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2013

J. Scudder
Juniper Networks
R. Fernando
Cisco Systems
S. Stuart
Google
October 22, 2012

BGP Monitoring Protocol
draft-ietf-grow-bmp-07

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on April 25, 2013.

Copyright Notice

Copyright (c) 2012 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 (<http://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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	4
1.1. Requirements Language	4
2. Definitions	4
3. Overview of BMP Operation	4
3.1. BMP Messages	4
3.2. Connection Establishment and Termination	5
3.3. Lifecycle of a BMP Session	6
4. BMP Message Format	7
4.1. Common Header	7
4.2. Per-Peer Header	7
4.3. Initiation Message	9
4.4. Termination Message	10
4.5. Route Monitoring	11
4.6. Stats Reports	11
4.7. Peer Down Notification	13
4.8. Peer Up Notification	14
5. Route Monitoring	15
6. Stat Reports	17
7. Other Considerations	17
8. Using BMP	17
9. IANA Considerations	17
9.1. BMP Message Types	18
9.2. BMP Statistics Types	18
9.3. BMP Initiation Message TLVs	18
9.4. BMP Termination Message TLVs	19
9.5. BMP Termination Message Reason Codes	19
10. Security Considerations	19
11. Acknowledgements	20
12. References	20
12.1. Normative References	20
12.2. Informative References	20
Appendix A. Changes Between BMP Versions 1 and 2	21
Appendix B. Changes Between BMP Versions 2 and 3	21
Authors' Addresses	21

1. Introduction

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

The BMP protocol provides access to the Adj-RIB-In of a peer on an ongoing basis and a periodic dump of certain statistics that the monitoring station can use for further analysis. From a high level, BMP can be thought of as the result of multiplexing together the messages received on the various monitored BGP sessions.

1.1. Requirements Language

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

2. Definitions

- o 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.
- o Post-Policy Adj-RIB-In: The result of applying inbound policy to an Adj-RIB-In, but prior to the application of route selection to form the Loc-RIB.

3. Overview of BMP Operation

3.1. BMP Messages

The following are the messages provided by BMP.

- o Route Monitoring (RM): An initial dump of all routes received from a peer as well as an ongoing mechanism that sends the incremental routes advertised and withdrawn by a peer to the monitoring station.
- o Peer Down Notification (PD): A message sent to indicate that a peering session has gone down with information indicating the reason for the session disconnect.

- o Stats Reports (SR): An ongoing dump of statistics that can be used by the monitoring station as a high level indication of the activity going on in the router.
- o Peer Up Notification (PU): A message sent to indicate that a peering session has come up. The message includes information regarding the data exchanged between the peers in their OPEN messages as well as information about the peering TCP session itself. In addition to being sent whenever a peer transitions to ESTABLISHED state, a Peer Up Notification is sent for each peer that is in ESTABLISHED state when the BMP session itself comes up.
- o Initiation: A means for the monitored router to inform the monitoring station of its vendor, software version, and so on.
- o Termination: A means for the monitored router to inform the monitoring station of why it is closing a BMP session.

3.2. Connection Establishment and Termination

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

The router may be monitored by one or more monitoring stations. With respect to each (router, monitoring station) pair, one party is active with respect to TCP session establishment, and the other party is passive. Which party is active and which is passive is controlled by configuration.

The passive party is configured to listen on a particular TCP port and the active party is configured to establish a connection to that port. If the active party is unable to connect to the passive party, it periodically retries the connection. Retries MUST be subject to some variety of backoff. Exponential backoff with a default initial backoff of 30 seconds and a maximum of 720 seconds is suggested.

The router MAY restrict the set of IP addresses from which it will accept connections. It SHOULD restrict the number of simultaneous connections it will permit from a given IP address. The default value for this restriction SHOULD be 1, though an implementation MAY permit this restriction to be disabled in configuration. The router MUST also restrict the rate at which sessions may be established. A suggested default is an establishment rate of 2 sessions per minute.

A router (or management station) MAY implement logic to detect redundant connections, as might occur if both parties are configured to be active, and MAY elect to terminate redundant connections. A Termination reason code is defined for this purpose.

Once a connection is established, the router sends messages over it. There is no initialization or handshaking phase, messages are simply sent as soon as the connection is established.

If the monitoring station intends to restart BMP processing, it simply drops the connection, optionally with a Termination message.

3.3. Lifecycle of a BMP Session

A router is configured to speak BMP with one more monitoring stations. It MAY be configured to send monitoring information for only a subset of its BGP peers. Otherwise, all BGP peers are assumed to be monitored.

A BMP session begins when the active party (either router or management station, as determined by configuration) successfully opens a TCP session (the "BMP session"). Once the session is up, the router begins to send BMP messages. It MUST begin by sending an Initiation message. It subsequently sends a Peer Up message over the BMP session for each of its monitored BGP peers which are in Established state. It follows by sending the contents of its Adj-RIBs-In (pre-policy, post-policy or both, see Section 5) encapsulated in Route Monitoring messages. Once it has sent all the routes for a given peer, it sends an End-of-RIB message for that peer; when End-of-RIB has been sent for each monitored peer, the initial table dump has completed. (A monitoring station that wishes only to gather a table dump could close the connection once it has gathered an End-of-RIB or Peer Down message corresponding to each Peer Up message.)

Following the initial table dump, the router sends incremental updates encapsulated in Route Monitoring messages. It MAY periodically send Stats Reports or even new Initiation messages, according to configuration. If any new monitored BGP peers become Established, corresponding Peer Up messages are sent. If any BGP peers for which Peer Up messages were sent transition out of the Established state, corresponding Peer Down messages are sent.

A BMP session ends when the TCP session that carries it is closed for any reason. The router MAY send a Termination message prior to closing the session.

4. BMP Message Format

4.1. Common Header

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

```

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 1 2 3 4 5 6 7 8
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   Version   |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Message Length                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   Msg. Type   |
+-----+

```

- o Version (1 byte): Indicates the BMP version. This is set to '3' for all messages defined in this specification. Version 0 is reserved and MUST NOT be sent.
- o Message Length (4 bytes): Length of the message in bytes (including headers, data and encapsulated messages, if any).
- o Message Type (1 byte): This identifies the type of the BMP message. A BMP implementation MUST ignore unrecognized message types upon receipt.
 - * Type = 0: Route Monitoring
 - * Type = 1: Statistics Report
 - * Type = 2: Peer Down Notification
 - * Type = 3: Peer Up Notification
 - * Type = 4: Initiation Message
 - * Type = 5: Termination Message

4.2. Per-Peer Header

The per-peer header follows the common header for most BMP messages. The rest of the data in a BMP message is dependent on the "Message Type" field in the common header.

```

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 1 2 3 4 5 6 7 8
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Peer Type      |  Peer Flags      |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Peer Distinguisher (present based on peer type)  |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Peer Address (16 bytes)  |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Peer AS  |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Peer BGP ID  |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Timestamp (seconds)  |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|  Timestamp (microseconds)  |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

- o Peer Type (1 byte): These bits identify the type of the peer. Currently only two types of peers are identified,
 - * Peer Type = 0: Global Instance Peer
 - * Peer Type = 1: L3 VPN Instance Peer
- o Peer Flags (1 byte): These flags provide more information about the peer. The flags are defined as follows.

```

0 1 2 3 4 5 6 7 8
+---+---+---+---+---+---+---+---+
|V|L| Reserved  |
+---+---+---+---+---+---+---+---+

```

- * The V flag indicates the the Peer address is an IPv6 address. For IPv4 peers this is set to 0.
 - * The L flag, if set to 1, indicates that the message reflects the post-policy Adj-RIB-In (i.e., it reflects the application of inbound policy). It is set to 0 if the message reflects the pre-policy Adj-RIB-In. See Section 5 for further detail.
 - * The remaining bits are reserved for future use.
- o Peer Distinguisher (8 bytes): Routers today can have multiple instances (example L3VPNs). This field is present to distinguish peers that belong to one address domain from the other.

If the peer is a "Global Instance Peer", this field is zero

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

- o Peer Address: The remote IP address associated with the TCP session over which the encapsulated PDU was received. It is 4 bytes long if an IPv4 address is carried in this field (with most significant bytes zero filled) and 16 bytes long if an IPv6 address is carried in this field.
- o Peer AS: The Autonomous System number of the peer from which the encapsulated PDU was received. If a 16 bit AS number is stored in this field [RFC4893], it should be padded with zeroes in the most significant bits.
- o Peer BGP ID: The BGP Identifier of the peer from which the encapsulated PDU was received.
- o Timestamp: The time when the encapsulated routes were received (one may also think of this as the time when they were installed in the Adj-RIB-In), expressed in seconds and microseconds since midnight (zero hour), January 1, 1970 (UTC). If zero, the time is unavailable. Precision of the timestamp is implementation-dependent.

4.3. Initiation Message

The initiation message provides a means for the monitored router to inform the monitoring station of its vendor, software version, and so on. An initiation message **MUST** be sent as the first message after the TCP session comes up. An initiation message **MAY** be sent at any point thereafter, if warranted by a change on the monitored router.

The initiation message consists of the common BMP header followed by two or more TLVs containing information about the monitored router, as follows:

```

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

```

- o Information Type (2 bytes): Type of information provided. Defined types are:

- * Type = 0: String. The Information field contains a free-form UTF-8 string whose length is given by the "Information Length" field. The value is administratively assigned. Inclusion of this TLV is optional. Multiple String TLVs MAY be included in the message.
 - * Type = 1: sysDescr. The Information field contains an ASCII string whose value MUST be set to be equal to the value of the sysDescr MIB-II [RFC1213] object. Inclusion of this TLV is mandatory.
 - * Type = 2: sysName. The Information field contains a ASCII string whose value MUST be set to be equal to the value of the sysName MIB-II [RFC1213] object. Inclusion of this TLV is mandatory.
- o Information Length (2 bytes): The length of the following Information field, in bytes.
 - o Information (variable): Information about the monitored router, according to the type.

4.4. Termination Message

The termination message provides a way for a monitored router to indicate why it is terminating a session. Although use of this message is RECOMMENDED, a monitoring station must always be prepared for the session to terminate with no message. Once the router has sent a termination message, it MUST close the TCP session without sending any further messages. Likewise, the monitoring station MUST close the TCP session after receiving a termination message.

The termination message consists of the common BMP header followed by one or more TLVs containing information about the reason for the termination, as follows:

```

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

```

- o Information Type (2 bytes): Type of information provided. Defined types are:

- * Type = 0: String. The Information field contains a free-form UTF-8 string whose length is given by the "Information Length" field. Inclusion of this TLV is optional. It MAY be used to provide further detail for any of the defined reasons. Multiple String TLVs MAY be included in the message.
- * Type = 1: Reason. The Information field contains a two-byte code indicating the reason the connection was terminated. Some reasons may have further TLVs associated with them. Inclusion of this TLV is not optional. Defined reasons are:
 - + Reason = 0: Session administratively closed.
 - + Reason = 1: Unspecified reason.
 - + Reason = 2: Out of resources. The router has exhausted resources available for the BMP session.
 - + Reason = 3: Redundant connection. The router has determined that this connection is redundant with another one.
- o Information Length (2 bytes): The length of the following Information field, in bytes.
- o Information (variable): Information about the monitored router, according to the type.

4.5. Route Monitoring

Route Monitoring messages are used for initial synchronization of ADJ-RIBs-In. They are also used for ongoing monitoring of received advertisements and withdraws. This is discussed in more detail in Section 5.

Following the common BMP header and per-peer header is a BGP Update PDU.

4.6. Stats Reports

These messages contain information that could be used by the monitoring station to observe interesting events that occur on the router.

Transmission of SR messages could be timer triggered or event driven (for example, when a significant event occurs or a threshold is reached). This specification does not impose any timing restrictions on when and on what event these reports have to be transmitted. It is left to the implementation to determine transmission timings --

however, configuration control should be provided of the timer and/or threshold values. This document only specifies the form and content of SR messages.

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

```

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 1 2 3 4 5 6 7 8
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Stats Count                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Each counter is encoded as follows,

```

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 1 2 3 4 5 6 7 8
+-----+-----+-----+-----+-----+-----+-----+-----+
|          Stat Type          |          Stat Len          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|          Stat Data          |
~                               ~
+-----+-----+-----+-----+-----+-----+-----+-----+

```

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

This specification defines the following statistics. A BMP implementation MUST ignore unrecognized stat types on receipt, and likewise MUST ignore unexpected data in the Stat Data field.

Stats are either counters or gauges, defined as follows after the examples of [RFC1155] Section 3.2.3.3 and [RFC2856] Section 4 respectively:

32-bit Counter: A non-negative integer which monotonically increases until it reaches a maximum value, when it wraps around and starts increasing again from zero. It has a maximum value of $2^{32}-1$ (4294967295 decimal).

64-bit Gauge: non-negative integer, which may increase or decrease, but shall never exceed a maximum value, nor fall below a minimum value. The maximum value can not be greater than $2^{64}-1$ (18446744073709551615 decimal), and the minimum value can not be smaller than 0. The value has its maximum value whenever the

information being modeled is greater than or equal to its maximum value, and has its minimum value whenever the information being modeled is smaller than or equal to its minimum value. If the information being modeled subsequently decreases below (increases above) the maximum (minimum) value, the 64-bit Gauge also decreases (increases).

- o Stat Type = 0: (32-bit Counter) Number of prefixes rejected by inbound policy.
- o Stat Type = 1: (32-bit Counter) Number of (known) duplicate prefix advertisements.
- o Stat Type = 2: (32-bit Counter) Number of (known) duplicate withdraws.
- o Stat Type = 3: (32-bit Counter) Number of updates invalidated due to CLUSTER_LIST loop.
- o Stat Type = 4: (32-bit Counter) Number of updates invalidated due to AS_PATH loop.
- o Stat Type = 5: (32-bit Counter) Number of updates invalidated due to ORIGINATOR_ID.
- o Stat Type = 6: (32-bit Counter) Number of updates invalidated due to AS_CONFED loop.
- o Stat Type = 7: (64-bit Gauge) Number of routes in Adj-RIBs-In.
- o Stat Type = 8: (64-bit Gauge) Number of routes in Loc-RIB.

Note that although the current specification only specifies 4-byte counters and 8-byte gauges as "Stat Data", this does not preclude future versions from incorporating more complex TLV-type "Stat Data" (for example, one which can carry prefix specific data). SR messages are optional. However if an SR message is transmitted, at least one statistic MUST be carried in it.

4.7. Peer Down Notification

This message is used to indicate that a peering session was terminated.

```

0 1 2 3 4 5 6 7 8
+---+---+---+---+---+
|   Reason   | 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8
+---+---+---+---+---+
|           Data (present if Reason = 1, 2 or 3)           |
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
+---+---+---+---+---+

```

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

- o Reason 1: The local system closed the session. Following the Reason is a BGP PDU containing a BGP NOTIFICATION message that would have been sent to the peer.
- o Reason 2: The local system closed the session. No notification message was sent. Following the reason code is a two-byte field containing the code corresponding to the FSM Event which caused the system to close the session (see Section 8.1 of [RFC4271]). Two bytes both set to zero are used to indicate that no relevant Event code is defined.
- o Reason 3: The remote system closed the session with a notification message. Following the Reason is a BGP PDU containing the BGP NOTIFICATION message as received from the peer.
- o Reason 4: The remote system closed the session without a notification message.

A Peer Down message implicitly withdraws all routes that had been associated with the peer in question. A BMP implementation MAY omit sending explicit withdraws for such routes.

4.8. Peer Up Notification

The Peer Up message is used to indicate that a peering session has come up (i.e., has transitioned into ESTABLISHED state). Following the common BMP header and per-peer header is the following:

```

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 1 2 3 4 5 6 7 8
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Local Address (16 bytes)                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Local Port      |      Remote Port      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Sent OPEN Message                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Received OPEN Message                                |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

- o Local Address: The local IP address associated with the peering TCP session. It is 4 bytes long if an IPv4 address is carried in this field, as determined by the V flag (with most significant bytes zero filled) and 16 bytes long if an IPv6 address is carried in this field.
- o Local Port: The local port number associated with the peering TCP session.
- o Remote Port: The remote port number associated with the peering TCP session. (Note that the remote address can be found in the Peer Address field of the fixed header.)
- o Sent OPEN Message: The full OPEN message transmitted by the monitored router to its peer.
- o Received OPEN Message: The full OPEN message received by the monitored router from its peer.

5. Route Monitoring

After the BMP session is up, Route Monitoring messages are used to provide a snapshot of the Adj-RIB-In of each monitored peer. This is done by sending all routes stored in the Adj-RIB-In of those peers using standard BGP Update messages, encapsulated in Route Monitoring messages. There is no requirement on the ordering of messages in the peer dumps. When the initial dump is completed for a given peer, this MUST be indicated by sending an End-of-RIB marker for that peer (as specified in Section 2 of [RFC4724], plus the BMP encapsulation header). See also Section 8.

A BMP speaker may send pre-policy routes, post-policy routes, or both. The selection may be due to implementation constraints (it is

possible that a BGP implementation may not store, for example, routes which have been filtered out by policy). Pre-policy routes MUST have their L flag clear in the BMP header (see Section 4), post-policy routes MUST have their L flag set. When an implementation chooses to send both pre- and post-policy routes, it is effectively multiplexing two update streams onto the BMP session. The streams are distinguished by their L flags.

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

AS Numbers in the BMP UPDATE message MUST be sent as 4-octet quantities, as described in [RFC4893]. This affects the AS_PATH and AGGREGATOR path attributes. AS4_PATH or AS4_AGGREGATOR path attributes MUST NOT be sent in a BMP UPDATE message, as it makes no sense to do so.

Ongoing monitoring is accomplished by propagating route changes in BGP Update PDUs and forwarding those PDUs to the monitoring station, again using RM messages. When a change occurs to a route, such as an attribute change, the router must update the monitor with the new attribute. As discussed above, it MAY generate either an update with the L flag clear, with it set, or two updates, one with the L flag clear and the other with the L flag set. When a route is withdrawn by a peer, a corresponding withdraw is sent to the monitor. The withdraw MUST have its L flag set to correspond to that of any previous announcement; if the route in question was previously announced with L flag both clear and set, the withdraw MUST similarly be sent twice, with L flag clear and set. Multiple changed routes MAY be grouped into a single BGP UPDATE PDU when feasible, exactly as in the standard BGP protocol.

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

6. Stat Reports

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

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

7. Other Considerations

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

8. Using BMP

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

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

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

9. IANA Considerations

IANA is requested to create the following registries.

9.1. BMP Message Types

This document defines five message types for transferring BGP messages between cooperating systems (Section 4):

- o Type 0: Route Monitor
- o Type 1: Statistics Report
- o Type 2: Peer Down Notification
- o Type 3: Peer Up Notification
- o Type 4: Initiation
- o Type 5: Termination

Type values 6 through 128 MUST be assigned using the "Standards Action" policy, and values 129 through 255 using the "Specification Required" policy defined in [RFC5226].

9.2. BMP Statistics Types

This document defines nine statistics types for statistics reporting (Section 4.6):

- o Stat Type = 0: Number of prefixes rejected by inbound policy.
- o Stat Type = 1: Number of (known) duplicate prefix advertisements.
- o Stat Type = 2: Number of (known) duplicate withdraws.
- o Stat Type = 3: Number of updates invalidated due to CLUSTER_LIST loop.
- o Stat Type = 4: Number of updates invalidated due to AS_PATH loop.
- o Stat Type = 5: Number of updates invalidated due to ORIGINATOR_ID.
- o Stat Type = 6: Number of updates invalidated due to a loop found in AS_CONFED_SEQUENCE or AS_CONFED_SET.
- o Stat Type = 7: Number of routes in Adj-RIBs-In.
- o Stat Type = 8: Number of routes in Loc-RIB.

Stat Type values 9 through 32767 MUST be assigned using the "Standards Action" policy, and values 32768 through 65535 using the "Specification Required" policy, defined in [RFC5226].

9.3. BMP Initiation Message TLVs

This document defines three types for information carried in the Initiation message (Section 4.3):

- o Type = 0: String.
- o Type = 1: sysDescr.
- o Type = 2: sysName.

Information type values 3 through 32767 MUST be assigned using the "Standards Action" policy, and values 32768 through 65535 using the

"Specification Required" policy, defined in [RFC5226].

9.4. BMP Termination Message TLVs

This document defines two types for information carried in the Termination message (Section 4.4):

- o Type = 0: String.
- o Type = 1: Reason.

Information type values 2 through 32767 MUST be assigned using the "Standards Action" policy, and values 32768 through 65535 using the "Specification Required" policy, defined in [RFC5226].

9.5. BMP Termination Message Reason Codes

This document defines four types for information carried in the Termination message (Section 4.4) Reason code,:

- o Type = 0: Administratively closed.
- o Type = 1: Unspecified reason.
- o Type = 2: Out of resources.
- o Type = 3: Redundant connection.

Information type values 4 through 32767 MUST be assigned using the "Standards Action" policy, and values 32768 through 65535 using the "Specification Required" policy, defined in [RFC5226].

10. Security Considerations

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

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

Users of this protocol MAY use some type of secure transport mechanism, such as IPSec [RFC4303] or TCP-AO [RFC5925], in order to provide mutual authentication, data integrity and transport protection.

Unless a transport that provides mutual authentication is used, an attacker could masquerade as the monitored router and trick a monitoring station into accepting false information.

11. Acknowledgements

Thanks to Tim Evens, John ji Ioannidis, Mack McBride, Danny McPherson, Dimitri Papadimitriou, Erik Romijn, and the members of the GROW working group for their comments.

12. References

12.1. Normative References

- [RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets:MIB-II", STD 17, RFC 1213, March 1991.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.
- [RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

12.2. Informative References

- [RFC1155] Rose, M. and K. McCloghrie, "Structure and identification of management information for TCP/IP-based internets", STD 16, RFC 1155, May 1990.
- [RFC2856] Bierman, A., McCloghrie, K., and R. Presuhn, "Textual Conventions for Additional High Capacity Data Types", RFC 2856, June 2000.
- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, December 2005.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

Appendix A. Changes Between BMP Versions 1 and 2

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

Appendix B. Changes Between BMP Versions 2 and 3

- o Added a 32-bit length field to the fixed header.
- o Clarified error handling.
- o Added new stat types: 5 (number of updates invalidated due to ORIGINATOR_ID), 6 (number of updates invalidated due to AS_CONFED_SEQUENCE/AS_CONFED_SET), 7 (number of routes in Adj-RIB-In) and 8 (number of routes in Loc-RIB).
- o Defined counters and gauges for use with stat types.
- o For peer down messages, the relevant FSM event is to be sent in type 2 messages.
- o Added local address and local and remote ports to the peer up message.
- o Require End-of-RIB marker after initial dump.
- o Added Initiation message with string content.
- o Permit multiplexing pre- and post-policy feeds onto a single BMP session.
- o Changed assignment policy for IANA registries.
- o Changed "Loc-RIB" references to refer to "Post-Policy Adj-RIB-In", plus other editorial changes.
- o Introduced option for monitoring station to be active party in initiating connection.
- o Introduced Termination message.

Authors' Addresses

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

Email: jgs@juniper.net

Rex Fernando
Cisco Systems
170 W. Tasman Dr.
San Jose, CA 95134
USA

Email: rex@cisco.com

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

Email: [sstuart@google.com](mailto:ssstuart@google.com)

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 31, 2013

R. Shakir
BT
July 30, 2012

Operational Requirements for Enhanced Error Handling Behaviour in BGP-4
draft-ietf-grow-ops-reqs-for-bgp-error-handling-05

Abstract

BGP-4 is utilised as a key intra- and inter-Autonomous System routing protocol in modern IP networks. The failure modes as defined by the original protocol standards are based on a number of assumptions around the impact of session failure. Numerous incidents both in the global Internet routing table and within Service Provider networks have been caused by strict handling of a single invalid UPDATE message causing large-scale failures in one or more Autonomous Systems.

This memo describes the current use of BGP-4 within Service Provider networks, and outlines a set of requirements for further work to enhance the mechanisms available to a BGP-4 implementation when erroneous data is detected. Whilst this document does not provide specification of any standard, it is intended as an overview of a set of enhancements to BGP-4 to improve the protocol's robustness to suit its current deployment.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 31, 2013.

Copyright Notice

Copyright (c) 2012 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 (<http://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. Role of BGP-4 in Service Provider Networks	3
1.2. Overview of Operator Requirements for BGP-4 Error Handling	5
2. Errors within BGP-4 UPDATE Messages	7
2.1. Classifying BGP Errors and Expected Error Handling	8
2.1.1. Critical BGP Errors	9
2.1.2. Semantic BGP Errors	9
3. Avoiding use of NOTIFICATION	11
4. Recovering RIB Consistency	13
5. Reducing the Impact of Session Reset	15
6. Operational Toolset for Monitoring BGP	17
7. Operational Complexities Introduced by Altering RFC4271	21
7.1. Reducing the Network Impact of Session Teardown	23
8. IANA Considerations	25
9. Security Considerations	26
10. Acknowledgements	27
11. References	28
11.1. Normative References	28
11.2. Informational References	28
Author's Address	30

1. Introduction

Where BGP-4 [RFC4271] is deployed in the Internet and Service Provider networks, numerous incidents have been recorded due to the manner in which [RFC4271] specifies errors in routing information should be handled. Whilst the behaviour defined in the existing standards retains utility, the deployments of the protocol have changed within modern networks, resulting in significantly different demands for protocol robustness. Whilst a number of Internet Drafts have been written to begin to enhance the behaviour of BGP-4 in terms of the handling of erroneous messages, this memo intends to define a set of requirements for ongoing work. These requirements are considered from the perspective of a Network Operator, and hence this draft does not intend to define the protocol mechanisms by which such error handling behaviour is to be implemented.

1.1. Role of BGP-4 in Service Provider Networks

BGP was designed as an inter-Autonomous System (AS) routing protocol and hence many of the error handling mechanisms within the protocol specification are designed to be conducive to this role. In general, this consideration as an inter-AS routing propagation mechanism results in the view that a BGP session propagates a relatively small amount of network-layer reachability information (NLRI) between two ASes. In this case, it is the expectation of session resilience for those adjacencies that are key to routing continuity (for example, it is expected that two networks peering via BGP would connect multiple times in order to safeguard equipment or protocol failure). In addition, there is some expectation of multiple paths to a particular NLRI being available - it would be expected that a network can fall back to utilising alternate, less direct, paths where a failure of a more direct path occurs.

Traditional network architectures would deploy an Interior Gateway Protocol (IGP) to carry infrastructure and customer routes, with an Exterior Gateway Protocol (EGP) such as BGP being utilised to propagate these routes to other Autonomous Systems. However, with the growth of IP-based services, this is no longer considered best practice. In order to ensure that convergence is within acceptable time bounds, the amount of routing information carried within the IGP is significantly reduced - and tends to be only infrastructure routes. iBGP is then utilised to propagate both customer, and external routes within an AS. As such, BGP has become an IGP, with traditional IGPs acting as a means by which to propagate the routing information which is required to establish a BGP session, and reach the egress node within the local routing domain. This change in role presents different requirements for the robustness of BGP as a routing protocol - with the expectation of similar level of

robustness to that of an IGP being set.

Along with this change in role, the nature of the IP routing information that is carried has changed. BGP has become a ubiquitous means by which service information can be propagated between devices. For instance, BGP is utilised to carry routing information for IP/MPLS VPN services as described in [RFC4364]. Since there is an existing deployment of the protocol between PE devices in numerous networks, it has been adapted to propagate this routing information, as its use limits the number of routing protocols required on each device. This additional information being propagated represents a large change in requirement for the error handling of the protocol - where session failure occurs, it is likely a complete service outage for at least a subset of a network's customers is experienced where an erroneous packet may have occurred within a different sub-topology or even service (a different address family for example). For this reason, there is a significant demand to avoid service affecting failures that may be triggered by routing information within a single sub-topology or service.

The combination of the increased number of deployments of BGP-4 as an intra-AS routing protocol, its use for the propagation of additional types of routing and service information, and the growth of IP services has resulted in a substantial increase in the volume of information carried within BGP-4. In numerous networks, RIB sizes of the order of millions of entries exist within individual BGP speakers, with particularly high-scale points exhibited at BGP speakers performing aggregation or functionality designed improve utilisation of network resources (e.g., route reflector hierarchies). Clearly an increase in the amount routing information carried in BGP results in greater impact to services during failures, which is only amplified by a corresponding increase in recovery times. Following a failure, there is a substantial recovery time to learn, compute and distribute new paths, which results in a greater observed impact to services affected, and hence adds further weight to the requirement to avoid failures altogether or, at least, mitigate their impact to the narrowest scope possible, (e.g., a specific NLRI). Whilst an argument could be made that convergence time of BGP-4 could potentially be reduced through deployment of additional computational resource, it is notable that solution is not necessarily straightforward from an implementation or deployment perspective, (e.g., scaling computation resources within a single address-family is difficult). Thus, significant challenges continue to exist for operators when scaling BGP-4 deployments, and hence mechanisms which improve the scalability of BGP-4 are very important.

Both within Internet and multi-service routing architectures, a number of BGP sessions propagate a large proportion of the required

routing information for network operation. For Internet routing, these are typically BGP sessions which propagate the global routing table to an AS - failure of these sessions may have a large impact on network service, based on a single erroneous update. In an multi-service environment, typical deployments utilise a small number of core-facing BGP sessions, typically towards route reflector devices. Failure of these sessions may also result in a large impact to network operation. Clearly, the avoidance of conditions requiring these sessions to fail is of great utility to any network operator, and provides further motivation for the revision of the existing behaviour.

Whilst the behaviour in [RFC4271] is suited to ensuring that BGP messages with erroneous routing information in are limited in scope (by means of session reset), with the above considerations, it is clear that this mechanism is not suited to all deployments. It should, however, be noted that the change in scope affects the handling only of errors occurring after BGP session establishment. There is no current operational requirement to amend the means by which error handling in session establishment, or liveness detection, are performed.

1.2. Overview of Operator Requirements for BGP-4 Error Handling

It is the intention of this document to define a set of criteria for the manner in which a revised error handling mechanism in BGP-4 is required to conform. The motivation for the definition of these requirements can be summarised based on certain behaviour currently present in the protocol that is not deemed acceptable within current operational deployments, or where there is a short-fall in the tool set available to an operator. These key requirements can be summarised as follows:

- o It is unacceptable within modern deployments of the BGP-4 protocol that a single erroneous UPDATE packet affects routes that it does not carry. This requirement therefore requires some modification to the means by which erroneous UPDATE packets are handled, and reacted to - with a particular focus on avoiding the use of the NOTIFICATION message.
- o It is recognised that some error conditions may occur within the BGP-4 protocol may not always be handled gracefully, and may result in conditions whereby an implementation cannot recover. In these (and similar) cases, it is undesirable for an operator that this reset of the BGP-4 session results in interruption to forwarding packets (by means of withdrawing routes installed by BGP-4 into a device's RIB, and subsequently FIB). To this end, there is a requirement to define a session reset mechanism which

provides session re-initialisation in a non-destructive manner.

- o Further to the requirements to provide a more robust protocol, the current visibility into error conditions within the BGP-4 protocol is extremely limited - where further modifications to this behaviour are to be made, complexity is likely to be added. Thus, to ensure that BGP-4 is manageable, there are requirements for mechanisms by which the protocol can be examined and monitored.

This document describes each of these requirements in further depth, along with an overview of means by which they are expected to be achieved. In addition, the mechanism by which the enhancements meeting these requirements are to interact is discussed.

2. Errors within BGP-4 UPDATE Messages

Both through analysis of incidents occurring with the Internet DFZ, and multi-service environments utilising BGP-4 to signal service or routing information, a number of different classes of errors within BGP-4 UPDATE messages have been observed. In order to consider the applicability of enhanced error handling mechanisms, it is possible to divide these errors into a number of sub-classes, particularly focusing around the location of the error within the UPDATE message.

Where an UPDATE message is considered invalid by a BGP speaker due to an error within a path attribute that is not the NLRI (where the definition of NLRI includes reachability information encoded in the MP_REACH_NLRI and MP_UNREACH_NLRI attributes as specified in [RFC4760]) it is a requirement of any enhanced error handling mechanism to handle the error in a manner focused on the NLRI contained within the message found to be erroneous. Since in this case, the message received from the remote peer is syntactically valid, it is considered that such an UPDATE is indicative of erroneous data within one or more path attributes. The impact of the current behaviour defined within the protocol makes the implication that the BGP speaker from whom the message is received is now an invalid path for all NLRI announced via the session - which results in a disproportionate impact to overall network operation. In particular scenarios (such as networks with centralised BGP route reflection) such action can result in a loss of all reachability to a network. In other contexts (such as the Internet DFZ), it cannot be assumed that the BGP speaker from whom the UPDATE message is received is directly responsible for the erroneous information contained within the message.

Two further error cases exist within UPDATE messages, both of which are related to the mechanisms that are applicable to messages received where some difficulty exists in parsing the entire BGP message. The two cases concern those cases where a valid NLRI attribute can be extracted, and those where such an attribute is not able to be parsed. In these cases, errors in the packing of attributes within a BGP message may have occurred. Such errors are likely indicative of an error specifically caused by the remote BGP speaker. It is, however, desirable to an operator that such errors are handled without affecting all NLRI across a BGP session. As such, there is a key requirement to maximise the number of cases in which it is possible to extract NLRI from a BGP UPDATE message. To this end, it is required that where possible the MP_REACH_NLRI and MP_UNREACH_NLRI attributes are utilised for encoding all NLRI (including IPv4 Unicast), and that this attribute is included as the first attribute of a BGP UPDATE message (as originally recommended in [I-D.chen-ebgp-error-handling]). Such a change to the order of

inclusion of this attribute maximises the number of cases in which NLRI can be extracted from an UPDATE. Where this is possible, it is again required that the error handling mechanisms utilised should be directly applied to the NLRI included in the UPDATE.

For all cases whereby NLRI can be obtained from an UPDATE message, it is expected that the requirements outlined in Section 3 should be considered by any enhancement to the BGP-4 protocol.

In the case that it is not possible to completely parse the NLRI attribute from the UPDATE message received from a peer, it is extremely likely that this is indicative of a serious error with either the process of attribute packing, or buffer usage on the remote BGP speaker. In this case, clearly, it is not possible to apply any error handling mechanism that is limited to a specific set of NLRI, since an implementation has no knowledge of the NLRI included within the UPDATE message. In addition, such errors are considered to be relatively fundamental to the operation of a BGP implementation, and hence may indicate a case whereby significant system errors have occurred. The current BGP-4 standard results in a BGP speaker restarting a session with the remote BGP speaker. However where such an error does occur, it is required that a graceful mechanism is utilised to provide a lower impact to network operation. The requirements for enhancements of this nature to BGP-4 are outlined in Section 5, with the requirements outlined therein focused on providing a means by which system integrity can be restored whilst allowing for continued network operation.

2.1. Classifying BGP Errors and Expected Error Handling

It is clearly of advantage for BGP-4 implementations to utilise a consistent set of error handling mechanisms for the different types of errors that are described in Section 2, and provide consistent nomenclature to refer to them. It is therefore suggested that errors that are indicative of larger scale failures of a BGP speaker, and hence require some error handling at the session level are referred to as 'critical' errors, whilst those errors that are identified based on incorrect content of one of more attributes of a message are referred to as 'semantic' errors.

The errors identified within the following sections consider only those errors within the specifications at the time of writing, it is recommended that in the definition of future extensions to the BGP-4 specification, the error handling behaviour (and the category within which errors within the extension should be considered by an implementation) is defined.

2.1.1. Critical BGP Errors

As described in this document, it is of advantage to limit the number of 'critical' errors that occur within the protocol, therefore, based on analysis of the processing of BGP UPDATE messages, it is required that 'critical' error handling behaviour is applied to:

- o UPDATE Message Length errors - whereby the specified overall UPDATE message length is inconsistent with sum of the Total Path Attribute and Withdrawn Routes length. In this case, this is indicative of message packing failure, whereby the NLRI may not be correctly extracted.
- o Errors Parsing the NLRI attributes of an UPDATE message - where NLRI is carried in either the IPv4-Unicast Advertised or Withdrawn routes, or in the MP_REACH_NLRI or MP_UNREACH_NLRI attributes [RFC2858], it is not possible to target error handling mechanisms to specific NLRI, and hence session level mechanisms must be utilised.

It is expected that those requirements outlined in Section 5 are utilised to provide session-level handling of those errors identified as 'critical'.

2.1.2. Semantic BGP Errors

Where a BGP message is correctly formed, a number of cases exist whereby the contents of the UPDATE are not valid - in these cases, this represents errors that can be identified to affect specific NLRI. The following cases are expected to be classified as semantic errors:

- o Zero or invalid length errors in path attributes excluding those containing NLRI, or where the length of all path attributes contained within the UPDATE does not correspond to the total path attributes length. In this case, the NLRI can be correctly extracted, and hence acted upon.
- o Messages where invalid data or flags are contained in a path attribute that does not relate to the NLRI.
- o UPDATE messages missing mandatory attributes, unrecognised non-optional attributes or those that contain duplicate or invalid attributes (be they unsupported or unexpected).
- o Those messages where the NEXT_HOP, or MP_REACH next-hop values are missing, length zero, or invalid for the relevant AFI/SAFI.

In these cases, it is expected that these errors can be handled gracefully, following the requirements detailed in Section 3 and Section 4 of this memo.

3. Avoiding use of NOTIFICATION

The error handling behaviour defined in RFC4271 is problematic due to the limited options that are available to an implementation. When an erroneous BGP message is received, at the current time, the implementation must either ignore the error, or send a NOTIFICATION message, after which it is mandatory to terminate the BGP session. It is apparent that this requirement is at odds with that of protocol robustness.

There is significant complexity to this requirement. The mechanism defined in [I-D.chen-ebgp-error-handling] describes a means by which no NOTIFICATION message is generated for all cases whereby NLRI can be extracted from an UPDATE. The NLRI contained within the erroneous UPDATE message is considered as though the remote BGP speaker has provided an UPDATE marking it as withdrawn. This results in a limit in the propagation of the invalid routing information, whilst also ensuring that no traffic is forwarded via a previously-known path that may no longer be valid. This mechanism is referred to as "treat-as-withdraw".

Whilst this behaviour results in avoiding a NOTIFICATION message, keeping other routing information advertised by the remote BGP speaker within the RIB, it may result in unreachability for a sub-set of the NLRI advertised by the remote speaker. Two cases should be considered - that where the entry for a route in the Adj-RIB-In of the neighbour propagating an erroneous packet is utilised, and that where the route installed in the device's RIB is learnt from another BGP speaker. In the former case, should the identified NLRI not be treated as withdrawn, the original NLRI is utilised within the global RIB. However, this information is potentially now invalid (i.e. it no longer provides a valid forwarding path), whilst an alternate (valid) path may exist in another Adj-RIB-In. By continuing to utilise the NLRI for which the UPDATE was considered invalid, traffic may be forwarded via an invalid path, resulting in routing loops, or black-holing. In the second case, no impact to the forwarding of traffic, or global RIB, is incurred, yet where treat-as-withdraw is implemented, possibly stale routing information is purged from the Adj-RIB-In of the neighbour propagating errors.

Whilst mechanisms such as "treat-as-withdraw" are currently documented, the proposals are limited in their scope - particularly in terms of restrictions to implementation only on eBGP sessions. This limitation is made based on the view that the BGP RIB must be consistent across an autonomous system. By implementing treat-as-withdraw for a iBGP session, one or more routers within the Autonomous System may not have reachability to a route, and hence blackholing of traffic, or routing loops, may occur. It should,

however, be considered if this view is valid, in light of the manner in which BGP is utilised within operator networks. Inconsistency in a RIB based on a single UPDATE being treated as withdrawn may cause a inconsistency in a single sub-topology (e.g. Layer 3 VPN service), or a service not operating completely (in the case of an UPDATE carrying service membership information). Where a NOTIFICATION and teardown is utilised this is destructive to all sub-topologies in all address family identifiers (AFIs) carried by the session in question. Even where mechanisms such as multi-session BGP are utilised, a whole AFI is affected by such a NOTIFICATION message. In terms of routing operation, it is therefore far less costly to endure a situation where a limited sub-set of routing information within an AS is invalid, than to consider all routing information as invalid based on a single trigger.

At the time of writing, error handling mechanisms related to optional, transitive attributes - such as [I-D.ietf-idr-optional-transitive] are restricted to handling only a subset of attribute errors - whereas the operational requirement is to expand this coverage to the widest set of errors possible (i.e., all semantic errors within UPDATE messages). Additionally, where approaches applicable to a greater number of attributes are proposed (e.g., [I-D.chen-ebgp-error-handling]), these are limited to deployment in eBGP applications only, where requirements also exist in intra-domain cases. As such, it is envisaged that if extended to cover these expanded cases, these mechanisms provide a means to avoid the transmission of a NOTIFICATION message to a remote BGP speaker, based on a single erroneous message, where at all possible, and hence meet this requirement. Critical errors, including those whereby the NLRI cannot be extracted from the UPDATE message, represent cases whereby the receiving system cannot handle the error gracefully based on this mechanism.

4. Recovering RIB Consistency

The recommendations described in Section 3 may result in the RIB for a topology within an AS being inconsistent across the AS' internal routers. Alternatively, where such mechanisms are deployed at an AS boundary, interconnects between two ASes may be inconsistent with each other. There are therefore risks of traffic blackholing, due to missing routing information, or forwarding loops. Whilst this is deemed an acceptable compromise in the short term, clearly, it is suboptimal. Therefore, a requirement exists to provide mechanisms by which a BGP speaker is able to recover the consistency of the Adj-RIB-In for a particular neighbour.

In the general case, the consistency of the BGP RIB can be recovered by re-requesting the entire Adj-RIB-Out of a remote BGP speaker is re-advertised. A mechanism to achieve this re-advertisement is defined within the ROUTE-REFRESH specification [RFC2918]. It is envisaged that by requesting a refresh of all NLRI advertised by a BGP speaker, any NLRI which has been withdrawn due to being contained within an invalid UPDATE message is re-learned. Where a ROUTE REFRESH is used to directly perform a consistency check between the Adj-RIB-Out of a remote device, and the Adj-RIB-In of the local BGP speaker, a demarcation between the ROUTE-REFRESH, and normal UPDATE messages is required (in order that an "end" of the refresh can be used to identify any 'stale' NLRI) - [I-D.ietf-idr-bgp-enhanced-route-refresh] provides a means by which the ROUTE-REFRESH mechanism can be extended to meet this requirement.

Whilst re-advertisement of the whole BGP RIB provides a means by which withdrawn NLRI can be re-advertised, there are some scaling implications that must be considered. In the case that a ROUTE-REFRESH is generated, all NLRI must be re-packed into UPDATE messages and advertised by one speaker on the BGP session, whilst the other must receive all UPDATE messages, and validate the RIB's consistency. In order to avoid the control-plane load, it is therefore a requirement to utilise targeted mechanisms where possible, rather than incurring the additional load on both the advertising and receiving speaker of building and processing UPDATES for the entire contents of the RIB.

It is envisaged that during routing inconsistencies caused by utilising the 'treat-as-withdraw' mechanism, the local BGP speaker is aware that some routing information was not able to be processed - due to the fact that an UPDATE message was not parsed correctly. Since this mechanism (as discussed in Section 3) requires the local BGP speaker to have determined the set of NLRI for which an erroneous UPDATE message was received, it is possible to use a targeted mechanisms to re-request the specific NLRI that was contained within

the erroneous UPDATE message. By re-requesting, this provides the remote BGP speaker an opportunity to re-transmit the NLRI - possibly providing an opportunity to leverage alternative methods to build the UPDATE message. Such a request requires extension to the existing BGP-4 protocol, in terms of specific UPDATE generation filters with a transient lifetime. It is envisaged that the work within [I-D.zeng-idr-one-time-prefix-orf] provides a mechanism allowing targeted elements of the Adj-RIB-In for a BGP neighbour to be recovered.

It is of particular note for both means of recovering RIB consistency described that these are effective only when considering transient errors within an implementation - for instance, should an RFC interpretation error within an implementation be present, regardless of the number of times a specific UPDATE is generated, it is likely that this error condition will persist (as it may with the existing behaviour defined by [RFC4271]). For this reason, there is an requirement to consider the means by which such consistency recovery mechanisms are utilised. It is not advisable that a dynamic filter and advertisement mechanism is triggered by all error handling events due to the load this is likely to place on the neighbour receiving such a request. Where this BGP speaker is a relatively centralised device - a route reflector (as described by [RFC4456]) for example - the act of generation of UPDATE messages with such frequency is likely to cause disproportionate load. It is therefore an operational requirement of such mechanisms that means of request dampening be required by any such extension.

In cases whereby the consistency of the Adj-RIB-In is to be restored (e.g., following the 'treat-as-withdraw' behaviour described in Section 3), and mechanisms such as those described herein are triggered, such a condition should be noted to an operator by means of a specific flag, SNMP trap, or other logging mechanism. In order to identify the subset of NLRI that are considered to be inconsistent, this information is of operational benefit and hence should be logged.

5. Reducing the Impact of Session Reset

Even where protocol enhancements allow errors in the BGP-4 protocol to cease to trigger NOTIFICATION messages, and hence reset a BGP session, it is clear that some error conditions may not be exited. In particular, errors due to existing state, or memory structures, associated with a specific BGP session will not be handled. It is therefore important to consider how these error conditions are currently handled by the protocol. It should be noted that the following discussion and analysis considers only those NOTIFICATION messages generated in response to errors in UPDATE messages (as defined by Section 6.3 in [RFC4271]).

The existing NOTIFICATION behaviour triggers a reset of all elements of the BGP-4 session, as described in Section 6 of [RFC4271]. It is expected that session teardown requires an implementation to re-initialise all structures and state required for session maintenance. Clearly, there is some utility to this requirement, as error conditions in BGP are, in general, exited from. However, this definition is responsible for the forwarding outages within networks utilising BGP for propagation of routing or service when each error is experienced. The requirement described in Section 3 is intended to reduce the cases whereby a NOTIFICATION is required, however, any mechanism implemented as a response to this requirement by definition cannot provide a session reset to the extent of that achieved by the current behaviour.

In order to address this, there is a requirement for a means by which a BGP speaker can signal that an unhandled error condition in an UPDATE message occurred - requiring a session reset - yet also continue to utilise the paths advertised by the neighbour that are currently in use within the RIB. In this case, the Adj-RIB-In received from the neighbour is not considered invalid, despite a NOTIFICATION, and session reset, being required. This set of requirements is akin to those answered by the BGP Graceful Restart mechanism described in [RFC4724]. Since the operational requirement in this case is to provide a means to achieve a complete session restart without disrupting the forwarding path of those routes in use within a BGP speaker's RIB, it is expected that utilising a procedure similar to the Graceful Restart mechanism meets the error handling requirement. By responding to an error condition (repeated or otherwise) with a message indicating that an error that cannot be handled has occurred, forcing session reset, whilst retaining forwarding information within the RIB allows forwarding to all routes within a system's RIB to continue during the period in which the session restarts. It is envisaged that the additional complexity introduced by the introduction of such a mechanism can be limited by extending existing BGP messages - one such approach is proposed in

[I-D.ietf-idr-bgp-gr-notification]. By placing a time bound on the restart lifetime, should an error condition not be transient - for example, should an error have occurred with the BGP process, rather than a specific of the BGP session - the remote BGP speaker is still detected as an invalid device for forwarding.

In some cases, the erroneous condition may be due to corruption of the Adj-RIB-Out on the advertising BGP speaker - rather than caused by the receiving speaker's state. In these cases, where existing structures are replayed whilst performing graceful restart functionality, the error condition is not necessarily resolved. Therefore, it is recommended that during a session restart event, as described within this section, the advertising speaker purge and rebuild RIB structures, in order to resolve any corruption within these structures.

It should be noted that a protocol enhancement meeting this requirement is not able to solve all error conditions - however, a complete restart of the BGP and TCP session between two BGP speakers implements an identical recovery mechanism to that which is achieved by the existing behaviour. Where an error condition such as memory or configuration corruption has occurred in a BGP implementation, it is expected that a mechanism meeting this requirement continues to detect this, by means of a bound on time for session restart to occur. Whilst there may be some consideration that packets continue to be forwarded through a device which can be in a failure mode of this nature for a longer period due to this requirement, the architecture of modern IP routers should be considered. A divided forwarding and control plane is common in many devices, as well as process separation for software-based devices - corruption of a specific protocol daemon does not necessarily imply forwarding is affected. Indeed, where forwarding behaviour of a device is affected, it is envisaged that a failure detection mechanism (be it Bidirectional Forwarding Detection, or indeed BGP KEEPALIVE packets) will detect such a failure in almost all cases, with the symptomatic behaviour of such a failure being an invalid UPDATE message in very few other cases.

6. Operational Toolset for Monitoring BGP

A significant complexity that is introduced through the requirements defined in this document is that of monitoring BGP session status for an operator. Although the existing error handling behaviour causes a disproportionate failure, session failure is extremely visible to most operational personnel within a Network Operator due to both existing definitions of SNMP trap mechanisms for BGP, along with the forwarding impact typically caused by such a failure. By introducing mechanisms by which errors of this nature are not as visible, this is no longer the case. There is a requirement that where subsets of the RIB on a device are no longer reachable from a BGP speaker, or indeed an AS, that some visibility of this situation, alongside a mechanism to determine the cause is available to an operator. Whilst, to some extent, this can be solved by mandating a sub-requirement of each of the aforementioned requirements that a BGP speaker must log where such errors occur, and are hence handled, this does not solve all cases. In order to clarify this requirement, the example of the transmission of an erroneous Optional Transitive attribute can be considered. Since, by definition, there is no requirement for all BGP speakers to parse such an attribute, a receiving router may treat NLRI as withdrawn based on an erroneous attribute not examined by its neighbour. In this case, the upstream device or network, propagating the UPDATE, has no visibility of this error. Operationally, however, it is of interest to the upstream router operator that such invalid information was propagated.

The requirement for logging of error conditions in transmitted BGP messages, which are visible to only the receiver, cannot be achieved by any existing BGP message, or capability. It is envisaged that each erroneous event should be transmitted to the remote peer - including the information as to the set of NLRI that were considered invalid. Whilst with some mechanisms this is achieved by default (for example, One-Time Prefix ORF [I-D.zeng-idr-one-time-prefix-orf] (Outbound Route Filtering) will transmit the set of routes that are required), the operator requirement is to know which routes may have been unreachable in all cases. It is envisaged that an extension to meet this requirement will allow for such information to be transmitted between peers, and hence logged. Such a mechanism may provide further utility as a either a diagnostic, or logging toolset.

As such, it is possible to divide the messages that are required in order to provide further visibility into BGP for an operator. Such a division can be made both due to the required means of message transmission, alongside the criticality of each request.

- o Messages required to replace NOTIFICATION - In cases where the error handling mechanisms defined by [RFC4271] currently result in

a NOTIFICATION message being generated, a number of the requirements detailed within this document result this message being suppressed. Despite this change, the error condition's occurrence is still of interest to an operator in order to provide both monitoring and troubleshooting capabilities, since some form of invalid data has been received on a session. It therefore considered that an implementation must generate a message both locally, and transmitted to the remote peer, based on the such a condition. Where such a message is transmitted to the remote peer, it is considered that the BGP session via which the erroneous UPDATE message was received should be used as transport to the remote peer. The information transmitted in such a message should be minimised to allow identification of the paths which were considered erroneous (i.e. restricting the information to that which is directly relevant to a network operator in the case of an error condition occurring). Any delay to convergence on the session in question is considered to be acceptable, given the suboptimal nature of the reception of invalid routing information via a BGP session. Further concerns regarding such a mechanism relate to the load generated on the BGP speaker in question, however, it must be considered that in the case of an erroneous UPDATE being received, and the 'treat-as-withdraw' mechanism being utilised, where the erroneous path is removed from the Loc-RIB, there is likely to be a requirement to generate UPDATE messages withdrawing the route from all further BGP speakers to which the prefix is advertised. The load generated by the generation of such UPDATES is likely to be much greater than that of transmitting error information via a logging message type back to the speaker from which it was received. It is envisaged that light-weight BGP message-based signalling mechanisms such as the ADVISORY message types detailed in [I-D.ietf-idr-operational-message] provide a suitable means to satisfy this requirement.

- o Additional Diagnostic Capabilities for BGP - In a number of cases, there is an operational requirement to further debug erroneous BGP UPDATE messages, along with the particulars of the state of a BGP speaker. For instance, where an invalid BGP UPDATE message is transmitted between two BGP speakers, the exact format of the UPDATE message is of interest to an operator, as this information provides a clear indication of a message considered to be erroneous by the BGP speaker to which it was transmitted. In this case, it is considered of great utility that the entire UPDATE message is transmitted back to the advertising speaker, in order to allow for further debugging to occur. Whilst such information is particularly useful to an operator, it clearly provides information that is not key to protocol operation - for this reason, it is expected that some of the concerns regarding the

additional complexity, and load that a BGP speaker is subjected to is not acceptable. For this reason, it is required that where mechanisms are developed to support this requirement, messages of this nature can be supported both within an existing BGP session, and via a dedicated separate session, be it BGP carrying messages such as those defined in [I-D.ietf-idr-operational-message] or a dedicated monitoring protocol akin to BMP described in [I-D.ietf-grow-bmp].

Whilst the operational requirement for such monitoring tools to allow for visibility into BGP is clearly agreed upon, the means by which such messages are transmitted between two BGP speakers is likely to be dependent upon both the positions of the speakers in question (for instances, the requirements for such a protocol may differ where a session is between two ASBRs under separate administration). The introduction of additional message types to the BGP protocol clearly introduces further complexity - and leaves room for further implementation and standardisation errors that may compromise the robustness of the BGP protocol. In addition, the queuing and scheduling of these BGP messages must be interleaved with the transmission of the key protocol messages - such as KEEPALIVE and UPDATE packets. It is therefore a concern that should a large number of messages specifically for operational visibility be transmitted, this will delay the transmission of UPDATE packets, and hence adversely affect the end-to-end convergence time for NLRI carried within BGP. The operational requirement for why messages are advantageous to be in-band to a protocol should also be considered. In particular, it should be noted that where such information is to be transmitted between administrative boundaries a BGP session represents an existing channel between the two ASes. This channel is considered to be secure insofar as the routing information, and requests sent via the session are considered to come from a trusted source. Since error information relates to both a particular attachment, and is key to ensuring that such a session is operating as expected, it is considered of great operational benefit that this information is transmitted over this channel. In addition, the overall system scalability is improved by such in-band transmission. It is expected that erroneous information resulting in the 'treat-as-withdraw' mechanism being utilised is relatively infrequently transmitted between two peers (when compared to the frequency of UPDATE messages transmission). The impact of including an additional BGP message type for such operational visibility is relatively small from a resource utilisation perspective - additional processing overhead is only experienced when such a message is received. Where a separate session is maintained, particular network elements within a service provider topology may require hundreds, or thousands, of additional sessions for the transmission of this information. Such an resource consumption overhead is likely to be unacceptable to some

network operators.

For the reasons explained above, it is expected that mechanisms specified to meet the requirements for event visibility consider the relative impacts of additional monitoring sessions, or message inclusion in band to BGP in order not to compromise the security, scalability and robustness of the BGP-4 protocol.

7. Operational Complexities Introduced by Altering RFC4271

The existing NOTIFICATION and subsequent teardown of a BGP session upon encountering an error has the advantage that a consistent approach to error handling is required of all implementations of the BGP-4 protocol. This is of operational advantage as it provides a clear expectation of the behaviour of the protocol. The requirements defined herein add further complexity to the error-handling within BGP, and hence are liable to compromise the existing deterministic protocol behaviour. It is therefore deemed that there is a further requirement to define a set of recommended behaviours based on the reception of a particular class of erroneous UPDATE message, alongside highlighting some of the implementation complexities that may need to be handled in the case that particular recommendations made within this memo are deployed.

Utilising the classes of erroneous UPDATE message described in Section 2, the recommended behaviour for a BGP-4 implementation can be divided into two branches. Primarily, where a semantic error is identified, an implementation is expected to utilise the reduced-impact error handling approach, as described in Section 3. In the case that such an approach results in known NLRI being withdrawn from the BGP speaker's RIB, and an implementation provides functionality such that these errors are recovered from through an automatically triggered means, such as those described within Section 4, some consideration of the scalability of these recovery mechanisms is required. Clearly, there is an computational and bandwidth overhead associated with the re-advertisement of NLRI between two BGP speakers - both due to the generation of UPDATE messages, their transmission between the two speakers, and the parsing and processing into the RIB required. This overhead is directly proportional to the number of UPDATE messages that are required. Where a semantic error is experienced, by definition the NLRI contained within the UPDATE can be extracted. It is therefore possible to minimise the proportion of the RIB that is re-advertised by targeting any recovery mechanism on the NLRI contained within the erroneous UPDATE. Such a targeted mechanism can be achieved through a means such as One-Time ORF, or other means of targeting UPDATE messages not discussed within this memo. It is recommended that where available, any automatic (or manual) triggered recovery mechanism behaviour utilises such targeted means in preference to any whole RIB refresh mechanism (such as ROUTE-REFRESH).

In the case that an erroneous UPDATE has been processed through a means such as treat-as-withdraw (described within Section 3), a recovering mechanism may be considered superfluous, if the assumption is made that the RIB inconsistency will only be recovered from based on a path re-convergence (or change in BGP attribute) for the

advertising BGP speaker. However, where this assumption is not considered to provide adequate recovery behaviour, and a mechanism to restore RIB consistency automatically is implemented, some consideration must be made for where repeated erroneous messages occur. In this case, in order to limit the impact to the BGP speaker's network operation, at a pre-defined point it is recommended that such automatic recovery mechanisms towards the BGP speaker from which erroneous UPDATES are repeatedly received are suppressed, and the fact that such suppression has occurred is highlighted to an operator. The point at which such behaviour is suppressed is to be defined on a per-implementation basis, taking into account feedback from the Network Operator community based on the deployment of the recommendations described in this document. It is expected that such trigger points are dependent upon the mechanisms implemented for a particular BGP-4 implementations, and the impact upon the speaker of these means of RIB recovery.

Where critical errors are experienced, such that a session reset is required, the mechanism discussed in Section 5 should be used. Again, since such a mechanism results in a restart of a BGP session, it is expected that all NLRI carried over the session is re-advertised as it is re-established, incurring processing overhead on both the advertising and receiving BGP speaker. In order to minimise the consumption of control-plane computational resource on both speakers, it is recommended that mechanisms allowing a reduced set of BGP UPDATE messages to be re-transmitted between two speakers are employed wherever possible - for instance through employing mechanisms such as those described in [I-D.ietf-idr-enhanced-gr].

In the case that repeated critical errors occur, the overhead of performing any mechanism implemented based on the requirements in Section 5 is incurred following each erroneous UPDATE message. Since these mechanisms are, by definition, performed automatically in response to the erroneous message being received similar considerations as to the impact to the BGP speaker must be taken into account. As such, it is expected that after a certain trigger level, the ongoing receipt of critical errors within BGP UPDATE messages is deemed to be indicative of a long-lasting failure, and a session no longer considered viable. Where such a case is experienced, it is expected that the BGP session reverts to the standard session failure behaviour, as described in [RFC4271] and documents updating this base standard. Where such a reversion is implemented this condition should be flagged to a network operator. The number of restart attempts before the session reverts to being shut down should be determined based on the overhead of the recovery mechanisms implemented (for instance, where [I-D.ietf-idr-enhanced-gr] is implemented, the impact of session restart may be significantly lower), and operational experience of the deployment of the

recommendations described in this document.

Since repeated erroneous UPDATE messages which experience critical errors may be indicative of long-lasting failure modes, it is recommended that a back-off from restarting BGP sessions experiencing such behaviour is implemented. As such, this is not applicable to restart behaviour through means such as those described in Section 5 since such restarts are time-bound based on the period for which the Adj-RIB-In from a BGP speaker is maintained as valid (e.g., when considering BGP Graceful Restart, such restarts are time-bound by the Restart Time described in [RFC4724]). However, following a session reverting to being pulled down based on repeated error conditions, it is recommended that following restart attempts are subject to an exponentially increasing interval between subsequent attempts. It is therefore recommended that in such cases an implementation implements the increasing values of IdleHoldTimer as described in the BGP-4 FSM documented in [RFC4271].

7.1. Reducing the Network Impact of Session Teardown

As discussed within the preceding section, where repeated critical UPDATE message errors are received, it is recommended that the impact to the both advertising and receiving BGP-4 speakers be limited by reverting to tearing the BGP-4 session experiencing such errors down. The BGP-4 specification presented in [RFC4271] achieves such a session shutdown by sending a NOTIFICATION message, however, this has the net result that all downstream BGP speakers (i.e. those to whom the routes carried over the now ceased BGP session was readvertised) must withdraw this route from their RIB, and perform a best-path selection if required. In some cases, there may be no alternate path available, and hence a period of time for which no valid BGP route exists. Particularly, this is very likely to occur where an upstream BGP speaker performs a best-path selection and advertises only a single path to its neighbours - there is a requirement for the upstream speaker to perform a best-path selection, and re-advertise a new set of NLRI before the downstream system is able to converge to a new path. It should be noted that where UPDATE messages withdrawing NLRI are not subject to the BGP session's configured MinRouteAdvertisementInterval (MRAI) [RFC4271], but re-advertisements are, this may result in a BGP speaker being without a path for a period up to the MRAI.

Clearly, it is advantageous to avoid this period of time for which there may be no reachability for a set of routes, especially since the BGP speaker terminating a particular session is doing so due to a particular error handling policy. The graceful shutdown mechanism detailed in [I-D.ietf-grow-bgp-gshut] provides a mechanism by which a BGP speaker is able to signal that a set of routes are to be

withdrawn, and hence allow downstream systems to pre-emptively perform a best-path selection, and hence advertise new reachability information in a make-before-break manner.

It is therefore envisaged, that where a session is to be shutdown, based on a trigger relating to erroneous UPDATE messages being received (be they repeated or not) that the graceful shutdown procedure is utilised, so as to reduce the forwarding impact of routes received on the session being withdrawn.

8. IANA Considerations

This memo includes no request to IANA.

9. Security Considerations

The requirements outlined in this document provide mechanisms by which erroneous BGP messages may be responded to with limited impact to forwarding operation. This is of benefit to the security of a BGP speaker in general. Where UPDATE messages may have been propagated by a single malicious Autonomous System or router within a network (or the Internet default free zone - DFZ), which are then propagated to all devices within the same routing domain, all other NLRI available over the same session become unreachable. This mechanism may provide means by which an Autonomous System can be isolated from required routing domains (such as the Internet), should the relevant UPDATE messages be propagated via specific paths. By reducing the impact of such failures, it is envisaged that this possibility may be constrained to a specific set of NLRI, or a specific topology.

Some mechanisms meeting the requirements specified in this document, particularly those within Section 6 may provide further security concerns, however, it is envisaged that these are addressed in per-enhancement memos.

10. Acknowledgements

The author would like to thank the following network operators for their insight, and valuable input in defining the requirements for a variety of operational deployments of the BGP-4 protocol; Shane Amante, Bruno Decraene, Rob Evans, David Freedman, Wes George, Tom Hodgson, Sven Huster, Jonathan Newton, Neil McRae, Thomas Mangin, Tom Scholl and Ilya Varlashkin.

In addition, many thanks are extended to Jeff Haas, Wim Hendrickx, Tony Li, Alton Lo, Keyur Patel, John Scudder, Adam Simpson and Robert Raszuk for their expertise relating to implementations of the BGP-4 protocol.

11. References

11.1. Normative References

- [RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.
- [RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

11.2. Informational References

- [I-D.chen-ebgp-error-handling]
Chen, E., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP Updates from External Neighbors", draft-chen-ebgp-error-handling-01 (work in progress), September 2011.
- [I-D.ietf-grow-bgp-gshut]
Francois, P., Decraene, B., Pelsser, C., Patel, K., and C. Filssils, "Graceful BGP session shutdown", draft-ietf-grow-bgp-gshut-03 (work in progress), December 2011.
- [I-D.ietf-grow-bmp]
Scudder, J., Fernando, R., and S. Stuart, "BGP Monitoring Protocol", draft-ietf-grow-bmp-06 (work in progress), December 2011.
- [I-D.ietf-idr-bgp-enhanced-route-refresh]

Patel, K., Chen, E., and B. Venkatachalapathy, "Enhanced Route Refresh Capability for BGP-4", draft-ietf-idr-bgp-enhanced-route-refresh-02 (work in progress), June 2012.

[I-D.ietf-idr-bgp-gr-notification]

Patel, K., Fernando, R., and J. Scudder, "Notification Message support for BGP Graceful Restart", draft-ietf-idr-bgp-gr-notification-00 (work in progress), December 2011.

[I-D.ietf-idr-enhanced-gr]

Patel, K., Chen, E., Fernando, R., and J. Scudder, "Accelerated Routing Convergence for BGP Graceful Restart", draft-ietf-idr-enhanced-gr-01 (work in progress), June 2012.

[I-D.ietf-idr-operational-message]

Freedman, D., Raszuk, R., and R. Shakir, "BGP OPERATIONAL Message", draft-ietf-idr-operational-message-00 (work in progress), March 2012.

[I-D.ietf-idr-optional-transitive]

Scudder, J., Chen, E., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", draft-ietf-idr-optional-transitive-04 (work in progress), October 2011.

[I-D.zeng-idr-one-time-prefix-orf]

Zeng, Q., Dong, J., Heitz, J., Patel, K., Shakir, R., and Z. Huang, "One-time Address-Prefix Based Outbound Route Filter for BGP-4", draft-zeng-idr-one-time-prefix-orf-02 (work in progress), July 2012.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.

Author's Address

Rob Shakir
BT
pp C3L
BT Centre
81, Newgate Street
London EC1A 7AJ
UK

Email: rob.shakir@bt.com
URI: <http://www.bt.com/>

