Network Working Group
Internet Draft
Intended Status: Standards Track
Expiration Date: Dec 30, 2011

K. Patel
E. Chen
R. Fernando
Cisco Systems
J. Scudder
Juniper Networks
June 29, 2011

Accelerated Routing Convergence for BGP Graceful Restart
draft-keyur-idr-enhanced-gr-00.txt

Status of this Memo

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

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/1id-abstracts.html

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

This Internet-Draft will expire on December 30, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as

described in the Simplified BSD License.

Abstract

In this document we specify extensions to BGP graceful restart in order to avoid unnecessary transmission of the routing information preserved across a session restart, thus accelerating the routing convergence.

1. Introduction

Currently the BGP graceful restart (GR) mechanism specified in [RFC4724] requires a complete re-advertisement of the routing information across a session restart, even though partial or complete routing information is usually preserved. For example, as described in [RFC4724], the "Receiving Speaker" temporarily maintains the routes received from its neighbor with the GR Capability. In addition, the "Restarting Speaker" may also be able to preserve partial or full routing information across a BGP restart by check-pointing routing information to a standby or secondary facility.

Clearly the routing re-convergence post a session restart would be faster if we can avoid unnecessary transmission of the routing information preserved across a session restart. That is the goal of this document.

In this document we specify extensions to BGP graceful restart in order to avoid unnecessary transmission of the routing information preserved across a session restart, thus accelerating the routing convergence. More specifically, we describe a "version number" based mechanism for keeping track of the routing information across a session restart. A new BGP message type, UPDATE-VERSION, is introduced for checkpointing the update version maintained for a neighbor. We also introduce the Enhanced Graceful Restart Capability, and specify procedures for handling routing update across a session restart.

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 [RFC2119].

2. Version Numbers for Routing Entities

In order to avoid unnecessary transmission of the routing information preserved across a session restart, a BGP speaker will need to identify exactly "what" has been preserved by a remote speaker.

The approach described here is "version number" (or "sequence number") based, and it consists of (a) assigning a unique, monotonically increasing number as the version number for each routing entity (e.g., route or message) when it is created or modified; and (b) maintaining an update version (for each neighbor) calculated as the maximum of the version numbers of all the routing entities that have been sent to the neighbor.

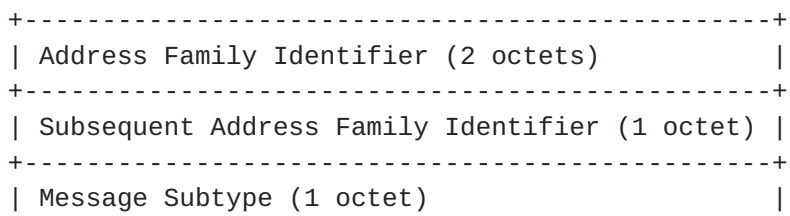
A BGP speaker can tell whether a given routing entity has been sent to a neighbor by comparing the version number of the entity with the update version for the neighbor. Thus by checkpointing the update version for a neighbor across a session restart, a BGP speaker would be able to identify exactly "what" has been preserved by a remote speaker, and also "what" remains to be sent.

In this document a version number is a 8-octet unsigned integer. Value 0 is used to indicate the beginning (or "epoch") of the update generation. The version number is not expected to wrap. However, in the unlikely scenario that it does wrap, the sender MUST maintain its internal consistency, and also MUST perform a route refresh [RFC2918, EH-RR] toward the receiver.

The number space for the version numbers should be AFI/SAFI [RFC4760] specific. Version numbers are also assigned (from the same number space) to other AFI/SAFI specific, non-update information (such as ROUTE-REFRESH [RFC2918]), and are included in the calculation of the update version for a neighbor.

3. UPDATE-VERSION Message

The UPDATE-VERSION message is a new BGP message type with type code <TBD>. In addition to the fixed-size BGP header [RFC4271], the UPDATE-VERSION message contains the following fields:



The "Address Family Identifier" (AFI) field and the "Subsequent Address Family Identifier" (SAFI) field are the same as the ones used in [RFC4760].

The "Message Subtype" field indicates whether the sender is (a) sending an update version (value 1), (b) acknowledging the receipt of an update version (value 2), or (c) requesting updates from the very last update version the sender has acknowledged (value 3).

The Version field contains an update version associated with the message subtypes 1 and 2. The value of this field is irrelevant for the message subtype 3. This value of the field is opaque to the receiver.

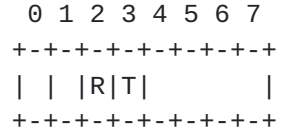
As detailed in the Operation section, the UPDATE-VERSION message can be used by a BGP speaker to either carry an update version, or acknowledge the receipt of an update version, or request updates from the very last update version acknowledged.

4. Enhanced Graceful Restart Capability

The Enhanced Graceful Restart (GR) Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is specified in the IANA Considerations section of this document. The Capability Length field of this capability is 0.

By advertising the Enhanced GR Capability to a peer, a BGP speaker conveys to the peer that the speaker is capable of receiving and properly handling the UPDATE-VERSION message from the peer, as well as recognizing the two new bit flags defined below for the GR Capability.

The two new bit flags for the "Flags for Address Family" field of the GR Capability are defined as follows:



The third most significant bit (R) is defined as the "RX Routing State", which is used to indicate whether during the previous session restart the routes of the given AFI/SAFI that were received have indeed been preserved up to the update version acknowledged by the speaker previously. When set (value 1), the bit indicates that the routes have been preserved.

The fourth most significant bit (T) is defined as the "TX Routing State", which is used to indicate whether the speaker has indeed preserved enough state to resume advertising routes of the given AFI/SAFI from the update version acknowledged by the neighbor previously. When set (value 1), the bit indicates that the state has been preserved.

5. Operation

In order for a BGP speaker to be able to resume sending routing information for an AFI/SAFI from the last update version that was previously acknowledged by a peer, the speaker MUST maintain enough state for all the routing information that has been sent until their acknowledgment is received by the speaker. The routing information includes reachable / unreachable information as well as other AFI/SAFI specific, non-update information. Furthermore, the route advertisement state needs to be maintained properly in order to minimize spurious route withdraws across a session restart.

An implementation SHOULD impose an upper bound on how much state it would maintain in the case that a receiver ("slow peer") is not able to generate an acknowledgment in a timely manner. The upper bound might be based on a number of factors such as the number of pending unacknowledged withdraws or more generally, the volume of unacknowledged state, and a timer. Once the acknowledgment from a peer is not received within the specified upper bound, and the maintained state is compromised, then the speaker MUST clear the "TX Routing State" in the GR Capability to be advertised to the peer in the next session restart.

A BGP speaker MAY advertise the Enhanced GR Capability to its peer if the speaker is capable of receiving and properly handling the UPDATE-VERSION message from the peer, and also recognizing the two new bit flags in the GR Capability. If the GR Capability is to be sent by the speaker, the "RX Routing State" for an AFI/SAFI in the GR Capability SHOULD be set if the speaker has preserved the routing information from the peer up to the update version that the speaker acknowledged previously. In addition, the "TX Routing State" for an AFI/SAFI in the GR Capability SHOULD be set if the speaker has preserved enough routing state to resume sending messages from the

update version acknowledged by the peer previously.

When both the GR Capability and the Enhanced GR Capability are to be included in an OPEN message, it is RECOMMENDED (though not required) that the Enhanced GR Capability be placed ahead of the GR Capability.

In processing the GR Capability in an OPEN message from a peer, a BGP speaker MUST NOT examine the two new bit flags defined in this document for the GR Capability unless the Enhanced GR Capability is also present in the OPEN message.

A BGP speaker MAY send an UPDATE-VERSION message to a peer only if the Enhanced GR Capability is received from the peer.

Once a BGP speaker receives the Enhanced GR Capability from its peer, the speaker SHOULD send an UPDATE-VERSION message carrying the update version after sending significant amount of routing information (including non-UPDATE messages) for an AFI/SAFI. This SHALL continue as long as routing information is being sent. To reduce the overhead by excessive number of UPDATE-VERSION messages, we highly recommend the "batching" approach, that is, use one UPDATE-VERSION message to cover a number of routing updates, and/or a meaningful duration of time.

When a BGP speaker receives an UPDATE-VERSION message carrying an update version, if the AFI/SAFI carried by the message does not match any AFI/SAFI that the speaker is willing to receive from the peer, the UPDATE-VERSION message SHALL be ignored. Otherwise, the speaker MUST send an UPDATE-VERSION message back promptly acknowledging the receipt of the update version. The UPDATE-VERSION messages carrying the acknowledgments MUST be sent in the same order as the received UPDATE-VERSION messages carrying the update versions.

When a BGP speakers receives an UPDATE-VERSION message acknowledging an update version, the speaker MUST record this latest update version being acknowledged for future use.

Consider the case that both the GR Capability and the Enhanced GR Capability are exchanged between Speaker A and Speaker B, and for an AFI/SAFI the "TX Routing State" is set in the GR advertised by A, and the "RX Routing State" is also set in the GR received from B. Then Speaker A SHALL send routing information from the last update version that was previously acknowledged by Speaker B. Note that it may be advantageous for Speaker B to send an UPDATE-VERSION message acknowledging the most recent update version immediately after the session is established. Also, Speaker B MUST not follow the procedures described in [RFC4724] for purging stale routes. If the conditions specified in this paragraph are not satisfied, then the

procedures described in [RFC4724] remain unchanged.

During the lifetime of an established session, if needed, a BGP speaker MAY use the UPDATE-VERSION message to request updates from the last update version that was previously acknowledged as long as the speaker has received the Enhanced GR Capability from its peer.

When a BGP speaker receives such a request, it SHALL try to send routing information from the last acknowledged update version that the speaker has recorded. If the speaker is unable to do so for some reason (e.g., "slow peer"), then it SHOULD perform a route refresh using mechanism defined in [EH-RR] if possible. Otherwise, the BGP speaker SHOULD reset the session.

6. Error Handling

This document defines a new NOTIFICATION error code:

Error Code	Symbolic Name
TBD	UPDATE-VERSION Message Error

The following error subcodes are defined as well:

Subcode	Symbolic Name
1	Invalid Message Length
2	Invalid Message Subtype

If a BGP speaker detects an error while processing an UPDATE-VERSION message, it MUST send a NOTIFICATION message with Error Code UPDATE-VERSION Message Error. The Data field of the NOTIFICATION message MUST contain the complete UPDATE-VERSION message.

If the Length field for the UPDATE-VERSION message is incorrect, then the error subcode is set to "Invalid Message Length".

If the Message Subtype in the UPDATE-VERSION message is not any of the defined value, then the error subcode is set to "Invalid Message Subtype".

7. IANA Considerations

This document introduces the Enhanced Graceful Restart Capability. The capability code needs to be assigned by IANA per [RFC5492].

This document introduce a new BGP message type, UPDATE-VERSION. The type code needs to be assigned by IANA.

In addition, this document defines an NOTIFICATION error code and several error subcodes for the UPDATE-VERSION message. They need to be registered with the IANA.

8. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP [RFC4271, RFC4724].

9. Acknowledgments

TBD.

10. Normative References

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

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4724] Sangli, S., E. Chen, R. Rernando, J. Scudder, and Y. Rekhter, "Graceful Restart Mechanism for BGP", January 2007

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

[EH-RR] Patel, K., E. Chen and B. Venkatachalapathy, "Enhanced

Route Refresh Capability for BGP-4", work in progress.

11. Authors' Addresses

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

Enke Chen
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134
USA

EMail: enkechen@cisco.com

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

Email: rex@cisco.com

John Scudder
Juniper Networks

Email: jgs@juniper.net