BESS WorkGroup
Internet-Draft
Intended status: Informational
Expires: September 16, 2021

S. Mohanty
Cisco Systems
A. Vayner
Google
A. Gattani
A. Kini
Arista Networks
March 15, 2021

Cumulative DMZ Link Bandwidth and load-balancing
draft-mohanty-bess-ebgp-dmz-03

Abstract

The DMZ Link Bandwidth draft provides a way to load-balance traffic to a destination (which is in a different AS than the source) which is reachable via more than one path. Typically, the link bandwidth (either configured on the link of the EBGP egress interface or set via a policy) is encoded in an extended community and then sent to the IBGP peer which employs multi-path. The link-bandwidth value is then extracted from the path extended community and is used as a weight in the FIB, which does the load-balancing. This draft extends the usage of the DMZ link bandwidth to another setting where the ingress BGP speaker requires knowledge of the cumulative bandwidth while doing the load-balancing. The draft also proposes neighbor-level knobs to enable the link bandwidth extended community to be regenerated and then advertised to EBGP peers to override the default behavior of not advertising optional non-transitive attributes to EBGP peers.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The Demilitarized Zone (DMZ) Link Bandwidth (LB) extended community along with the multi-path feature can be used to provide unequal cost load-balancing as per user control. In [I-D.ietf-idr-link-bandwidth] the EBGP egress link bandwidth is encoded in the link bandwidth extended community and sent along with the BGP update to the IBGP peer. It is assumed that either a labeled path exists to each of the EBGP links or alternatively the IGP cost to each link is the same. When the same prefix/net is advertised into the receiving AS via different egress-points or next-hops, the receiving IBGP peer that employs multi-path will use the value of the DMZ LB to load-balance traffic to the egress BGP speakers (ASBRs) in the proportion of the link-bandwidths.

The link bandwidth extended community cannot be advertised over EBGP peers as it is defined to be optional non-transitive. This draft

discusses a new use-case where we need to advertise the link bandwidth over EBGP peers. The new use-case mandates that the router calculates the aggregated link-bandwidth, regenerate the DMZ link bandwidth extended community, and advertise it to EBGP peers. The new use case also negates the [I-D.ietf-idr-link-bandwidth] restriction that the DMZ link bandwidth extended community not be sent when the the advertising router sets the next-hop to itself.

In draft [I-D.ietf-idr-link-bandwidth], the DMZ link bandwidth advertised by EBGP egress BGP speaker to the IBGP BGP speaker represents the Link Bandwidth of the EBGP link. However, sometimes there is a need to aggregate the link bandwidth of all the paths that are advertising a given net and then send it to an upstream neighbor. This is represented pictorially in Figure 1. The aggregated link bandwidth is used by the upstream router to do load-balancing as it may also receive several such paths for the same net which in turn carry the accumulated bandwidth.

```
R1- -20 - - |
            R3- -100 - -|
R2- -10 - - |           |
                        |
R6- -40 - - |           |- - R4
            |           |
            R5- -100 - -|
R7- -30 - - |
```

EBGP Network with cumulative DMZ requirement

Figure 1

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. Problem Description

Figure 1 above represents an all-EBGP network. Router R3 is peering with two other EBGP downstream routers, R1 and R2, over the eBGP link and another upstream EBGP router R4. There is another router, R5,

which is peering with two downstream routers R6 and R7. R5 peers with R4. A net, p/m, is learnt by R1, R2, R6, and R7 from their downstream routers (not shown). From the perspective of R4, the topology looks like a directed tree. The link bandwidths of the EBGP links are shown alongside the links (The exact units are not really important and for simplicity these can be assumed to be weights proportional to the operational link bandwidths). It is assumed that R3, R4 and R5 have multi-path configured and paths having different value as-path attributes can still be considered as multi-path (knobs exist in many implementations for this). When the ingress router, R4, sends traffic to the destination p/m, the traffic needs to be spread amongst the links in the ratio of their link bandwidths. Today this is not possible as there is no way to signal the link bandwidth extended community over the EBGP session from R3 to R4. In absence of a mechanism to regenerate the link bandwidth over the EBGP session from R3 to R4 and from R5 to R4, the assumed link bandwidth for paths received over the R3 to R4 and R5 to R4 EBGP sessions would be equal to the operational link bandwidth of the corresponding EBGP links.

As per EBGP rules at the advertising router, the next-hop will be set to the advertising router itself. Accordingly, R3 computes the best-path from the advertisements received from R1 and R2 and R5 computes the best-path from advertisements received from R6 and R7 respectively. R4 receives the update from R3 and R5 and in-turn computes the best-path and may advertises it upstream (not shown). The expected behavior is that when R4 sends traffic for p/m towards R3 and R5, and then on to to R1, R2, R6, and R7, the traffic should be load-balanced based on the calculated weights at the routers which employ multi-path. R4 should send 30% of the traffic to R3 and the remaining 70% to R5. R3 in turn should send 67% of the traffic that it received from R4 to R1 and 33% to R2. Similarly, R5 should send 57% of the traffic received from R4 to R6 and the remaining 43% to R7. Instead what is happening is that R4 sends 50% of the traffic towards both R3 and R5. R3 in turn sends more traffic than is desired towards R1 and R2. R4 in turn sends less traffic than is desired towards R6 and R7. Effectively the load balancing is getting skewed towards R1 and R2 even as R1 and R2's egress link bandwidth relative to R6 and R7 is less.

```
   R1- -20 - - |
               R3- -30 (100) - -|
   R2- -10 - - |                |
                                |
   R6- -40 - - |                |- - R4
               |                |
               R5- -70 (100) - -|
   R7- -30 - - |
```

EBGP Network showing advertisement of cumulative link bandwidth

Figure 2

With the existing rules for the DMZ link bandwidth, this is not possible. First the LB extended community is not sent over EBGP. Secondly the DMZ does not have a notion of conveying the cumulative link bandwidth (of the directed tree rooted at a node) to an upstream router. To enable the use case described above, the cumulative link bandwidth of R1 and R2 has to be advertised by R3 to R4, and, similarly, the cumulative bandwidth of R6 and R7 has to be advertised by R5 to R4. This will enable R4 to load-balance based on the proportion of the cumulative link bandwidth that it receives from its downstream routers R3 and R5. Figure 2 shows the cumulative link bandwidth advertised by R3 towards R4 and R5 towards R4 with the original link bandwidth values in '()' for comparison.

To address cases like the above example, rather than introducing a new attribute for aggregate link bandwidth, we will reuse the link bandwidth extended community attribute and relax a few assumptions. With neighbor-specific knobs or policy configuration applied to the neighbor outbound or inbound as may be the case, we can regenerate and advertise and/or accept the link bandwidth extended community over the EBGP link. In addition, we can define neighbor specific knobs that will aggregate the link bandwidth values from the LB extended communities learnt from the downstream routers (either received as link bandwidth extended community in the path update or assigned at ingress using a neighbor inbound policy configuration or derived from the operational link-speed of the peer link) and then regenerate and advertise (via neighbor outbound policy knob) this aggregate link bandwidth value in the form of the LB extended community to the upstream EBGP router. Since the advertisement is being made to EBGP neighbors, the next-hop is going to be reset at the advertising router.

Speaking of overall traffic profile, if we assume that on ingress at R4 traffic flow for net p/m is received at a data rate of 'x', then in absence of link bandwidth regeneration at R3 and R5 the resultant traffic profile is below:

link	ratio	percent approximation(~)
R4-R3	1/2x	50%
R4-R5	1/2x	50%
R3-R1	1/3x (1/2 * 2/3)	33%
R3-R2	1/6x (1/2 * 1/3)	17%
R5-R6	2/7x (1/2 * 4/7)	29%
R5-R7	3/14x (1/2 * 3/7)	21%

For comparison the resultant traffic profile in presence of cumulative link bandwidth regeneration at R3 and R5 is as below:

link	ratio	percent approximation(~)
R4-R3	3/10x	30%
R4-R5	7/10x	70%
R3-R1	1/5x (3/10 * 2/3)	20%
R3-R2	1/10x (3/10 * 1/3)	10%
R5-R6	2/5x (7/10 * 4/7)	40%
R5-R7	3/10x (7/10 * 3/7)	30%

As is evident, the second table is closer to the desired traffic profile that shoud be received by the leaf nodes (R1, R2, R6, R7) compared to the first one.

4. Large Scale Data Centers Use Case

The "Use of BGP for Routing in Large-Scale Data Centers" [RFC7938] describes a way to design large scale data centers using EBGP across the different routing layers. [RFC7938] section 6.3 ("Weighted ECMP") describes a use case in which a service (most likely represented using an anycast virtual IP) has an unequal set of resources serving across the data center regions. Figure 3 shows a

typical data center topology as described in section 3.1 of [RFC7938] where an unequal number of servers are deployed advertising a certain BGP prefix. As can be seen in the figure, the left side of the data center hosts only 3 servers while the right side hosts 10 servers.

```
                   +------+  +------+
                   |      |  |      |
                   | AS1  |  | AS1  |           Tier 1
                   |      |  |      |
                   +------+  +------+
                     |  |      |  |
           +---------+  |      |  +----------+
           | +-------+--+------+--+-------+  |
           | |       |  |      |  |       |  |
         +----+     +----+    +----+     +----+
         |    |     |    |    |    |     |    |
         |AS2 |     |AS2 |    |AS3 |     |AS3 | Tier 2
         |    |     |    |    |    |     |    |
         +----+     +----+    +----+     +----+
            |         |          |         |
            |         |          |         |
            | +-----+ |          | +-----+ |
            +-| AS4 |-+          +-| AS5 |-+    Tier 3
              +-----+              +-----+
               | | |                | | |

          <- 3 Servers ->    <- 10 Servers ->
```

Typical Data Center Topology (RFC7938)

Figure 3

In a regular ECMP environment, the tier 1 layer would see an ECMP path equally load-sharing across all 4 tier 2 paths. This would cause the servers on the left part of the data center to be potentially overloaded, while the servers on the right to be underutilized. Using link bandwidth advertisements the servers could add a link bandwidth extended community to the advertised service prefix. Another option is to add the extended community on the tier 3 network devices as the routes are received from the servers or generated locally on the network devices. If the link bandwidth value advertised for the service represents the server capacity for that service, each data center tier would aggregate the values up when sending the update to the higher tier. The result would be a set of weighted load-sharing metrics at each tier allowing the network to distribute the flow load among the different servers in

the most optimal way. If a server is added or removed to the service prefix, it would add or remove its link bandwidth value and the network would adjust accordingly.

Figure 4 shows a more popular Spine Leaf architecture similar to [RFC7938] section 3.2. Tor1, Tor2 and Tor3 are in the same tier, i.e. the leaf tier (The representation shown in Figure 3 here is the unfolded Clos). Using the same example above, it is clear that the LB extended community value received by each of Spine1 and Spine2 from Tor1 and Tor2 is in the ratio 3 to 10 respectively. The Spines will then aggregate the bandwidth, regenerate and advertise the LB extended-community to Tor3. Tor3 will do equal cost sharing to both the spines which in turn will do the traffic-splitting in the ratio 3 to 10 when forwarding the traffic to the Tor1 and Tor2 respectively.

```
                       +------+
                       | Tor3 |      Tier 1
                       +------+
                            |
                  +- - - - -+- - - - +
                  |                  |
               +-----+             +-----+
               |    |              |    |
               |Spine1             |Spine2
               |    |              |    |
               +----+--+         +-+-----+
                 |      \       /     |
                         - + - -
                 |      /       \     |
              +-----+- +          -+-----+
              |Tor1 |              |Tor2 |   Tier 1
              +-----+              +-----+
               | | |                | | |
        <- 3 Servers ->     <- 10 Servers ->
```

Two-tier Clos Data Center Topology

Figure 4

5. Non-Conforming BGP Topologies

This use-case will not readily apply to all topologies. Figure 5 shows a all EBGP topology: R1, R2, R3, R4, R5 and R6 are in AS1, AS2, AS3, AS4, AS5 and AS6 respectively. A net p/m, is being advertised

from a server S1 with LB extended-community value 10 to R1 and R5. R1 advertises p/m to R2 and R3 and also regenerates the LB extended-community with value 10. R4 receives the advertisements from R2, R3 and R5 and computes the aggregate bandwidth to be 30. R4 advertises p/m to R6 with LB extended-community value 30. The link bandwidths are as shown in the figure.

In the example as can be seen, R4 will do the cumulative bandwidth of the LB that it receives from R2, R3 and R5 which is 30. When R4 receives the traffic from R6, it will load-balance it across R2, R3 and R5. As a result R1 will receive twice the volume of traffic that R5 does. This is not desirable because the bandwidth from R1 to S1 and the bandwidth from S1 to R5 is the same i.e. 10. The discrepancy arose because when R4 aggregated the link bandwidth values from the received advertisements, the contribution from R1 was actually factored in twice.

```
              |- - R2 - 10  --|
              |               |
              |               |
    S1- - 10- R1              R4- - - --30 - -R6
     |        |               |
     |        |               |
    10        |- - -R3- 10 - -|
     |                        |
     |- - - R5 - - -- - -- - - -|
```

A non-conforming topology for the Cumulative DMZ

Figure 5

One way to make the topology in the figure above conforming would be to regenerate a normalized value of the aggregate link bandwidth when the aggregate link bandwidth is being advertised over more than one eBGP peer link. Such normalization can be achieved through outbound policy application on top of the aggregate link bandwidth value. A couple of options in this context are:

1. divide the aggregate link bandwidth across the eBGP peers equally

2. divide the aggregate link bandwidth across the eBGP peers as per the ratio of the operational link capacity of the eBGP peer links

These and similar options for regeneration of link-bandwidth to cater to load-balancing requirements in such topologies are outside the

scope of this document and can be implementated as additional outbound policy enhancements on top of a computed aggregate link bandwidth.

6. Protocol Considerations

[I-D.ietf-idr-link-bandwidth] needs to be refreshed. No Protocol Changes are necessary if the knobs are implemented as recommended. The other way to achieve the same purpose would be to use some complicated policy frameworks. But that is only a conjecture.

7. Operational Considerations

A note may be made that these solutions also are applicable to many address families such as L3VPN [RFC2547] , IPv4 with labeled unicast [RFC8277] and EVPN [RFC7432].

In topologies and implementation where there is an option to advertise all multipath (equal cost) eligible paths to eBGP peers (i.e. 'ecmp' form of additional-path advertisement is enabled), aggregate link bandwidth advertisement may not be required or may be redundant since the receiving BGP speaker receives the link bandwidth extended community values with all eligible paths, so the aggregate link bandwidth is effectively received by the downstream eBGP speaker and can be used in the local computation to affect the forwarding behaviour. This assumes the additional paths are advertised with next-hop self.

8. Security Considerations

This document raises no new security issues.

9. Acknowledgements

Viral Patel did substantial work on an implementation along with the first author. The authors would like to thank Acee Lindem and Jakob Heitz for their help in reviewing the draft and valuable suggestions. The authors would like to thank Shyam Sethuram, Sameer Gulrajani, Nitin Kumar, Keyur Patel and Juan Alcaide for discussions related to the draft.

10. References

10.1. Normative References

[I-D.ietf-idr-link-bandwidth]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-06 (work in progress), January 2013.

[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>.

[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>.

10.2. Informative References

[RFC2547] Rosen, E. and Y. Rekhter, "BGP/MPLS VPNs", RFC 2547, DOI 10.17487/RFC2547, March 1999, <https://www.rfc-editor.org/info/rfc2547>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

Authors' Addresses

Satya Ranjan Mohanty
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: satyamoh@cisco.com

Arie Vayner
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
USA

Email: avayner@google.com

Akshay Gattani
Arista Networks
5453 Great America Parkway
Santa Clara, CA 95054
USA

Email: akshay@arista.com

Ajay Kini
Arista Networks
5453 Great America Parkway
Santa Clara, CA 95054
USA

Email: ajkini@arista.com