Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: May 13, 2015

R. Shakir
BT

November 9, 2014

Operational Requirements for Enhanced Error Handling Behaviour in BGP-4

draft-ietf-grow-ops-reqs-for-bgp-error-handling-07

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 May 13, 2015.

Table of Contents

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 developed into a key intra- and inter-domain routing protocol, deployed within both the Internet and private networks. The changing deployments of the protocol have resulted in increased demand for robustness of the routing system - with the error handling behaviour defined in [RFC4271] having been shown to have caused numerous incidents within live network deployments. This document intends to provide 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 propagated through a particular neighbour, with the expectation that alternate, less direct, paths can be selected where a failure occurs. These assumptions 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.

BGP's deployments have evolved with the growth of IP-based services. 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, within modern deployments 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.

The volume and nature of the information carried within BGP has also changed - it 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.

In addition, there has been significant growth in the volume of routing information carried in BGP. In numerous networks, the RIB size of individual BGP speakers can be of the order of millions of paths. Particularly large volumes 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. Particularly, 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 providing IP/MPLS Provider Edge devices adjacency with route reflector hierarchy providing signalling for elements of services connected elsewhere within the routing domain. In both cases, the failure of these sessions can result in a significant outage to customer services.

2.2. Service Requirements for Amended BGP Error Handling

Alongside the infrastructure requirements outlined above, service provider customer requirements continue to evolve. In particular, there are increasing requirements for robustness and fault isolation based on:

- The increasing reliance on public IP service instead of private networks - resulting is requirements for greater availability of Internet services. The diversity of autonomous systems has resulted in individual BGP sessions within the Internet carrying

more routing information (e.g., IP transit, or large peering interconnections), which is originated from more individual networks - increasing both the impact of an individual session failure, and the number of different sources of error which can lead to its failure. To meet the requirement of high-availability Internet services, it is therefore an expectation that the error handling behaviour MUST affect only the those routes, or autonomous systems, that are are impacted by the erroneous messages, rather than all routes received by a particular session, such that the maximum service availability is maintained.

- The requirement to support multiple services. In multi-service environments such as those that support L3VPNs, multiple customer VPNs are isolated from one another, and from other IP environments (such as the Internet). There is an expectation from a service perspective therefore that the customer service is within its own fault domain (even when carried via a shared set of signalling), hence an error on routes or BGP messages related to one VPN SHOULD NOT negatively impact other VPNs. Further to this, an error relating to another service (i.e., another address family, such as Internet or L2VPN services) SHOULD NOT impact the availability of the VPN service. Both of these principles of fault separation are

required in order to support multiple services and segregated customer infrastructures over a common network infrastructure whilst meeting the availability required of them.

It should be noted that the requirements for fault isolation and high-availability do not imply that routing information that is potentially erroneous (through being carried in an UPDATE message that cannot be parsed for example) is always maintained despite questions as to its integrity, particularly as such routing information may result in leakage between services - but merely that there is a requirement to reconsider the balance between protocol correctness, and robustness.

In addition to these service requirements, an increasing requirement to minimise the time taken to recover from incidents exists. In some cases, this may require an operator to compromise on correctness in order to maintain integrity of a subset of routing information or services. To meet this requirement, mechanisms allowing an operator to ignore all errors or maintain "known good" routing information MAY be required. The implementation of such mechanisms is a business consideration of the service provider in question, and MUST consider the balance between the risk of incorrectness and the overall impact to a network platform. Such mechanisms are particularly of use where lack of routing information violates an operator's policies (e.g., filtering rules distributed via BGP FlowSpec are no longer installed), or fault isolation requires significant external liaison

(such as contacting a third-party autonomous system to amend or filter route announcement).

3. Classes of Errors within UPDATE Messages

To meet the requirement to provide more targeted error handling, errors are therefore classified into the following scopes:

- Attribute Scope - in this case, an error can be localised to a particular attribute within the message. For instance, such errors may occur when invalid flags are set within an individual attribute within a message, which is otherwise well-formed.

- Message Scope - errors resulting in the inability to parse a single UPDATE message, but not affecting the ability of an

implementation to parse subsequent BGP messages. For instance, where the overall length of an UPDATE message is correct, but the length of a single attribute contained within it is erroneously specified.

- Session Scope - where errors occur such that an error in an UPDATE message results in the inability to the parse subsequent messages. In this case, attribute length errors may result in the inability for a BGP implementation to locate the bounds of an UPDATE, and hence the subsequent message from a peer.

For session-scope errors, the error handling approach implemented MUST conform with the requirements described in Section 5 of this document (generically referred to as "Critical" error handling mechanisms). Session-scope errors requiring Critical error handling MUST be the only case whereby the impact of error handling mechanisms should be allowed to impact entire BGP sessions between two BGP speakers.

For message- and attribute-level errors, "Non-Critical" error handling mechanisms SHOULD be used, which MUST meet the specification described in Section 4. In the case of attribute-scope errors, a BGP speaker MUST limit the impact of error-handling mechanisms to the NLRI carried within the message, and MAY (where applicable) limit to the scope of error handling to the individual attribute. Where a message-scope error occurs, a BGP speaker MUST limit the impact of error handling to the NLRI contained within the affected UPDATE.

3.1. Characteristics of Session Scope Errors

Based on analysis of existing BGP implementations, and incidents within the Internet and private network routing tables, it is expected that errors with a session level scope are restricted 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.

3.2. Characteristics of Message Scope Errors

Message scope errors are restricted to those whereby erroneous encoding results in the ability to parse and determine the NLRI carried by the message - but the carried attributes are invalid. These errors (based on existing attributes) are limited to:

- Errors where the length of all path attributes contained within the UPDATE does not correspond to the total path attribute length.
- 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.

3.3. Characteristics of Attribute Scope Errors

Attribute scope errors are defined to be those that relate to an individual attribute (not related to the NLRI) carried by an UPDATE message. Particularly, where:

- Zero- or invalid-length errors in path attributes, excluding those containing NLRI.
- Invalid data or flags are contained in a path attribute that does not relate to the NLRI.

3.4. Avoiding Session Scope Errors

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

3.5. Future Attributes introduced to BGP

Where attributes are introduced by future extensions to the BGP protocol error handling behaviour SHOULD be assumed to be be at a message- or attribute-scope, unless otherwise specified within the per-extension memo, or the attribute relates directly to carrying NLRI. It is recommended that authors of future BGP extensions SHOULD specify the error handling behaviour required 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 SHOULD be considered as being withdrawn by the neighbour (referred to as treat-as-withdraw), until they are updated by a subsequent UPDATE message. Where defined is acceptable by the relevant memo, for the specific-case of attribute-scope errors, the erroneous attribute MAY be discarded by an implementation. This attribute-discard approach MUST only be used for attributes that do not impact best-path selection within an implementation. An operator SHOULD consider the impact of implementing policies considering such attributes as part of the route selection algorithm, such that operator configuration does not result in unexpected consequences should such an attribute be discarded.

Network operators SHOULD recognise that where treat-as-withdraw 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 had treat-as-withdraw applied to it 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 direct neighbour. In order to allow operators to select sessions on which this risk of inconsistency is acceptable, an implementation SHOULD provide means by which Non-Critical error handling 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 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.

4.1.1. Notifying the Remote Peer of Non-Critical Errors

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 invalid, 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 such 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 information can be utilised to identify the root cause of errors. However, in many cases these devices 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 source 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 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

In order to recover consistency of Adj-RIBs following Non-Critical error handling, a means by which a validation and recovery of consistency can be achieved SHOULD be provided to an operator. This functionality MAY be provided through extension of the ROUTE-REFRESH [RFC2918] mechanism - providing 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 may 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 automatic mechanisms are implemented, those focused on 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

Critical error handling MUST be used where session-scope errors occur. In such cases, a NOTIFICATION message MUST be sent to the remote peer. In order to limit the impact to network operation, during such events the mechanisms applied MUST allow for the paths

NLRI received from the remote speaker to continue to be utilised 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 RIB where possible.

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]). If 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. By default, 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 upon session-re-establishment (or via a separate session if implemented). 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 Critical error handling does not resolve the failure condition, 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 MUST 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].

5.1. Long-Lived Critical Errors

Where Critical error handling mechanisms are required to be utilised, significant impact to an operator's network or services may still be experienced. In order to allow an operator to avoid such scenarios:

- An implementation MAY provide functionality whereby all future Critical errors result in UPDATE messages being discarded. Such functionality MUST be disabled by default, and SHOULD be configurable on a per-address-family basis. An operator MUST

consider such mechanisms as a tool of last-resort to maintain service for a subset of NLRI, whilst the root cause of a such errors is investigated and resolved. This MAY be achieved by filtering erroneous NLRI at an upstream peer.

- Provide means by which a the restart timer for Graceful Restart can be configured to be a long period (order of days, or weeks) such that a critical failure can be resolved whilst maintaining operation for a subset of NLRI. This restart period MUST be configured separately to standard graceful-restart timers and MUST be configurable per-address-family. Long-lived restart mechanisms MAY be configurable to be utilised by default. An operator MUST configure the impact to forwarding correctness of such configuration, based on the expected rate of change of NLRI within a particular <AFI,SAFI>.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

The requirements outlined in this document provide mechanisms which limit the forwarding 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 a means by which an AS can be 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

Many thanks are extended to Bruno Decraene and David Freedman for their numerous detailed reviews, and significant contribution towards the refinement of the requirements in this document.

In addition, 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 BGP: Shane Amante, Colin Bookham, Rob Evans, Wes George, Tom Hodgson, Sven Huster, Jonathan Newton, Neil McRae, Thomas Mangin, Tom Scholl and Ilya Varlashkin. 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-06 (work in progress), August 2014.

[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-10 (work in progress), June 2014.

[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 plc.
pp. C3L,
BT Centre,
81, Newgate Street,
London. EC1A 7AJ
UK

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