IDR Working Group
Internet-Draft
Intended status: Informational
Expires: September 24, 2021

A. Bhagat
Amazon
March 23, 2021

BGP Multiple Nexthops
draft-bhagat-bgp-multiple-nexthops-00

Abstract

This document presents a new feature in BGP that allows grouping of multiple BGP sessions between a pair of speakers and sending multiple nexthops for a single prefix. This helps avoid sending and receiving duplicate routes across all sessions.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

In Data Center networks where CLOS fabrics are built solely using BGP [RFC4271], it is very common to have topology where a pair of routers have multiple BGP sessions between them - a single BGP session over every link. Each BGP session is independent of the others and BGP messages are sent and received over every BGP session. There are various reasons for following this design pattern but the main reason is that when links within the LAG interfaces go down, that results in inconsistent bandwidth availability which is not reflected at the routing level. This causes the capacity models to not work correctly and can also result in network congestion.

While the maintenance of these independent BGP sessions is trivial, routers sending and receiving duplicate BGP UPDATE messages for hundreds or thousands of routes, leads to unnecessarily generating, processing and storing of routes. These duplicate messages provide no extra information except capability to select and install multiple paths for routes. Every route in the BGP UPDATE messages has same BGP path attributes except the NEXT_HOP attribute.

This document provides a way to advertise the route only one time with multiple NEXT_HOP attributes to achieve the same benefits as having the same route advertised multiple times over multiple BGP sessions with different NEXT_HOP attributes.

2. Capability Support

A new Capability Optional parameter will be communicated in BGP Open message. A BGP speaker SHOULD use Capability Advertisement procedure in [RFC3392] to announce the support. The Capability Code is to be assigned by IANA.

```
       0                   1                   2                   3
       0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |     Type      |    Length     |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |      AFI                      |  Reserved     |  SAFI         |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       ~                                                               ~
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |      AFI                      |  Reserved     |  SAFI         |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: BGP Multiple Nexthops Capability

Capability Type: TBA by IANA

Capability Length: Variable length.

Capability Value: Specifies all AFI/SAFI configured on the BGP speaker that support the feature.

3. Operation

During BGP session establishment, BGP Multiple Nexthops capability for every supported AFI/SAFI is advertised and received in BGP Open message. When two BGP speakers have multiple BGP sessions between themselves and if they support BGP Multiple Nexthops capability, BGP Identifier and AS of the peer are used to identify all BGP sessions that can be logically grouped together per AFI/SAFI. BGP UPDATE messages sent over one BGP session applies to all other BGP sessions within this logical group. The BGP peers MUST share same configuration settings to be treated as a group on the speaker.

When BGP UPDATE message is advertised, the rules for the next hop information are as follows:

When sending a message to an external peer:

o The BGP speaker SHOULD add multiple NEXT_HOP attributes - each NEXT_HOP attribute carrying the IP address of the interface that the speaker uses to establish BGP session to peer.

When sending a message to an internal peer:

o If the route is not locally originated, the BGP speaker SHOULD NOT modify the NEXT_HOP attributes unless it has been explicitly configured to announce its own IP address(es) as next-hop(s).

- o If the route is locally originated, the BGP speaker SHOULD add multiple NEXT_HOP attributes - each NEXT_HOP attribute carrying the IP address of the interface that the speaker uses to establish BGP session to peer.

When withdrawing routes, next-hop information is not carried in the message. In that case, the peer SHOULD remove the route with any number of NEXT_HOP attributes attached to it even when the withdraw message is received over a different BGP session than the original BGP session over which the update message was sent.

When the link or BGP session associated with the logical group goes down, the speakers SHOULD remove only the NEXT_HOP associated with routes.

Note that the BGP UPDATE message is sent over a single BGP session in the logical group. For example, if there are 8 independent BGP sessions between two speakers, the speaker chooses only 1 out of 8 sessions over which it sends the BGP UPDATE message. The speaker can choose one BGP session at random, or in round-robin fashion, or some other means and hence is out-of-scope of this document.

4. Multiprotocol Extensions

[RFC4760] defines MP_REACH_NLRI path attribute which carries routes as well as next-hop information, grouped together. Details of next-hop information for MP_REACH_NLRI in section 3 of [RFC4760]. This document allows adding multiple NEXT_HOP attributes when advertising routes with MP_REACH_NLRI path attribute using the same mechanism described in section 3 of this document.

5. IANA Considerations

As specified in the document, the IANA will assign a new Capability Code for BGP Multiple Nexthops capability support.

6. Acknowledgements

The authors would like to thank members of IDR Working Group for their review and comments.

7. Normative References

[RFC3392] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 3392, DOI 10.17487/RFC3392, November 2002, <https://www.rfc-editor.org/info/rfc3392>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

Author's Address

Amit Bhagat
Amazon
Seattle
USA

Email: abhagat@amazon.com