Network Working Group H. Zhang
Internet-Draft HangZhou H3C Co. Limited
Updates: RFC 4724 (if approved) A. Retana
Intended status: Standards Track Cisco Systems, Inc.
Expires: September 26, 2013 March 25, 2013

Transitive BGP Graceful Restart
draft-zhang-idr-transitive-gr-02

Abstract

This document defines an extension to BGP Graceful Restart that reduces the negative impact of multiple inter-connected routers restarting. The proposed mechanism does not require any changes to the BGP protocol.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

The BGP Graceful Restart [RFC4724] process defines a mechanism that a restarting router can use with its non-restarting peers. The existence of other restarting routers results in the use of the base route exchange mechanism [RFC4271] with them, even if the forwarding state has indeed been preserved for (and by) those peers during the restart. As a result, traffic forwarding between restarting routers is disrupted.

This document defines an extension to BGP Graceful Restart that reduces the negative impact of multiple inter-connected restarting routers. The proposed mechanism does not require any changes to the BGP protocol.

The current process [RFC4724] states that routes from restarting peers are to be removed from the local forwarding state when the non-restarting peers converge (the End-of-RIB marker is received from all of them). Assuming a simple topology:

```
NR1 - R2 - R3 - NR2
```

where NRx are non-restarting routers, Rx are restarting routers and the lines between them represent BGP sessions.

There are two types of routes affected (from R2's point of view) by the current process:

1. Routes that are only reachable through R3. These routes will be removed from the forwarding table when the non-restarting routers converge, and installed back in when the convergence with R3 is done.

2. Routes that are reachable through both R3 and NR1. These routes will first change to NR1 when the non-restarting routers converge, and later back to R3 (assuming that is in fact still the preferred path).

Both types can clearly cause disruption in traffic forwarding, micro-loops, traffic loss, etc.

2. Requirements Language

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].

3. Proposed Solution

The extension proposed to BGP Graceful Restart to accommodate for multiple restarting routers, when the forwarding state has been preserved between them, is simply to delay sending the End-of-RIB marker to non-restarting routers.

Specifically, to allow a restarting router the ability to reduce the impact due to other restarting routers, the following paragraph is added as the fifth one in section 4.1 (Procedures for the Restarting Speaker) [RFC4724]:

> Before updating the corresponding forwarding states, the Restarting Speaker MAY start a path calculation after all non-retarting peers's End-Of-RIB marker have been received, and advertise the Adj-RIB-Out to its restarting peers (ones with the "Restart State" bit set in the received capability), including the End-of-RIB marker, and wait for the corresponding End-of-RIB marker from them.

In order to maintain the transitive property when more than two BGP speakers peering with each other restart, the following paragraph is added as the sixth one in section 4.1 (Procedures for the Restarting Speaker) [RFC4724]:

> If the Restarting Speaker has multiple restarting peers, sending the End-of-RIB marker SHOULD be delayed until all the markers from those restarting peers have been received. The BGP speaker on a given connection SHOULD send its End-of-RIB marker if the pair hasn't sent or received UPDATES for a locally configured time period (which SHOULD be significantly less than the Selection_Deferral_Timer).

During the recovery period of multiple restarting routers, a BGP speaker may advertise routing information that is not being used at the time. Because the forwarding state of the speakers remains unchanged (from that at the restart), it is clear that this transitive property of sharing routing information between restarting routers doesn't cause any issues in the actual forwarding of traffic. Furthermore, it has the advantage if avoiding further disruptions in the forwarding of traffic through the restarting routers.

4. Security Considerations

This document proposes an extension to an existing mechanism. The same security considerations explained there apply to this extension.

The propagation of routing information that is not in use may cause forwarding loops and an inconsistent state in a network. However, the risk in this document is mitigated by the fact that the information is validated by all peers once the convergence process completes.

5. IANA Considerations

This document has no IANA actions.

6. Acknowledgements

The authors would like to thank Enke Chen, John Scudder, Robert Raszuk and Abhay Roy for their feedback.

7. References

7.1. Normative References

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

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.

7.2. Informative References

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

Authors' Addresses

Haifeng Zhang
HangZhou H3C Co. Limited
310 Liuhe Road, Zhijiang Science Park
Hangzhou
P.R. China

Email: zhanghf@h3c.com

Alvaro Retana
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
USA

Email: aretana@cisco.com