Internet Engineering Task Force
Internet-Draft
Updates: 4724 (if approved)
Intended status: Standards Track
Expires: May 31, 2019

K. Patel
Arrcus
R. Fernando
Cisco Systems
J. Scudder
J. Haas
Juniper Networks
November 27, 2018

Notification Message support for BGP Graceful Restart
draft-ietf-idr-bgp-gr-notification-16.txt

Abstract

The BGP Graceful Restart mechanism defined in RFC 4724 limits the usage of BGP Graceful Restart to BGP protocol messages other than a BGP NOTIFICATION message. This document updates RFC 4724 by defining an extension that permits the Graceful Restart procedures to be performed when the BGP speaker receives a BGP NOTIFICATION Message or the Hold Time expires. This document also defines a new BGP NOTIFICATION Cease Error subcode whose effect is to request a full session restart instead of a Graceful Restart.

Status of This Memo

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

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

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

This Internet-Draft will expire on May 31, 2019.

Table of Contents

1. Introduction

For many classes of errors, the BGP protocol must send a NOTIFICATION message and reset the peering session to handle the error condition. The BGP Graceful Restart extension defined in [RFC4724] requires that normal BGP procedures defined in [RFC4271] be followed when a NOTIFICATION message is sent or received. This document defines an extension to BGP Graceful Restart that permits the Graceful Restart procedures to be performed when the BGP speaker receives a NOTIFICATION message or the Hold Time expires. This permits the BGP speaker to avoid flapping reachability and continue forwarding while the BGP speaker restarts the session to handle errors detected in the BGP protocol.

At a high level, this document can be summed up as follows. When a BGP session is reset, both speakers operate as "Receiving Speakers" according to [RFC4724], meaning they retain each other's routes. This is also true for HOLDTIME expiration. The functionality can be defeated using a "Hard Reset" subcode for the BGP NOTIFICATION Cease

Error code. If a Hard Reset is used, a full session reset is performed.

1.1. Requirements Language

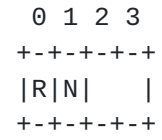
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Modifications to BGP Graceful Restart Capability

The BGP Graceful Restart Capability is augmented to signal the Graceful Restart support for BGP NOTIFICATION messages. The Restart Flags field is augmented as follows (following the diagram from section 3 of [RFC4724]):

Restart Flags:

This field contains bit flags relating to restart.



The most significant ("Restart State", or "R") bit is defined in [RFC4724].

The second most significant bit ("N") is defined as the BGP Graceful Notification bit, which is used to indicate Graceful Restart support for BGP NOTIFICATION messages. A BGP speaker indicates support for the procedures of this document, by advertising a Graceful Restart Capability with its Graceful Notification bit set (value 1).

If a BGP speaker that previously advertised a given set of Graceful Restart parameters opens a new session with a different set of parameters, these new parameters apply once the session has transitioned into ESTABLISHED state.

3. BGP Hard Reset Subcode

We define a new BGP NOTIFICATION Cease message subcode, called the BGP Hard Reset Subcode. The value of this subcode is discussed in Section 9. We refer to a BGP NOTIFICATION Cease message with the Hard Reset subcode as a Hard Reset message, or just a Hard Reset.

When the "N" bit has been exchanged by two peers, to distinguish them from Hard Reset we refer to any NOTIFICATION messages other than Hard Reset as "Graceful", since such messages invoke Graceful Restart semantics.

3.1. Sending a Hard Reset

A Hard Reset message is used to indicate to a peer with which the Graceful Notification bit has been exchanged, that the session is to be fully terminated.

When sending a Hard Reset, the data portion of the NOTIFICATION is encoded as follows:

```
+--------+--------+--------
| ErrCode| Subcode| Data
+--------+--------+--------
```

ErrCode is a BGP Error Code (as documented in the IANA BGP Error Codes registry) that indicates the reason for the Hard Reset. Subcode is a BGP Error Subcode (as documented in the IANA BGP Error Subcodes registry) as appropriate for the ErrCode. Similarly, Data is as appropriate for the ErrCode and Subcode. In short, the Hard Reset encapsulates another NOTIFICATION message in its data portion.

3.2. Receiving a Hard Reset

Whenever a BGP speaker receives a Hard Reset, the speaker MUST terminate the BGP session following the standard procedures in [RFC4271].

4. Operation

A BGP speaker that is willing to receive and send BGP NOTIFICATION messages according to the procedures of this document MUST advertise the BGP Graceful Notification "N" bit using the Graceful Restart Capability as defined in [RFC4724].

When such a BGP speaker has received the "N" bit from its peer, and receives from that peer a BGP NOTIFICATION message other than a Hard Reset, it MUST follow the rules for the Receiving Speaker mentioned in Section 4.1. The BGP speaker generating the BGP NOTIFICATION message MUST also follow the rules for the Receiving Speaker.

When a BGP speaker resets its session due to a HOLDTIME expiry, it should generate the relevant BGP NOTIFICATION message as mentioned in [RFC4271], but subsequently it MUST follow the rules for the Receiving Speaker mentioned in Section 4.1.

A BGP speaker SHOULD NOT send a Hard Reset to a peer from which it has not received the "N" bit. We note, however, that if it did so the effect would be as desired in any case, since according to [RFC4271] and [RFC4724] any NOTIFICATION message, whether recognized or not, results in a session reset. Thus the only negative effect to be expected from sending the Hard Reset to a peer that hasn't advertised compliance to this specification would be that the peer would be unable to properly log the associated information.

Once the session is re-established, both BGP speakers SHOULD set their "Forwarding State" bit to 1. If the "Forwarding State" bit is not set, then according to the procedures of [RFC4724] section 4.2, the relevant routes will be flushed, defeating the goals of this specification.

4.1. Rules for the Receiving Speaker

[RFC4724] section 4.2 defines rules for the Receiving Speaker. These are modified as follows.

The sentence "To deal with possible consecutive restarts, a route (from the peer) previously marked as stale MUST be deleted" only applies when the "N" bit has not been exchanged with the peer:

OLD: When the Receiving Speaker detects termination of the TCP session for a BGP session with a peer that has advertised the Graceful Restart Capability, it MUST retain the routes received from the peer for all the address families that were previously received in the Graceful Restart Capability and MUST mark them as stale routing information. To deal with possible consecutive restarts, a route (from the peer) previously marked as stale MUST be deleted. The router MUST NOT differentiate between stale and other routing information during forwarding.

NEW: When the Receiving Speaker detects termination of the TCP session for a BGP session with a peer that has advertised the Graceful Restart Capability, it MUST retain the routes received from the peer for all the address families that were previously received in the Graceful Restart Capability and MUST mark them as stale routing information. The router MUST NOT differentiate between stale and other routing information during forwarding. If the "N" bit has not been exchanged with the peer, then to deal with possible consecutive restarts, a route (from the peer) previously marked as stale MUST be deleted.

The stale timer is given a formal name and made mandatory:

OLD: To put an upper bound on the amount of time a router retains the stale routes, an implementation MAY support a (configurable) timer that imposes this upper bound.

NEW: To put an upper bound on the amount of time a router retains the stale routes, an implementation MUST support a (configurable) timer, called the "stale timer", that imposes this upper bound. A suggested default value for the stale timer is 180 seconds. An implementation MAY provide the option to disable the timer (i.e., to provide an infinite retention time) but MUST NOT do so by default.

5. Use of Hard Reset

5.1. When to Send Hard Reset

Although when to send a Hard Reset is an implementation-specific decision, we offer some advice. Many Cease notification subcodes represent permanent or long-term rather than transient session termination, and as such it's appropriate to use Hard Reset with them. At time of publication, Cease subcodes 1-9 were defined.

Value	Name	Suggested Behavior
1	Maximum Number of Prefixes Reached	Hard Reset
2	Administrative Shutdown	Hard Reset
3	Peer De-configured	Hard Reset
4	Administrative Reset	Provide user control
5	Connection Rejected	Graceful Cease
6	Other Configuration Change	Graceful Cease
7	Connection Collision Resolution	Graceful Cease
8	Out of Resources	Graceful Cease
9	Hard Reset	Hard Reset

Suggestions for Cease Subcode Behavior

These suggestions are only that, suggestions, not requirements. It's the nature of BGP implementations that the mapping of internal states to BGP NOTIFICATION codes and subcodes is not always perfect. The guiding principle for the implementor should be that if there is no realistic hope that forwarding can continue or that the session will be re-established within the deadline, Hard Reset should be used.

For all other NOTIFICATION codes other than Cease, use of Hard Reset does not appear to be indicated.

5.2. Interaction With Other Specifications

"BGP Administrative Shutdown Communication" [RFC8203] specifies use of the data portion of the Administrative Shutdown or Administrative Reset Cease to convey a short message. When [RFC8203] is used in conjunction with Hard Reset, the subcode of the outermost Cease MUST be Hard Reset, with the Administrative Shutdown or Reset Cease encapsulated within. The encapsulated administrative shutdown message MUST subsequently be processed according to [RFC8203].

6. Management Considerations

When reporting a Hard Reset to network management, the error code and subcode reported MUST be Cease, Hard Reset. If the network management layer in use permits it, the information carried in the Data portion SHOULD be reported as well.

7. Operational Considerations

Note that long (or infinite) retention time may cause operational issues, and should be enabled with care.

8. Acknowledgements

The authors would like to thank Jim Uttaro for the suggestion, and Emmanuel Baccelli, Bruno Decraene, Chris Hall, Warren Kumari, Paul Mattes, Robert Raszuk, and Alvaro Retana for their review and comments.

9. IANA Considerations

IANA has temporarily assigned subcode 9, named "Hard Reset", in the "BGP Cease NOTIFICATION message subcodes" registry. Upon publication of this document as an RFC, IANA is requested to make this allocation permanent.

IANA is requested to establish a registry within the "Border Gateway Protocol (BGP) Parameters" grouping, to be called "BGP Graceful Restart Flags". The Registration Procedure should be Standards Action, the reference this document and [RFC4724], and the initial values as follows:

Bit Position	Name	Short Name	Reference
0	Restart State	R	[RFC4724]
1	Notification	N	this document
2, 3	unassigned		this document

IANA is requested to establish a registry within the "Border Gateway Protocol (BGP) Parameters" grouping, to be called "BGP Graceful Restart Flags for Address Family". The Registration Procedure should be Standards Action, the reference this document and [RFC4724], and the initial values as follows:

Bit Position	Name	Short Name	Reference
0	Forwarding State	F	[RFC4724]
1-7	unassigned		this document

10. Security Considerations

This specification doesn't change the basic security model inherent in [RFC4724], with the exception that the protection against repeated resets is relaxed. To mitigate the consequent risk that an attacker could use repeated session resets to prevent stale routes from ever being deleted, we make the stale routes timer mandatory (in practice it is already ubiquitous). To the extent [RFC4724] might be said to help defend against denials of service by making the control plane more resilient, this extension may modestly increase that resilience; however, there are enough confounding and deployment-specific factors that no general claims can be made.

11. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <https://www.rfc-editor.org/info/rfc4724>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8203] Snijders, J., Heitz, J., and J. Scudder, "BGP Administrative Shutdown Communication", RFC 8203, DOI 10.17487/RFC8203, July 2017, <https://www.rfc-editor.org/info/rfc8203>.

Authors' Addresses

Keyur Patel
Arrcus

Email: keyur@arrcus.com

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

Email: rex@cisco.com

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

Email: jgs@juniper.net

Jeff Haas
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: jhaas@juniper.net