Network Working Group
Request for Comments: 4724
Category: Standards Track

S. Sangli
E. Chen
Cisco Systems
R. Fernando
J. Scudder
Y. Rekhter
Juniper Networks

January 2007

Graceful Restart Mechanism for BGP

Status of This Memo

Copyright Notice

Abstract

This document describes a mechanism for BGP that would help minimize the negative effects on routing caused by BGP restart. An End-of-RIB marker is specified and can be used to convey routing convergence information. A new BGP capability, termed "Graceful Restart Capability", is defined that would allow a BGP speaker to express its ability to preserve forwarding state during BGP restart. Finally, procedures are outlined for temporarily retaining routing information across a TCP session termination/re-establishment.

The mechanisms described in this document are applicable to all routers, both those with the ability to preserve forwarding state during BGP restart and those without (although the latter need to implement only a subset of the mechanisms described in this document).

Table of Contents

1. Introduction

Usually, when BGP on a router restarts, all the BGP peers detect that the session went down and then came up. This "down/up" transition results in a "routing flap" and causes BGP route re-computation, generation of BGP routing updates, and unnecessary churn to the forwarding tables. It could spread across multiple routing domains. Such routing flaps may create transient forwarding blackholes and/or transient forwarding loops. They also consume resources on the control plane of the routers affected by the flap. As such, they are detrimental to the overall network performance.

This document describes a mechanism for BGP that would help minimize the negative effects on routing caused by BGP restart. An End-of-RIB marker is specified and can be used to convey routing convergence information. A new BGP capability, termed "Graceful Restart Capability", is defined that would allow a BGP speaker to express its ability to preserve forwarding state during BGP restart. Finally, procedures are outlined for temporarily retaining routing information across a TCP session termination/re-establishment.

1.1 Specification of Requirements

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. Marker for End-of-RIB

An UPDATE message with no reachable Network Layer Reachability Information (NLRI) and empty withdrawn NLRI is specified as the End-of-RIB marker that can be used by a BGP speaker to indicate to its peer the completion of the initial routing update after the session is established. For the IPv4 unicast address family, the End-of-RIB marker is an UPDATE message with the minimum length [BGP-4]. For any other address family, it is an UPDATE message that contains only the MP_UNREACH_NLRI attribute [BGP-MP] with no withdrawn routes for that <AFI, SAFI>.

Although the End-of-RIB marker is specified for the purpose of BGP graceful restart, it is noted that the generation of such a marker upon completion of the initial update would be useful for routing convergence in general, and thus the practice is recommended.

In addition, it would be beneficial for routing convergence if a BGP speaker can indicate to its peer up-front that it will generate the End-of-RIB marker, regardless of its ability to preserve its forwarding state during BGP restart. This can be accomplished using the Graceful Restart Capability described in the next section.

3. Graceful Restart Capability

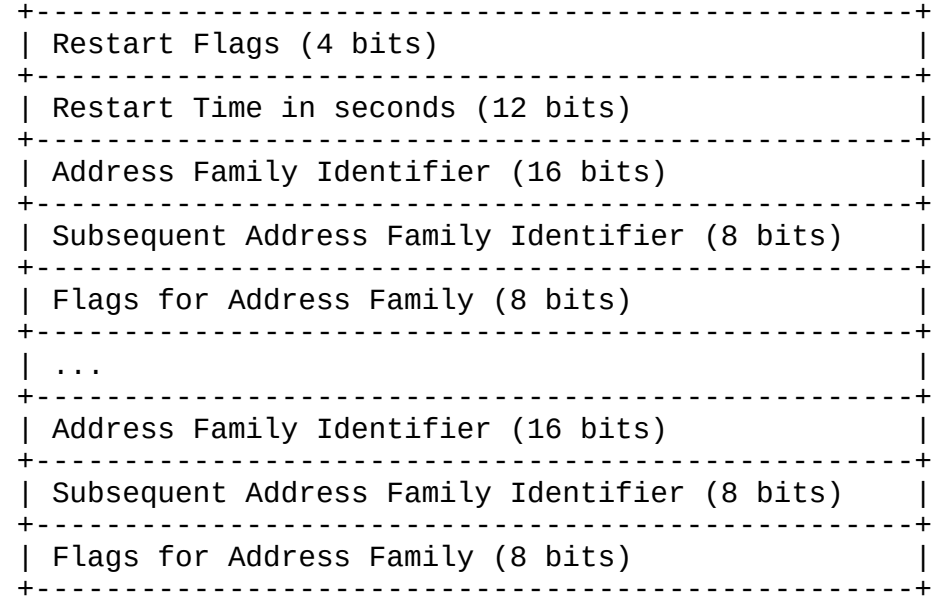
The Graceful Restart Capability is a new BGP capability [BGP-CAP] that can be used by a BGP speaker to indicate its ability to preserve its forwarding state during BGP restart. It can also be used to convey to its peer its intention of generating the End-of-RIB marker upon the completion of its initial routing updates.

This capability is defined as follows:

Capability code: 64

Capability length: variable

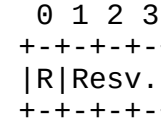
Capability value: Consists of the "Restart Flags" field, "Restart Time" field, and 0 to 63 of the tuples <AFI, SAFI, Flags for address family> as follows:



The use and meaning of the fields are as follows:

Restart Flags:

This field contains bit flags related to restart.



The most significant bit is defined as the Restart State (R) bit, which can be used to avoid possible deadlock caused by waiting for the End-of-RIB marker when multiple BGP speakers peering with each other restart. When set (value 1), this bit indicates that the BGP speaker has restarted, and its peer MUST NOT wait for the End-of-RIB marker from the speaker before advertising routing information to the speaker.

The remaining bits are reserved and MUST be set to zero by the sender and ignored by the receiver.

Restart Time:

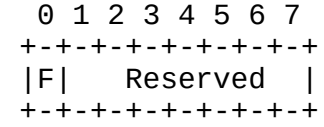
This is the estimated time (in seconds) it will take for the BGP session to be re-established after a restart. This can be used to speed up routing convergence by its peer in case that the BGP speaker does not come back after a restart.

Address Family Identifier (AFI), Subsequent Address Family Identifier (SAFI):

The AFI and SAFI, taken in combination, indicate that Graceful Restart is supported for routes that are advertised with the same AFI and SAFI. Routes may be explicitly associated with a particular AFI and SAFI using the encoding of [BGP-MP] or implicitly associated with <AFI=IPv4, SAFI=Unicast> if using the encoding of [BGP-4].

Flags for Address Family:

This field contains bit flags relating to routes that were advertised with the given AFI and SAFI.



The most significant bit is defined as the Forwarding State (F) bit, which can be used to indicate whether the forwarding state for routes that were advertised with the given AFI and SAFI has indeed been preserved during the previous BGP restart. When set (value 1), the bit indicates that the forwarding state has been preserved.

The remaining bits are reserved and MUST be set to zero by the sender and ignored by the receiver.

When a sender of this capability does not include any <AFI, SAFI> in the capability, it means that the sender is not capable of preserving its forwarding state during BGP restart, but supports procedures for the Receiving Speaker (as defined in Section 4.2 of this document). In that case, the value of the "Restart Time" field advertised by the sender is irrelevant.

A BGP speaker MUST NOT include more than one instance of the Graceful Restart Capability in the capability advertisement [BGP-CAP]. If more than one instance of the Graceful Restart Capability is carried in the capability advertisement, the receiver of the advertisement MUST ignore all but the last instance of the Graceful Restart Capability.

Including <AFI=IPv4, SAFI=unicast> in the Graceful Restart Capability does not imply that the IPv4 unicast routing information should be carried by using the BGP multiprotocol extensions [BGP-MP] -- it could be carried in the NLRI field of the BGP UPDATE message.

4. Operation

A BGP speaker MAY advertise the Graceful Restart Capability for an address family to its peer if it has the ability to preserve its forwarding state for the address family when BGP restarts. In addition, even if the speaker does not have the ability to preserve its forwarding state for any address family during BGP restart, it is still recommended that the speaker advertise the Graceful Restart Capability to its peer (as mentioned before this is done by not including any <AFI, SAFI> in the advertised capability). There are two reasons for doing this. The first is to indicate its intention of generating the End-of-RIB marker upon the completion of its initial routing updates, as doing this would be useful for routing convergence in general. The second is to indicate its support for a peer which wishes to perform a graceful restart.

The End-of-RIB marker MUST be sent by a BGP speaker to its peer once it completes the initial routing update (including the case when there is no update to send) for an address family after the BGP session is established.

It is noted that the normal BGP procedures MUST be followed when the TCP session terminates due to the sending or receiving of a BGP NOTIFICATION message.

A suggested default for the Restart Time is a value less than or equal to the HOLDTIME carried in the OPEN.

In the following sections, "Restarting Speaker" refers to a router whose BGP has restarted, and "Receiving Speaker" refers to a router that peers with the restarting speaker.

Consider that the Graceful Restart Capability for an address family is advertised by the Restarting Speaker, and is understood by the Receiving Speaker, and a BGP session between them is established. The following sections detail the procedures that MUST be followed by the Restarting Speaker as well as the Receiving Speaker once the Restarting Speaker restarts.

4.1. Procedures for the Restarting Speaker

When the Restarting Speaker restarts, it MUST retain, if possible, the forwarding state for the BGP routes in the Loc-RIB and MUST mark them as stale. It MUST NOT differentiate between stale and other information during forwarding.

To re-establish the session with its peer, the Restarting Speaker MUST set the "Restart State" bit in the Graceful Restart Capability

of the OPEN message. Unless allowed via configuration, the "Forwarding State" bit for an address family in the capability can be set only if the forwarding state has indeed been preserved for that address family during the restart.

Once the session between the Restarting Speaker and the Receiving Speaker is re-established, the Restarting Speaker will receive and process BGP messages from its peers. However, it MUST defer route selection for an address family until it either (a) receives the End-of-RIB marker from all its peers (excluding the ones with the "Restart State" bit set in the received capability and excluding the ones that do not advertise the graceful restart capability) or (b) the Selection_Deferral_Timer referred to below has expired. It is noted that prior to route selection, the speaker has no routes to advertise to its peers and no routes to update the forwarding state.

In situations where both Interior Gateway Protocol (IGP) and BGP have restarted, it might be advantageous to wait for IGP to converge before the BGP speaker performs route selection.

After the BGP speaker performs route selection, the forwarding state of the speaker MUST be updated and any previously marked stale information MUST be removed. The Adj-RIB-Out can then be advertised to its peers. Once the initial update is complete for an address family (including the case that there is no routing update to send), the End-of-RIB marker MUST be sent.

To put an upper bound on the amount of time a router defers its route selection, an implementation MUST support a (configurable) timer that imposes this upper bound. This timer is referred to as the "Selection_Deferral_Timer". The value of this timer should be large enough, so as to provide all the peers of the Restarting Speaker with enough time to send all the routes to the Restarting Speaker.

If one wants to apply graceful restart only when the restart is planned (as opposed to both planned and unplanned restart), then one way to accomplish this would be to set the Forwarding State bit to 1 after a planned restart, and to 0 in all other cases. Other approaches to accomplish this are outside the scope of this document.

4.2. Procedures for the Receiving Speaker

When the Restarting Speaker restarts, the Receiving Speaker may or may not detect the termination of the TCP session with the Restarting Speaker, depending on the underlying TCP implementation, whether or not [BGP-AUTH] is in use, and the specific circumstances of the restart. In case it does not detect the termination of the old TCP session and still considers the BGP session as being established, it

MUST treat the subsequent open connection from the peer as an indication of the termination of the old TCP session and act accordingly (when the Graceful Restart Capability has been received from the peer). See Section 8 for a description of this behavior in terms of the BGP finite state machine.

"Acting accordingly" in this context means that the previous TCP session MUST be closed, and the new one retained. Note that this behavior differs from the default behavior, as specified in [BGP-4], Section 6.8. Since the previous connection is considered to be terminated, no NOTIFICATION message should be sent -- the previous TCP session is simply closed.

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.

In re-establishing the session, the "Restart State" bit in the Graceful Restart Capability of the OPEN message sent by the Receiving Speaker MUST NOT be set unless the Receiving Speaker has restarted. The presence and the setting of the "Forwarding State" bit for an address family depend upon the actual forwarding state and configuration.

If the session does not get re-established within the "Restart Time" that the peer advertised previously, the Receiving Speaker MUST delete all the stale routes from the peer that it is retaining.

A BGP speaker could have some way of determining whether its peer’s forwarding state is still viable, for example through Bidirectional Forwarding Detection [BFD] or through monitoring layer two information. Specifics of such mechanisms are beyond the scope of this document. In the event that it determines that its peer’s forwarding state is not viable prior to the re-establishment of the session, the speaker MAY delete all the stale routes from the peer that it is retaining.

Once the session is re-established, if the "Forwarding State" bit for a specific address family is not set in the newly received Graceful Restart Capability, or if a specific address family is not included in the newly received Graceful Restart Capability, or if the Graceful Restart Capability is not received in the re-established session at

all, then the Receiving Speaker MUST immediately remove all the stale routes from the peer that it is retaining for that address family.

The Receiving Speaker MUST send the End-of-RIB marker once it completes the initial update for an address family (including the case that it has no routes to send) to the peer.

The Receiving Speaker MUST replace the stale routes by the routing updates received from the peer. Once the End-of-RIB marker for an address family is received from the peer, it MUST immediately remove any routes from the peer that are still marked as stale for that address family.

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.

5. Changes to BGP Finite State Machine

As mentioned under "Procedures for the Receiving Speaker" above, this specification modifies the BGP finite state machine.

The specific state machine modifications to [BGP-4], Section 8.2.2, are as follows.

In the Idle state, make the following changes.

Replace this text:

- initializes all BGP resources for the peer connection,

with

- initializes all BGP resources for the peer connection, other than those resources required in order to retain routes according to section "Procedures for the Receiving Speaker" of this (Graceful Restart) specification,

In the Established state, make the following changes.

Replace this text:

> In response to an indication that the TCP connection is successfully established (Event 16 or Event 17), the second connection SHALL be tracked until it sends an OPEN message.

with

If the Graceful Restart Capability with one or more AFIs/SAFIs has not been received for the session, then in response to an indication that a TCP connection is successfully established (Event 16 or Event 17), the second connection SHALL be tracked until it sends an OPEN message.

However, if the Graceful Restart Capability with one or more AFIs/SAFIs has been received for the session, then in response to Event 16 or Event 17 the local system:

- retains all routes associated with this connection according to section "Procedures for the Receiving Speaker" of this (Graceful Restart) specification,

- releases all other BGP resources,

- drops the TCP connection associated with the ESTABLISHED session,

- initializes all BGP resources for the peer connection, other than those required in order to retain routes according to section "Procedures for the Receiving Speaker" of this specification,

- sets ConnectRetryCounter to zero,

- starts the ConnectRetryTimer with the initial value, and

- changes its state to Connect.

Replace this text:

If the local system receives a NOTIFICATION message (Event 24 or Event 25), or a TcpConnectionFails (Event 18) from the underlying TCP, the local system:

- sets the ConnectRetryTimer to zero,

- deletes all routes associated with this connection,

- releases all the BGP resources,

- drops the TCP connection,

- increments the ConnectRetryCounter by 1,

- changes its state to Idle.

with

If the local system receives a NOTIFICATION message (Event 24 or Event 25), or if the local system receives a TcpConnectionFails (Event 18) from the underlying TCP and the Graceful Restart capability with one or more AFIs/SAFIs has not been received for the session, the local system:

- sets the ConnectRetryTimer to zero,
- deletes all routes associated with this connection,
- releases all the BGP resources,
- drops the TCP connection,
- increments the ConnectRetryCounter by 1, and
- changes its state to Idle.

However, if the local system receives a TcpConnectionFails (Event 18) from the underlying TCP, and the Graceful Restart Capability with one or more AFIs/SAFIs has been received for the session, the local system:

- sets the ConnectRetryTimer to zero,
- retains all routes associated with this connection according to section "Procedures for the Receiving Speaker" of this (Graceful Restart) specification,
- releases all other BGP resources,
- drops the TCP connection,
- increments the ConnectRetryCounter by 1, and
- changes its state to Idle.

6. Deployment Considerations

Although the procedures described in this document would help minimize the effect of routing flaps, it is noted that when a BGP Graceful Restart-capable router restarts, or if it restarts without preserving its forwarding state (e.g., due to a power failure), there is a potential for transient routing loops or blackholes in the network if routing information changes before the involved routers complete routing updates and convergence. Also, depending on the

network topology, if not all IBGP speakers are Graceful Restart capable, there could be an increased exposure to transient routing loops or blackholes when the Graceful Restart procedures are exercised.

The Restart Time, the upper bound for retaining routes, and the upper bound for deferring route selection may need to be tuned as more deployment experience is gained.

Finally, it is noted that the benefits of deploying BGP Graceful Restart in an Autonomous System (AS) whose IGPs and BGP are tightly coupled (i.e., BGP and IGPs would both restart) and IGPs have no similar Graceful Restart Capability are reduced relative to the scenario where IGPs do have similar Graceful Restart Capability.

7. Security Considerations

Since with this proposal a new connection can cause an old one to be terminated, it might seem to open the door to denial of service attacks. However, it is noted that unauthenticated BGP is already known to be vulnerable to denials of service through attacks on the TCP transport. The TCP transport is commonly protected through use of [BGP-AUTH]. Such authentication will equally protect against denials of service through spurious new connections.

If an attacker is able to successfully open a TCP connection impersonating a legitimate peer, the attacker’s connection will replace the legitimate one, potentially enabling the attacker to advertise bogus routes. We note, however, that the window for such a route insertion attack is small since through normal operation of the protocol the legitimate peer would open a new connection, in turn causing the attacker’s connection to be terminated. Thus, this attack devolves to a form of denial of service.

It is thus concluded that this proposal does not change the underlying security model (and issues) of BGP-4.

We also note that implementations may allow use of graceful restart to be controlled by configuration. If graceful restart is not enabled, naturally the underlying security model of BGP-4 is unchanged.

8. Acknowledgments

The authors would like to thank Bruce Cole, Lars Eggert, Bill Fenner, Eric Gray, Jeffrey Haas, Sam Hartman, Alvaro Retana, Pekka Savola Naiming Shen, Satinder Singh, Mark Townsley, David Ward, Shane Wright, and Alex Zinin for their review and comments.

9. IANA Considerations

This document defines a new BGP capability - Graceful Restart Capability. The Capability Code for Graceful Restart Capability is 64.

10. References

10.1. Normative References

[BGP-4] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[BGP-MP] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.

[BGP-CAP] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 3392, November 2002.

[BGP-AUTH] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

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

[IANA-AFI] http://www.iana.org/assignments/address-family-numbers

[IANA-SAFI] http://www.iana.org/assignments/safi-namespace

10.2. Informative References

[BFD] Katz, D. and D. Ward, "Bidirectional Forwarding Detection", Work in Progress.

Authors' Addresses

Srihari R. Sangli
Cisco Systems, Inc.

EMail: rsrihari@cisco.com

Yakov Rekhter
Juniper Networks, Inc.

EMail: yakov@juniper.net

Rex Fernando
Juniper Networks, Inc.

EMail: rex@juniper.net

John G. Scudder
Juniper Networks, Inc.

EMail: jgs@juniper.net

Enke Chen
Cisco Systems, Inc.

EMail: enkechen@cisco.com

Full Copyright Statement

Intellectual Property

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.