Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 9, 2018

X. Xu
Alibaba Inc.
K. Bi
Huawei
J. Tantsura
Nuage Networks
N. Triantafillis
Linked-in
K. Talaulikar
Cisco

April 7, 2018

BGP Neighbor Autodiscovery
draft-xu-idr-neighbor-autodiscovery-04

Abstract

BGP has been used as the underlay routing protocol in many hyper-scale data centers. This document proposes a BGP neighbor autodiscovery mechanism that greatly simplifies BGP deployments. This mechanism is very useful for those hyper-scale data centers where BGP is used as the underlay routing protocol.

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 RFC 2119 [RFC2119].

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

BGP has been used as the underlay routing protocol instead of IGP in many hyper-scale data centers [RFC7938]. Furthermore, there is an ongoing effort to leverage BGP link-state distribution mechanism to achieve BGP-SPF [I-D.keyupate-lsvr-bgp-spf]. However, BGP is not good as an IGP from the perspective of deployment automation and simplicity. For instance, the IP address and the Autonomous System Number (ASN) of each and every BGP neighbor have to be manually configured on BGP routers although these BGP peers are directly connected. In addition, for those directly connected BGP routers, it's usually not ideal to establish BGP sessions over their directly connected interface addresses due to the following reasons: 1) it's not convient to do trouble-shooting; 2) the BGP update volume is unnecessarily increased when there are multiple physical links

between them and those links couldn't be configured as a Link Aggregtion Group (LAG) due to whatever reason (e.g., diffferent link type or speed). As a result, it's more common that loopback interface addresses of those directly connected BGP peers are used for BGP session establishment. To make those loopback addresses of directly connected BGP peers reachable from one another, either static routes have to be configured or some kind of IGP has to be enabled. The former is not good from the automation perspective while the latter is in conflict with the original intention of using BGP as an IGP.

This draft specifies a BGP neighbor autodiscovery mechanism by borrowing some ideas from the Label Distribution Protocol (LDP) [RFC5036] . More specifically, directly connected BGP routers could automatically discovery the loopback address and the ASN of one other through the exchange of the to-be-defined BGP messages. The BGP session establishment process as defined in [RFC4271] could be triggered once directly connected BGP neighbors are discovered from one another. Note that the BGP session should be established over the discovered loopback address of the BGP neighbor. In addition, to elimnate the need of configuring static routes or enabling IGP for the loopback addresses, a certain type of routes towards the BGP neighbor's loopback addresses are dynatically instantiated once the BGP neighbor has been discovered. The administritive distance of such type of routes MUST be smaller than their equivalents that are learnt by the regular BGP update messages . Otherwise, circular dependency would occur once these loopback addresses are advertised via the regular BGP updates.

2. Terminology

This memo makes use of the terms defined in [RFC4271].

3. BGP Hello Message Format

To automatically discover directly connected BGP neighbors, a BGP router periodically sends BGP HELLO messages out those interfaces on which BGP neighbor autodiscovery are enabled. The BGP HELLO message is a new BGP message which has the same fixed-size BGP header as the exiting BGP messages. However, the HELLO message MUST sent as UDP packets addressed to the to-be-assigned BGP discovery port (179 is the suggested port value) for the "all routers on this subnet" group multicast address (i.e., 224.0.0.2 in the IPv4 case and FF02::2 in the IPv6 case). The IP source address is set to the address of the interface over which the message is sent out.

In addition to the fixed-size BGP header, the HELLO message contains the following fields:

```
       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
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |     Version   |   Hold Time   |      Message Length           |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |                           AS number                           |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |                             TLVs                              |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: BGP Hello Message

Version: This 1-octet unsigned integer indicates the protocol version number of the message. The current BGP version number is 4.

Hold Time: Hello hold timer in seconds. Hello Hold Time specifies the time the sending BGP peer will maintain its record of Hellos from the receiving BGP peer without receipt of another Hello. A pair of BGP peers negotiates the hold times they use for Hellos from each other. Each proposes a hold time. The hold time used is the minimum of the hold times proposed in their Hellos. A value of 0 means use the default 15 seconds.

Message Length: This 2-octet unsigned integer specifies the length in octects of the Connection Address TLV and other TLVs.

AS number: AS Number of the Hello message sender.

TLVs: This field contains Connection Address TLV and other TLVs.

The Accepted ASN List TLV format is shown as follows:

```
        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=TBD1            |      Length                   |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                 Accepted ASN List(variable)                   |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Accepted ASN List TLV

Type: TBD1

Length:Specifies the length of the Value field in octets.

Accepted ASN-List: This variable-length field contains one or more accepted 4-octet ASNs.

The Connection Address TLV format is shown as follows:

```
     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=TBD2            |      Length                   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |      Connection Address (4-octet or 16-octet)                 |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: Connection Address TLV

Type: TBD2

Length:Specifies the length of the Value field in octets.

Connection Address: This variable-length field indicates the IPv4 or IPv6 loopback address which is used for establishing BGP sessions.

The Router ID TLV format is shown as follows:

```
     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=TBD3            |      Length                   |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |               Router ID (4-octet or 16-octet)                 |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Router ID TLV

Type: TBD3

Length:Specifies the length of the Value field in octets and it's set to 4 for the IPv4-address-formatted BGP Router ID.

Router ID: This variable-length field indicates the BGP router ID which could be used for performing the BGP-SPF algorithm as described in [I-D.keyupate-lsvr-bgp-spf].

4. Hello Message Procedure

A BGP peer receiving Hellos from another peer maintains a Hello adjacency corresponding to the Hellos. The peer maintains a hold timer with the Hello adjacency, which it restarts whenever it receives a Hello that matches the Hello adjacency. If the hold timer for a Hello adjacency expires the peer discards the Hello adjacency.

We recommend that the interval between Hello transmissions be at most one third of the Hello hold time.

A BGP session with a peer has one or more Hello adjacencies.

A BGP session has multiple Hello adjacencies when a pair of BGP peers is connected by multiple links that have the same connection address (e.g., multiple PPP links between a pair of routers). In this situation, the Hellos a BGP peer sends on each such link carry the same Connection Address. In addition, to elimnate the need of configuring static routes or enabling IGP for advertising the loopback addresses, a certain type of routes towards the BGP neighbor's loopback addresses (e.g., carried in the Connection Address TLV) could be dymatically created once the BGP neighbor has been discovered. The administritive distance of such type of routes MUST be smaller than their equivalents which are learnt via the normal BGP update messages. Otherwise, circular dependency problem would occur once these loopback addresses are advertised via the normal BGP update messages as well.

BGP uses the regular receipt of BGP Hellos to indicate a peer's intent to keep BGP session identified by the Hello. A BGP peer maintains a hold timer with each Hello adjacency that it restarts when it receives a Hello that matches the adjacency. If the timer expires without receipt of a matching Hello from the peer, BGP concludes that the peer no longer wishes to keep BGP session for that link or that the peer has failed. The BGP peer then deletes the Hello adjacency. When the last Hello adjacency for an BGP session is deleted, the BGP peer terminates the BGP session by sending a Notification message and closing the transport connection. Meanwhile, the routes towards the BGP neighbor's loopback addresses that had been dynamically created due to the BGP Hello adjacency SHOULD be deleted accordingly.

5. HELLO Message Error Handling

TBD

6. Contributors

Satya Mohanty
Cisco
Email: satyamoh@cisco.com

7. Acknowledgements

The authors would like to thank Enke Chen for his valuable comments and suggestions on this document.

8. IANA Considerations

8.1. BGP Hello Message

This document requests IANA to allocate a new UDP port for BGP Hello message.

Value	TLV Name	Reference

Service Name: BGP-HELLO
Transport Protocol(s): UDP
Assignee: IESG <iesg@ietf.org>
Contact: IETF Chair <chair@ietf.org>.
Description: BGP Hello Message.
Reference: This document -- draft-xu-idr-neighbor-autodiscovery.
Port Number: TBD1 (179 is the suggested value) -- To be assigned by IANA.

8.2. TLVs of BGP Hello Message

This document requests IANA to create a new registry "TLVs of BGP Hello Message" with the following registration procedure:

Registry Name: TLVs of BGP Hello Message.

Value	TLV Name	Reference
0	Reserved	This document
1	Accepted ASN List	This document
2	Connection Address	This document
3	Router ID	This document
4-65500	Unassigned	
65501-65534	Experimental	This document
65535	Reserved	This document

9. Security Considerations

For security purposes, BGP speakers usually only accept TCP connection attempts to port 179 from the specified BGP peers or those within the configured address range. With the BGP auto-discovery mechanism, it's configurable to enable or disable sending/receiving BGP hello messages on the per-interface basis and BGP hello messages are only exchanged between physically connected peers that are

trustworthy. Therefore, the BGP auto-discovery mechanism doesn't introduce additional security risks associated with BGP.

In addition, for the BGP sessions with the automatically discovered peers via the BGP hello messages, the TTL of the TCP/BGP messages (dest port=179) MUST be set to 255. Any received TCP/BGP message with TTL being less than 254 MUST be dropped according to [RFC5082].

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[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>.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <https://www.rfc-editor.org/info/rfc5036>.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, DOI 10.17487/RFC5082, October 2007, <https://www.rfc-editor.org/info/rfc5082>.

[RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <https://www.rfc-editor.org/info/rfc8279>.

10.2. Informative References

[I-D.keyupate-lsvr-bgp-spf] Patel, K., Lindem, A., Zandi, S., and W. Henderickx, "Shortest Path Routing Extensions for BGP Protocol", draft-keyupate-lsvr-bgp-spf-00 (work in progress), March 2018.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

Authors' Addresses

Xiaohu Xu
Alibaba Inc.

Email: xiaohu.xxh@alibaba-inc.com

Kunyang Bi
Huawei

Email: bikunyang@huawei.com

Jeff Tantsura
Nuage Networks

Email: jefftant.ietf@gmail.com

Nikos Triantafillis
Linked-in

Fax: Nikos Triantafillis<nikos@linkedin.com>

Ketan Talaulikar
Cisco

Email: ketant@cisco.com