Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: June 30, 2013

R. Shakir
BT

December 27, 2012

Operational Requirements for Enhanced Error Handling Behaviour in BGP-4

draft-ietf-grow-ops-reqs-for-bgp-error-handling-06

Abstract

BGP 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 within service provider networks, and outlines a set of requirements for further work to enhance the mechanisms available to a BGP 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 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 June 30, 2013.

Copyright Notice

Table of Contents

1. Requirements Language 3
2. Problem Statement 4
2.1. Role of BGP-4 in Service Provider Networks 4
3. Critical and Non-Critical Errors 7
4. Error Handling for Non-Critical Errors 9
4.1. NLRI-level Error Handling Requirements 9
4.2. Recovering RIB Consistency following NLRI-level Error Handling 10
5. Error Handling for Critical Errors 12
6. IANA Considerations 14
7. Security Considerations 15
8. Acknowledgements 16
9. References 17
9.1. Normative References 17
9.2. Informational References 17
Author's Address 19

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. Problem Statement

BGP has become a key intra- and inter-domain routing protocol, deployed within both the Internet and private networks. The increased reliance on the protocol has resulted in increased demand for robustness - with the error handling behaviour defined in [RFC4271] having been shown to have caused numerous incidents within live network deployments. This document provides an overview of the current deployment cases for BGP-4, and define a set of requirements (from the perspective of a network operator) for enhancing error handling within the protocol.

2.1. Role of BGP-4 in Service Provider Networks

BGP was designed as an inter-autonomous system (AS) routing protocol. Many of the error handling mechanisms within the protocol are defined in order to be guarantee consistency, and correctness of information between two neighbouring speakers. The assumption is made that each AS operates with many adjacencies, each propagating a relatively small amount of routing information. Through focusing on information consistency, the protocol specification prefers failure of an individual routing adjacency to maintaining reachability to all NLRI received from a particular neighbour, with the expectation that alternate, less direct, paths can be selected where a failure occurs. The assumptions of the nature of BGP deployments resulted in the specification made in [RFC4271] whereby the receipt of an erroneous UPDATE message is reacted to by sending a NOTIFICATION message, and tearing down the adjacency with the remote speaker from whom the error was observed.

Historically, a network would deploy an interior gateway protocol (IGP) to carry infrastructure and customer routes, and utilise an external gateway protocol (EGP) such as BGP to propagate routes to other autonomous systems. However, BGP's deployments have evolved with the growth of IP-based services. To ensure route convergence within an AS is within acceptable time bounds the amount of information within the IGP has been minimised (typically to only infrastructure routes). iBGP is then utilised to carry both internal, customer and external routes within an AS. As such, this has resulted in BGP having become an IGP, with traditional IGPs providing only reachability between nodes within the AS for packet forwarding and to establish iBGP sessions. This change in role within the overall architecture of an AS has resulted in an increased robustness requirement for BGP, with the expectation of a similar level of robustness to that of an IGP being set. The loss of an iBGP session can result in significant levels of unreachability internally to an AS, especially since there are typically limited (when compared to the Internet) signalling and forwarding paths available.

In parallel with this change of deployment, the volume and nature of the information carried within BGP has also changed. BGP has become the ubiquitous means through which service information can be propagated between devices. For instance, being utilised to carry IP/MPLS service information such as Layer 3 IP VPN routes [RFC4364] , and Layer 2 Virtual Private LAN Service device membership [RFC4761]. Since these extensions to the protocol allow signalling of multiple services (represented by address families within BGP), and multiple customer topologies (i.e., subsets of routes within each address family) via the BGP protocol, the impact of session failure is increased. The tear down of a single BGP session can result in a complete outage to all customer services signalled via the session, even where the triggering event is related to only one service or topology being carried - reflecting a disproportional impact to all other services and routing topologies.

The convergence of services to IP, and BGP's changing deployment has resulted in a significant growth in the volume of routing information carried in the protocol. In numerous networks, the RIB size of individual BGP speakers can be of the order of millions of paths. Particularly large RIBs are observed at BGP speakers performing aggregation and border roles (such as ASBR, or route reflector hierarchies). This increased volume of routes results not only in a significant number of services being impacted during a protocol failure, but also increases the time to recovery after re-establishing a BGP session. The time taken to learn, compute and distribute new paths increases the impact of failures on services carried by the network - adding further weight to the requirement to avoid failures, or limit the extent of their impact. Furthermore, the impact of individual session failures is increased due to the existence of a relatively small number of highly-critical BGP sessions within Internet and multi-service network deployments. These sessions propagate a high-proportion of the reachability information - for instance, providing an Internet AS with the global routing table from upstream providers, or connecting IP/MPLS Provider Edge devices to route reflector hierarchies from which they are signalled reachability for services connected elsewhere within the routing domain. In both cases, the failure of these sessions can result in a significant outage to customer services.

For the current deployments of BGP, the behaviour described in [RFC4271] related to handling errors in UPDATE messages is suboptimal, and results in significant disruption to services in modern network deployments. This document defines a set of requirements for protocol developments, and revisions to [RFC4271] to address these concerns through a set of generalised definitions. It should be noted that the scope of these requirements is limited to the handling of UPDATE messages as, at the time of writing, there is

no operational requirement to amend the means by which error handling in session establishment, or liveliness detection are performed.

3. Critical and Non-Critical Errors

As described in Section 2.1, the error handling behaviour described in [RFC4271] is applied at a per-session level, affecting all NLRI signalled via the adjacency on which an erroneous message is observed. In order to reduce the impact of error handling to those NLRI affected by an erroneous UPDATE, a BGP speaker MUST limit the error handling mechanisms implemented to those NLRI contained within an erroneous UPDATE message where it is possible to do so. Clearly, some errors within the formation of BGP UPDATE messages may result in it being impossible to reliably extract NLRI from the received message, and hence the same error handling procedures may not apply. There is therefore a requirement to classify errors based on their impact to the BGP UPDATE message, hence messages whereby the NLRI attribute cannot be extracted or parsed are referred to throughout this document as Critical errors. These Critical errors are limited to:

- UPDATE Message Length errors - where the specified UPDATE message length is inconsistent with the sum of the Total Path Attribute and Withdrawn Routes length. These errors relate to message packing or framing, and result in cases whereby the NLRI attribute cannot be correctly extracted from the message.

- Errors parsing the NLRI attribute of an UPDATE message - where the contents of the IPv4 Unicast Advertised or Withdrawn Routes attributes, or multi-protocol BGP NLRI attributes (MP_REACH_NLRI and/or MP_UNREACH_NLRI as defined in [RFC2858]), cannot be successfully parsed.

In the case of Critical errors is expected that error handling is applied at a session level as per Section 5 of this document.

All errors whereby the contained NLRI can be extracted, are referred to as Non-Critical. It is expected that the following cases fall within this category:

- 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 attribute length.

- Messages where invalid data or flags are contained in a path attribute that does not relate to the NLRI.

- UPDATE messages missing mandatory attributes, unrecognised non-optional attributes, or those that contain duplicate or invalid attributes (be they unsupported, or unexpected).

- Those messages where the NEXT_HOP, the MP_REACH_NLRI next-hop values are missing, zero-length, or invalid for the relevant address family.

For these Non-Critical errors, the NLRI-targeted error handling requirements described in Section 4 should be followed.

In order to maximise the number of cases whereby the NLRI attributes can be reliably extracted from a received message, where a BGP speaker supports multi-protocol extensions, the MP_REACH_NLRI and MP_UNREACH_NLRI attributes SHOULD be utilised for all address families (including IPv4 Unicast) and these attributes should be the first attribute contained within the UPDATE message.

Where attributes are introduced by future extensions to the BGP protocol the error handling behaviour applied MUST be assumed that applied to Non-Critical errors, unless otherwise specified within the per-extension memo, or the attribute relates directly to carrying NLRI. Authors of future BGP extensions SHOULD specify the error handling behaviour required for new attributes in terms of the classification into a Critical or Non-Critical error on a per-attribute error basis.

4. Error Handling for Non-Critical Errors

4.1. NLRI-level Error Handling Requirements

When a Non-Critical error is detected within an UPDATE message a BGP speaker MUST NOT send a NOTIFICATION message to the remote neighbour. Instead, the NLRI contained within the message MUST be considered as no longer viable until they are updated by a subsequent UPDATE message, thus treating the NLRI as withdrawn as per the treat-as-withdraw mechanism described in [I-D.chen-ebgp-error-handling].

Network operators SHOULD recognise that where such behaviour is implemented black-holing or looping of traffic may occur in the period between the NLRI being treated as withdrawn, and subsequent updates, dependent upon the routing topology. It SHOULD be noted that such periods of RIB inconsistency (where one speaker has advertised a prefix, which has been treated as withdrawn by the receiving speaker) may be relatively long lived, based on situations such as an erroneous implementation at the receiver, or the error occurring within an optional, transitive attribute not examined by the advertising device. In order to allow operators to select sessions on which this risk of inconsistency is acceptable, an implementation SHOULD provide means by which NLRI-level error handling for Non-Critical errors can be disabled on a per-session basis.

Since the Non-Critical error handling required within this section results in no NOTIFICATION message being transmitted, the fact that an error has occurred and hence there may be inconsistency between the local and remote BGP speaker MUST be flagged to the network operator through standard operational interfaces (e.g., SNMP, syslog). The information highlighted MUST include the NLRI identified to be contained within the error message, and SHOULD contain a exact copy of the received message for further analysis.

In order that the operator of the BGP speaker from whom an erroneous UPDATE message has been advertised is aware of the fact that some NLRI advertised to the remote speaker have been considered withdrawn due to being contained within an erroneous UPDATE, a BGP speaker SHOULD support mechanisms to report the occurrence of Non-Critical error handling to the remote speaker. The receiving speaker SHOULD transmit the NLRI contained within the erroneous message to the advertising speaker. An exact copy of the received UPDATE message SHOULD also be sent.

The exchange of information related to events occurring as a result of BGP messages is not currently supported by any extension to the protocol. Clearly, where the two speakers reside within the same

administrative domain, shared logging infrastructure can be utilised to identify the root cause of errors, however, in many cases neighbouring BGP speakers reside within separate administrative domains (e.g., are ASBRs for Internet or private networks). In this case, mechanisms allowing transmission in-band to the BGP session SHOULD be utilised (e.g., the OPERATIONAL message described in [I-D.ietf-idr-operational-message]). Such an in-band channel is preferred based on the BGP session representing a pre-established trusted channel which is related to a specific BGP-speaking device within a network. It is expected that the overall system scalability of a BGP speaker is improved through utilising the existing channel, rather than incurring overhead for maintaining many additional logging-specific protocol sessions for relatively infrequent messaging events when errors occur. However, the extensions providing such a channel MUST consider their impact to base BGP protocol functions such as the transmission of UPDATE or KEEPALIVE messages, and SHOULD limit the volume of messaging to direct reactions to Non-Critical errors occurring. These considerations SHOULD be made in order to ensure that no compromise is made to the security, scalability and robustness of BGP. Where additional BGP monitoring information that is not suitable to be carried in-band is required, out-of-band mechanisms such as the BMP protocol described in [I-D.ietf-grow-bmp] could be utilised to provide further information relating to erroneous messages.

4.2. Recovering RIB Consistency following NLRI-level Error Handling

Following NLRI being treated as withdrawn due to Non-Critical error handling, inconsistencies exist between the Adj-RIB-Out of the advertising BGP speaker, and the Adj-RIB-In of the receiving device. These inconsistencies may result in forwarding loops or blackholing of traffic in some routing topologies. In order to ensure that such cases can be recovered from a means by which a validation and recovery of consistency can be achieved SHOULD be provided to an operator. This function may be provided through enhancing the ROUTE-REFRESH [RFC2918] mechanism to add means to identify the beginning and end of a replay of the entire Adj-RIB-Out of the advertising speaker (as per the suggestion in [I-D.ietf-idr-bgp-enhanced-route-refresh]).

As Non-Critical error handling is localised to the NLRI contained within the erroneous UPDATE message, a targeted recovery mechanism MAY be provided allowing a speaker to request re-advertisement of a particular subset of the Adj-RIB-Out. Where such targeted refresh functions are available, they SHOULD be preferred to mechanisms requesting re-advertisement of the whole Adj-RIB-Out based on their more limited use of CPU and network resources.

A BGP speaker may automatically trigger recovery mechanisms such as those described in this section following the receipt of an erroneous UPDATE message identified as Non-Critical to expedite recovery. It should be noted that if automatic recovery mechanisms trigger only re-advertisement of an identical erroneous message, they are likely to be ineffective. Additionally, where the best-path to be advertised by remote speaker changes, this will be advertised directly, without a requirement for a request from the receiver. However, in some cases, RIB consistency recovery mechanisms may prompt alternate UPDATE message packing, and hence allow quicker recovery. Where such mechanisms are implemented, mechanisms focused to smaller sets of NLRI SHOULD be preferred over those requesting the entire RIB. In addition, such mechanisms SHOULD have dampening mechanisms to ensure that their impact to computational and network resources is limited.

5. Error Handling for Critical Errors

Where an UPDATE message containing a Critical error is received, since the NLRI cannot be extracted, error handling mechanisms must be applied at the per-session level. In order to limit the impact to network operation, these session-level mechanisms MUST be applied in a manner which allows the paths NLRI received from the remote speaker to continue to be utilised for forwarding during the session reset and re-establishment. It is envisaged that this requirement may be met through extension of the BGP Graceful Restart mechanism ([RFC4724]) to be triggered by NOTIFICATION messages indicating the occurrence of a Critical error. Such an extension allows a restart of the TCP and BGP sessions between two speakers, in a similar manner to the current session restart behaviour triggered by a NOTIFICATION message. In order to maximise the level of re-initialisation which occurs during such a restart triggered by a Critical error, BGP speakers MAY re-initialise memory structures related to the Adj-RIB-In and Adj-RIB-Out associated with the session on which the erroneous UPDATE was observed.

Where such a restart event occurs, the continued liveliness of the remote device MAY be verified by BGP KEEPALIVE packets or other OAM functions such as Bidirectional Forwarding Detection ([RFC5880]). In cases where the observed Critical BGP error is indicative of a wider device failure of the remote speaker, it is expected that a BGP sessions will not re-establish correctly. Each BGP speaker SHOULD maintain a limited time window in which session restart is expected in order to mitigate this possibility.

When a Critical error occurs, the network operator MUST be made aware of its occurrence through local logging mechanisms (e.g., SNMP traps or syslog). The BGP speaker receiving an UPDATE message identified as a Critical error MUST log its occurrence and a copy of the UPDATE message. Where a inter-device messaging mechanism is implemented (as discussed in Section Section 4.1) a copy of the erroneous UPDATE message SHOULD be transmitted to the remote speaker. Both BGP speakers MUST indicate to an operator the cause of a session restart was a Critical error in an UPDATE message.

Since repeated critical errors (and session restarts) may have an impact in overall device scaling if the failure condition is not resolved by session restart, a BGP speaker MAY choose to revert to the session tear down behaviour described in the base BGP specification. This reversion SHOULD only be utilised after a number of attempts which SHOULD be controllable by the network operator. Where a session is shut down, the implementation MAY utilise a back-off from session restart attempts (as per the IdleHoldTimer described in the BGP FSM [RFC4271]). Where reversion to tearing down the BGP

session is performed, a speaker SHOULD limit the impact of withdrawing prefixes from downstream speakers where possible. It is envisaged that this can be achieved by utilising a mechanism such as the BGP Graceful Shutdown procedure as described in [I-D.ietf-grow-bgp-gshut].

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

The requirements outlined in this document provide mechanisms which limit the overall impact of the response to an error in a BGP UPDATE message. This is of benefit to the security of a BGP speaker. Without these mechanisms, where erroneous UPDATE messages relating to a single NLRI entry can be propagated to a BGP speaker, all other NLRI carried via the same session are affected by the resulting session tear-down. This may result in an AS being isolated from particular routing domains (such as the Internet) should an UPDATE message be propagated via targeted specific paths. It is envisaged by reducing the impact of the reaction of the receiving speaker to these messages, the isolation can be constrained to specific sets of NLRI, or a specific topology.

A number of the mechanisms meeting the requirements specified within the document (particularly those relating to operational monitoring) may raise further security concerns. Such concerns will be addressed during the specification of these mechanisms.

8. Acknowledgements

The author would like to thank the following network operators for their insight, and valuable input into defining the requirements for a variety of deployments of the BGP 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 protocol.

9. References

9.1. Normative References

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

[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.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

9.2. 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. Filsfils, "Graceful BGP session shutdown", draft-ietf-grow-bgp-gshut-04 (work in progress), October 2012.

[I-D.ietf-grow-bmp]
Scudder, J., Fernando, R., and S. Stuart, "BGP Monitoring Protocol", draft-ietf-grow-bmp-07 (work in progress), October 2012.

[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-03 (work in progress), December 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.

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/