Network Working Group
Internet Draft
Expiration Date: January 2002

Srihari Ramachandra (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-01.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.

3. 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.

4. Marker for End-of-RIB

An UPDATE message with 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 MP_UNREACH_NLRI [BGP-MP] with no withdrawn routes for that <AFI, Sub-AFI>.

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.

5. Graceful Restart Capability

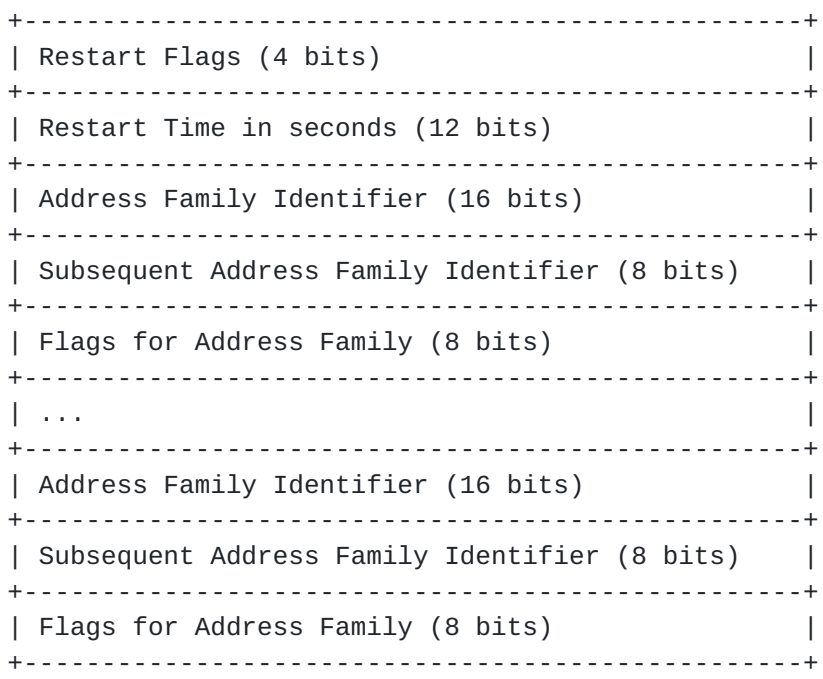
The Graceful Restart Capability is a new BGP capability [BGP-CAP] that can be used by a BGP speaker to indicate its ability to preserve its forwarding state during BGP restart. It can also be used to convey to its peer its intention of generating the End-Of-RIB marker upon the completion of its initial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

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 RFC1700 (see the Address Family Numbers section).

Subsequent Address Family Identifier (Sub-AFI):

This field provides additional information about the type of the Network Layer Reachability Information carried in the attribute.

Flags for Address Family:

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

The most significant bit is defined as the Forwarding State bit which can be used to indicate if the forwarding state for the <AFI, Sub-AFI> 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.

The advertisement of this capability by a BGP speaker also implies that it will generate the End-of-RIB marker upon completion of its initial routing update to its peer. The value of the "Restart Time" field is irrelevant in the case that the capability does not carry any <AFI, Sub-AFI>.

6. Operation

A BGP speaker may advertise the Graceful Restart Capability for an address family to its peer only if it has the ability to preserve its forwarding state for the address family when BGP restarts.

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 to indicate its intention of generating the End-of-RIB marker upon the completion of its initial routing updates.

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.

6.1. Procedures for the Restarting Speaker

When the Restarting Speaker restarts, if possible it shall retain 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). 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.

6.2. 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 Restarting Speaker as an indication of TCP reset and act accordingly.

When the TCP reset is detected by the Receiving Speaker, it shall retain the routes received from the Restarting Speaker 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 Restarting Speaker) 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 also 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 Restarting Speaker advertised previously, the Receiving Speaker shall delete all the stale routes from the Restarting Speaker that it is retaining.

Once the session is re-established, if the "Forwarding State" bit for an address family is not set in the received Graceful Restart Capability, or if the capability is not received for an address family, the Receiving Speaker shall immediately remove all the stale routes from the Restarting Speaker 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 Restarting Speaker.

The Receiving Speaker shall replace the stale routes by the routing updates received from the Restarting Speaker. Once the End-of-RIB marker for an address family is received from the Restarting Speaker, it shall immediately remove any routes from the Restarting Speaker 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.

7. 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 there is little benefit 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.

8. 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.

9. Acknowledgments

The authors would like to thank Alvaro Retana, Satinder Singh, David Ward, Naiming Shen and Bruce Cole for their review and comments.

10. References

[BGP-4] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[BGP-MP] Bates, T., Chandra, R., Katz, D., and Rekhter, Y., "Multiprotocol Extensions for BGP-4", RFC 2283, March 1998.

[BGP-CAP] Chandra, R., Scudder, J., "Capabilities Advertisement with BGP-4", RFC 2842, May 2000.

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

11. Author Information

Srihari Ramachandra
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