Network Working Group
Internet Draft
Expiration Date: December 2004

Srihari R. Sangli (Procket Networks)
Yakov Rekhter (Juniper Networks)
Rex Fernando (Procket Networks)
John G. Scudder (Cisco Systems)
Enke Chen (Redback Networks)

Graceful Restart Mechanism for BGP

draft-ietf-idr-restart-10.txt

1. Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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.''

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

2. Abstract

This document proposes 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 which 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 transport reset.

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).

3. 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 RFC2119 [RFC2119].

4. 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 flap 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 proposes 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 which 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 transport reset.

5. Marker for End-of-RIB

An UPDATE message with no reachable 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 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.

6. Graceful Restart Capability

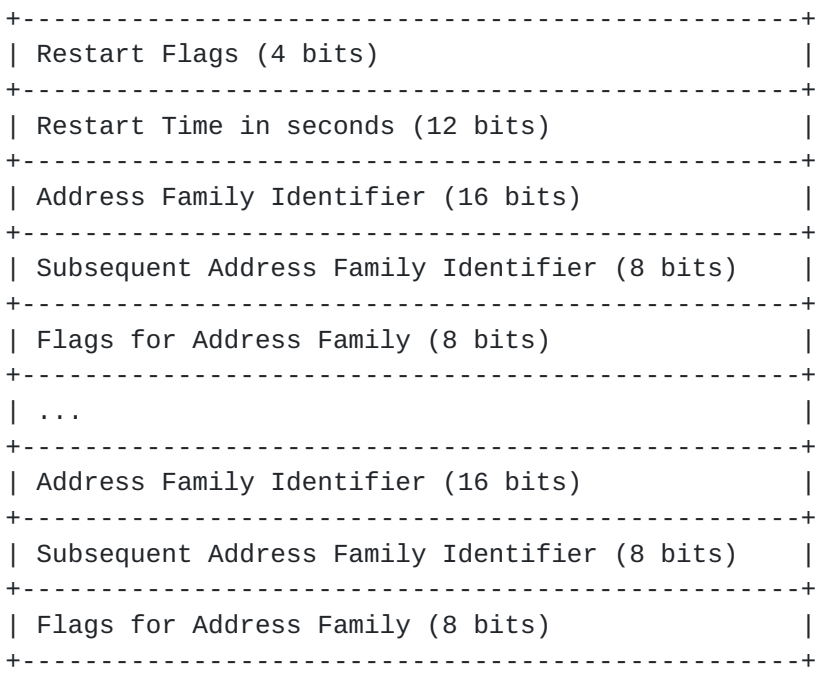
The Graceful Restart Capability is a new BGP capability [BGP-CAP] that can be used by a BGP speaker to indicate its ability to preserve its forwarding state during BGP restart. It can also be used to convey to its peer its intention of generating the End-Of-RIB marker upon the completion of its initial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

This capability is defined as follows:

Capability code: 64

Capability length: variable

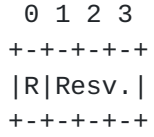
Capability value: Consists of the "Restart Flags" field, "Restart Time" field, and zero or more of the tuples <AFI, SAFI, Flags for address family> as follows:



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 SHOULD NOT wait for the End-of-RIB marker from the speaker before advertising routing information to the speaker.

The remaining bits are reserved, and SHOULD 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):

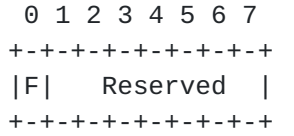
This field carries the identity of the Network Layer protocol for which the Graceful Restart support is advertised. Presently defined values for this field are specified in [IANA-AFI].

Subsequent Address Family Identifier (SAFI):

This field provides additional information about the type of the Network Layer Reachability Information carried in the attribute. Presently defined values for this field are specified in [IANA-SAFI].

Flags for Address Family:

This field contains bit flags for the <AFI, SAFI>.



The most significant bit is defined as the Forwarding State (F)

bit which can be used to indicate if the forwarding state for the <AFI, 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 SHOULD be set to zero by the sender and ignored by the receiver.

When a sender of this capability doesn't 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 6.2 of this document). In that case the value of the "Restart Time" field advertised by the sender is irrelevant.

A BGP speaker SHOULD 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 SHOULD ignore all but the last instance of the Graceful Restart Capability.

Including <AFI=IPv4, SAFI=unicast> into the Graceful Restart Capability doesn't 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.

7. 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. First, 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. Second, to indicate its support for a peer which wishes to perform a graceful restart.

The End-of-RIB marker SHOULD 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.

In general the Restart Time SHOULD NOT be greater than 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 SHALL be followed by the Restarting Speaker as well as the Receiving Speaker once the Restarting Speaker restarts.

7.1. Procedures for the Restarting Speaker

When the Restarting Speaker restarts, possible it SHOULD retain, if possible, the forwarding state for the BGP routes in the Loc-RIB, and SHALL mark them as stale. It SHOULD 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 SHALL defer route selection for an address family until it 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 which do not advertise the graceful restart capability). 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 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 SHALL be updated and any previously marked stale information SHALL 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 SHALL 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.

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.

7.2. 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 TCP reset and still considers the BGP session as being established, it SHALL treat the subsequent open connection from the peer as an indication of TCP reset 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 SHOULD 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 reset, no NOTIFICATION message should be sent -- the previous TCP session is simply closed.

When the Receiving Speaker detects TCP reset for a BGP session with a peer that has advertised the Graceful Restart Capability, it SHALL retain the routes received from the peer for all the address families that were previously received in the Graceful Restart Capability, and SHALL mark them as stale routing information. To deal with possible consecutive restarts, a route (from the peer) previously marked as stale SHALL be deleted. The router SHOULD 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 SHALL NOT be set unless the Receiving Speaker has restarted. The presence and the setting of the "Forwarding State" bit for an address family depends 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 SHALL 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 isn't received in the re-established session at all, then Receiving Speaker SHALL immediately remove all the stale routes from the peer that it is retaining for that address family.

The Receiving Speaker SHALL 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 SHALL 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 SHALL 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.

8. 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 Established state, replace this text:

> If a TcpConnection_Valid (Event 14) or Tcp_CR_Acked (Event 16) is received, or a TcpConnectionConfirmed event (Event 17) is received, a second TCP connection may be in progress. This second TCP connection is tracked per Connection Collision processing (Section 6.8) until an OPEN message is received.

with this:

If a TcpConnection_Valid (Event 14) or Tcp_CR_Acked (Event 16) is received, a second TCP connection may be in progress. This second TCP connection is tracked per Connection Collision processing (Section 6.8) until an OPEN message is received.

If the Graceful Restart capability with one or more AFI/SAFI has been received for the session, then TcpConnectionConfirmed (Event 17) is treated as TcpConnectionFails (Event 18).

If a TcpConnectionConfirmed event (Event 17) is received and if the Graceful Restart capability with one or more AFI/SAFI has not been received for the session, a second TCP connection may be in progress. This second TCP connection is tracked per Connection Collision processing (Section 6.8) until an OPEN message is received.

9. Deployment Considerations

While the procedures described in this document would help minimize the effect of routing flaps, it is noted, however, that when a BGP Graceful Restart capable router restarts, 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 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.

10. 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.

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

11. Acknowledgments

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

12. Normative References

[BGP-4] Rekhter, Y., T. Li, Hares, S., "A Border Gateway Protocol 4 (BGP- 4)", work in progress.

[BGP-MP] Bates, T., Chandra, R., Katz, D., and Rekhter, Y., "Multiprotocol Extensions for BGP-4", RFC2858, June 2000.

[BGP-CAP] Chandra, R., Scudder, J., "Capabilities Advertisement with BGP-4", draft-ietf-idr-rfc2842bis-02.txt, April 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.

13. Author Information

Srihari R. Sangli
Procket Networks, Inc.
1100 Cadillac Court
Milpitas, CA 95035
e-mail: srihari@procket.com

Yakov Rekhter
Juniper Networks, Inc.
1194 N. Mathilda Avenue
Sunnyvale, CA 94089
e-mail: yakov@juniper.net

Rex Fernando
Procket Networks, Inc.
1100 Cadillac Court
Milpitas, CA 95035
e-mail: rex@procket.com

John G. Scudder
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
e-mail: jgs@cisco.com

Enke Chen
Redback Networks, Inc.
350 Holger Way
San Jose, CA 95134
e-mail: enke@redback.com