BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: 28 August 2026

S. Litkowski, Ed.
Cisco
S R. Mohanty, Ed.
Zscaler
A. Vayner
Nvidia
A. Gattani
A. Kini
Arista Networks
J. Tantsura
Nvidia
R. Das
HPE

24 February 2026

BGP link bandwidth extended community use cases
draft-ietf-bess-ebgp-dmz-09

Abstract

BGP link bandwidth extended community provides a way to signal a value along with a BGP path that can be used to perform weighted load-balancing in multipath scenarios. This document details various use cases of the BGP link bandwidth extended community. It also describes local mechanisms to dynamically adjust the BGP link bandwidth value or the multipath weights based on different considerations.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on 28 August 2026.

Table of Contents

1. Introduction 2
2. Requirements Language 4
3. Use cases 4
3.1. Internet exit points 4
3.2. Optimizing server load-balancing 5
3.3. External connectivity and top-down LB extended community propagation 8
4. Mechanisms to adjust BGP link bandwidth and path weights . . 10
4.1. Link bandwidth discovery 10
4.2. Contributing link bandwidth computation 11
4.3. Cumulating link bandwidth 12
5. Operational Considerations 13
6. Security Considerations 14
7. Acknowledgements 15
8. References 15
8.1. Normative References 15
8.2. Informative References 15
Authors' Addresses 16

1. Introduction

BGP link bandwidth (LB) extended community (defined in [I-D.ietf-idr-link-bandwidth]) provides a way to perform weighted load-balancing in multipaths scenarios.

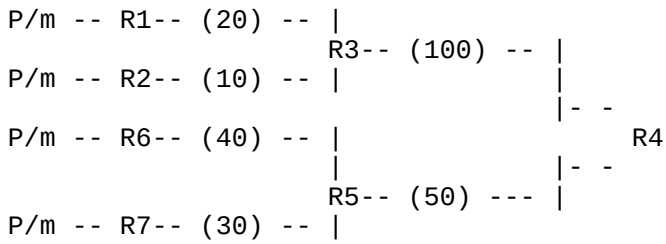

Figure 1: EBGP based network with BGP link bandwidth

Figure 1 represents an all-EBGP network using link address based sessions. Router R3 has eBGP sessions with two downstream routers R1 and R2 and with another upstream router R4. Similarly, R5 has eBGP sessions to downstream routers R6 and R7 and with upstream router R4. A prefix 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 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 multipath configured and paths having different value as-path attributes can still be considered as multipath (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. BGP link bandwidth extended community can be used for this purpose.

R3 can reach P/m via R1 through a link having a bandwidth of 20Gbps and R2 through a link having 10Gbps. As R3 has multipath configured, weighted ECMP is performed with a ratio 2:1 between R1 and R2. Hence, R3 has a total bandwidth of 30Gbps available to reach P/m. When advertising P/m to R4, R3 sets itself as nexthop and can add a BGP link bandwidth extended community. To enable R4 to perform proper weighted load-balancing, R3 could advertise 30Gbps of bandwidth into the BGP link bandwidth extended community. Similarly, R5 performs weighted load-balancing with a ratio of 4:3 between R6 and R7 and could advertise 70Gbps of bandwidth into the BGP link bandwidth extended community when advertising to R4.

R4 receives two BGP paths for P/m:

* one path from R3 with a BGP link bandwidth of 30Gbps. R3 is locally reachable through a link having 100Gbps of capacity.

* one path from R5 with a BGP link bandwidth of 70Gbps. R5 is locally reachable through a link having 50Gbps only of capacity.

R4 may decide to consider only the BGP link bandwidth attribute when performing weighted load-balancing. Then, R4 will use a ratio of 30%/70% between R3 and R5. However, with a local link capacity of only 50Gbps to R5, sending 70% of the traffic to R5 may create congestion. It may be then interesting for R4 to consider both the BGP link bandwidth attribute and the local link bandwidth when computing the weighted load-balancing ratios.

The example above shows that the BGP link bandwidth extended community may not be limited to carry only a "link" bandwidth value. In our example, the value 30Gbps advertised by R3 represents an aggregated path bandwidth.

This document details various use cases related to the usage of BGP link bandwidth extended community and also deployed local mechanisms to dynamically adjust BGP link bandwidth value or weights used during load-balancing.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Use cases

3.1. Internet exit points

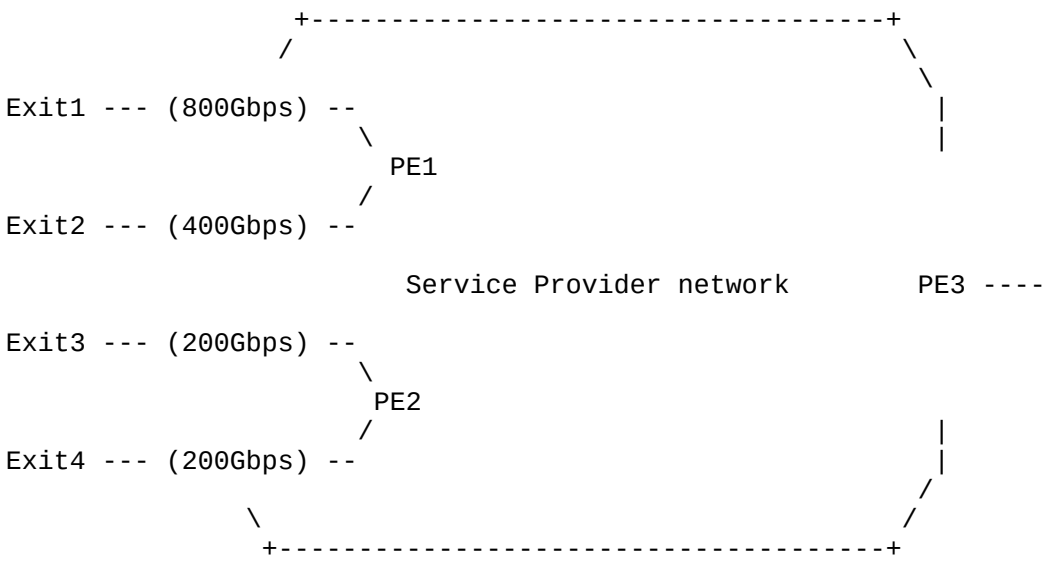

Figure 2: Internet exit points

Figure 2 describes a service provider network with multiple Internet exit points. Each exit point has a different link bandwidth represented in the figure. For best efficiency, PE3 should load-balance Internet traffic between PE1 and PE2 according to the bandwidth each one of them has available. PE1 is connected to two exit points, one with 800Gbps, the other with 400Gbps, it can reach Internet prefixes with a total capacity of 1.2Tbps. PE2 is connected to two exit points, one with 200Gbps, the other with 200Gbps, it can reach Internet prefixes with a total capacity of 400Gbps. We should then expect PE3 to load-balance between PE1 and PE2 with a ratio of 3:1.

BGP link bandwidth extended community can be leverage by PE1 and PE2 to advertise the bandwidth available to PE3.

3.2. Optimizing server load-balancing

[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 (anycast IP). As displayed in the figure, the left side of the data center hosts only 5 servers while the right side hosts 10 servers.

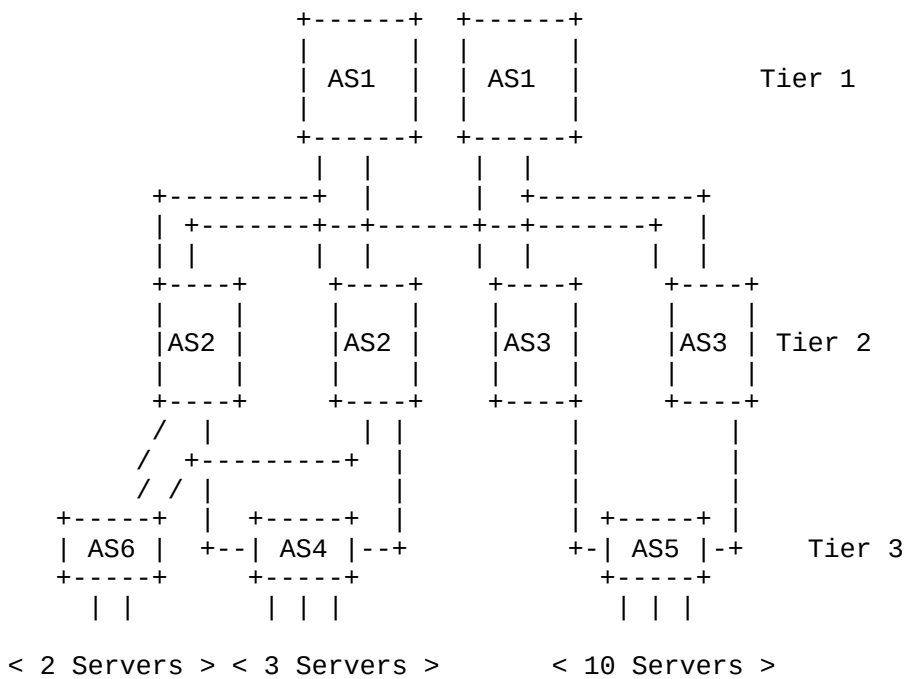

Figure 3: Clos topology with unequal number of servers per cluster

In a regular ECMP environment, the tier 1 layer would see an ECMP path equally load-sharing across all four 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 will be underutilized. By leveraging BGP link bandwidth extended community, the servers could add a link bandwidth 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.

* AS4 has 3 servers and will advertise a link bandwidth of 3.

* AS6 has 2 servers and will advertise a link bandwidth of 2.

* AS2 will aggregate link bandwidth values from AS6 and AS4 and will advertise a link bandwidth of 5 to AS1.

* AS5 has 10 servers and will advertise a link bandwidth of 10. As AS5 is the only tier 3 cluster for AS3, AS3 will advertise the same value of 10 to AS1.

* AS1 can then perform a better load-balancing by using a ratio of 1:2 between AS2 and AS3.

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.

Same use case can be applied to a two tier clos-topology as described in Figure 4. Tor1, Tor2 and Tor3 are in the same tier, i.e. the leaf tier. Using the same considerations as in the previous example, 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.

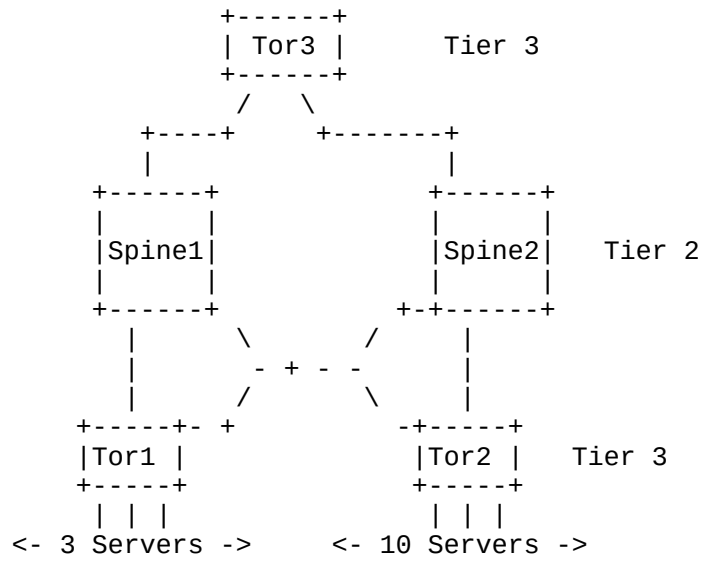

Figure 4: Clos Data Center Topology

3.3. External connectivity and top-down LB extended community propagation

While, in [RFC7938] section 5.2.4., external connectivity module is described as a separate cluster with Tier 3 devices being WAN routers, it is much more common to extend connectivity to a regional aggregation block over Tier 1 layer, where a number of multiplanar Tier 1 blocks connect into regional aggregation blocks, extending commonly used 5-7 stages fabric by a number of additional stages instantiated within the regional aggregation block. Consequently, external connectivity is implemented within the regional aggregation block. The total BW available towards WAN is significantly lower than the total BW within the fabric.

In Figure 5, we consider a datacenter topology with three tiers. Links in DC2 are not displayed to simplify the figure (DC1 topology is also collapsed). All T3s connect to the two local T2s, all T2 connects to both T1s. All T1s connect to all aggregation routers for external connectivity. In order to be able to load-share traffic in accordance to the capacity available towards DC1 aggregation blocks or the WAN, DC1 aggregate blocks and WAN routes are tagged with a BGP link bandwidth extended community. For instance, each aggregation router advertises:

* WAN routes with a link bandwidth value corresponding to the bandwidth available to reach the WAN (for instance, the bandwidth of the link to the upstream WAN router).

* DC1 aggregate blocks with a link bandwidth value corresponding to the bandwidth available to reach DC1.

The LB extended community is propagated top-down (to the tier 3) to DC2, reflecting the bandwidth available towards DC1 aggregation blocks and the WAN. Any partial loss of connectivity to DC1 or the WAN will be reflected to DC2 devices which will adapt their weighted load-balancing.

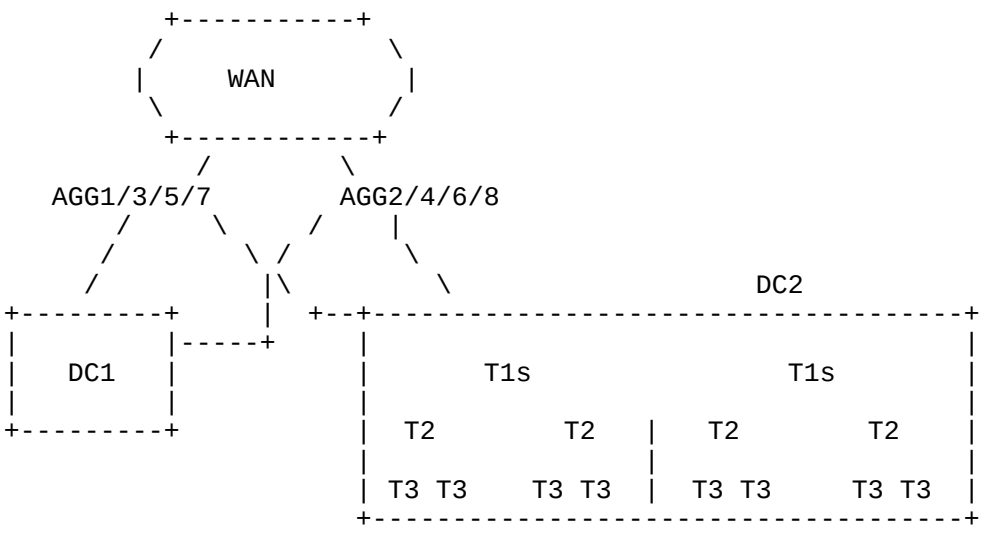

Figure 5: Data center external connectivity

Considering Figure 6, each T1 router is connected to multiple aggregation router. Each aggregation router advertises WAN routes with a BGP LB corresponding the bandwidth available to reach the WAN. If some T1 - AGG or AGG - WAN links fail, the weighted load-balancing is automatically adjusted on T1s according to the remaining BGP paths. Each T1 updates the cumulative available bandwidth down to the T2s (sum of bandwidth of all paths from aggregation routers). T2s will do a similar processing when receiving paths from T1s, adjust their weighted load-balancing based on bandwidth present in the paths available from T1 and cumulate the bandwidth before advertising down to T3s. Depending on the mesh density and the total bandwidth available within the fabric, DC devices may also consider their local link bandwidth in regards to the received BGP LB when computing the weighted load-balancing and performing the cumulation.

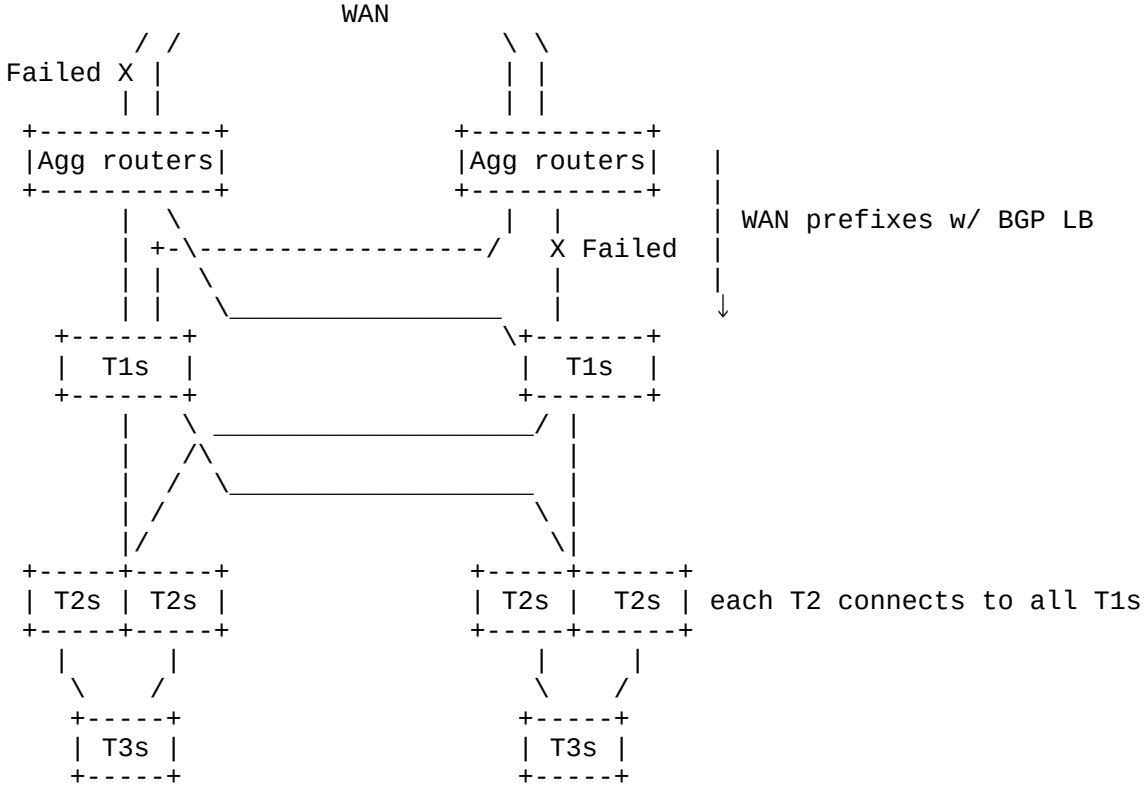

Figure 6: Data center external connectivity and internal failures

4. Mechanisms to adjust BGP link bandwidth and path weights

The use cases described earlier involve multiple ways to dynamically adjust the BGP link bandwidth value advertised or the value to be used for weighted load-balancing. This section describes these mechanisms.

4.1. Link bandwidth discovery

BGP link extended community is usually set by either inbound or outbound route policy. In case of directly connected BGP session, it may be desirable to set the link bandwidth value dynamically based on the actual bandwidth of interface used by the session. The setting of the link bandwidth value based on interface bandwidth MAY be activated through a configuration knob and MAY be done for all prefixes or a subset of prefixes learned over the BGP session. Upon

interface bandwidth change, BGP should update the link bandwidth value for the affected prefixes and adjust the weights accordingly.

4.2. Contributing link bandwidth computation

With the link bandwidth discovery, BGP may get link bandwidth value from two sources for a particular path:

* from BGP link bandwidth extended community received in the BGP update from the neighbor. We call it the remote bandwidth in this document.

* from the discovered interface bandwidth over which the BGP session is established. We call it the local link bandwidth in this document.

In addition, as illustrated in the previous sections, BGP may have to consider a combination of the local link and remote bandwidth when computing the weights for weighted load-balancing. Any function of the two may be used like for instance a "minimum" function that was highlighted in Section 1. The weight of each path may then be based on either:

* Only the remote bandwidth

* Only the local link bandwidth

* A function of both

This is controlled by a local configuration. The resulting value is called contributing link bandwidth. If contributing bandwidth is set to local link bandwidth but bandwidth is not learned from interface, the contributing bandwidth is considered as zero. If contributing bandwidth is set to a function of local link and remote bandwidth, but the function cannot return a result, the contributing bandwidth is considered as zero. If contributing bandwidth is set to remote bandwidth but no remote bandwidth is received, the contributing bandwidth is considered as zero. By default, the contributing bandwidth is set to the remote bandwidth if present, if not, it is set to the local link bandwidth and if local link bandwidth is not available, it is set to zero.

[I-D.ietf-idr-link-bandwidth] mandates all BGP paths of a prefix and being part of the multipath to have BGP LB extended community with a non-zero value to perform weighted load-balancing. Otherwise the behavior is determined by local policy. This document updates this behavior as follows. If all of these conditions are met, weighted load-balancing can be done otherwise behavior is determined by local policy:

* all BGP paths of a prefix MUST have a non-zero contributing bandwidth.

* for consistency reason, all BGP paths of a prefix MUST use a contributing bandwidth based on the same source (e.g.: all paths are using contributing bandwidth from the same function).

4.3. Cumulating link bandwidth

Cumulation of link bandwidth is the action of performing the sum of the bandwidth associated to each BGP path of a prefix (being part of multipath). The cumulation is done when advertising the LB to a neighbor and if nexthop is modified in accordance to [I-D.ietf-idr-link-bandwidth]. Cumulation is optional and MAY be enabled for all prefixes or only a subset of prefixes by using configuration knobs. It may also be controlled on a per-neighbor basis.

By default, cumulation SHOULD be done based on the contributing bandwidth of each path, hence, the device will advertise a bandwidth value which is aligned with the total bandwidth that it takes into account during weighted load-balancing. This behavior may be modified by a local configuration, so cumulation can be done only based on local link bandwidth for instance.

When performing cumulation, if one of the paths has no bandwidth information on which the cumulation is based (e.g. contributing bandwidth or local link bandwidth), its bandwidth is considered as zero.

```
P/m
   Path1:
     NH=R1, Link-BW extended-community: 1000Mbps,
     local-link-bandwidth: 2000Mbps, contributing-bandwidth: 2000
     best, multipath

   Path2:
     NH=R2, Link-BW extended-community: 2000Mbps,
     local-link-bandwidth: 2000Mbps, contributing-bandwidth: 4000
     multipath

   Path3:
     NH=R3, Link-BW extended-community: 3000Mbps,
     multipath
```

Considering the prefix P/m and BGP paths described above and that P/m must be advertised to an eBGP neighbor R4 with cumulation of link bandwidth:

* if cumulation is configured to be based on remote bandwidth, P/m will be advertised with a BGP link bandwidth extended community with a value of 6000 Mbps

* if cumulation is configured to be based on local link bandwidth, P/m will be advertised with a BGP link bandwidth extended community with a value of 4000 Mbps. Path3 does not have a local link bandwidth so its local link bandwidth is considered as zero.

* if cumulation is configured to be based on contributing bandwidth, P/m will be advertised with a BGP link bandwidth extended community with a value of 6000. Path3 does not have a contributing bandwidth so its is considered as zero.

5. Operational Considerations

Some of the use cases present earlier are applicable to many address families such as L3VPN [RFC4364] , IPv4 with labeled unicast [RFC8277] and EVPN [RFC7432].

In topologies and implementation where there is an option to advertise all multipaths eligible paths ([RFC7911]) to BGP peers (i.e. 'ecmp' form of additional-path advertisement is enabled), when next-hop is changed during advertisement of these multiple paths, cumulated 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 BGP 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.

Propagation of the BGP link bandwidth, taking into account local link bandwidth and with cumulation happening at multiple stages (e.g.: at every tier of a DC) has the benefit of providing the end-to-end view of the bandwidth available through the path. However, it has the drawback of creating additional churn within the BGP control plane. In Figure 6, any link failure (e.g: aggregation router to WAN, T1 - T2,...) will trigger a link bandwidth attribute value change and a new BGP update to be sent. Implementation MAY provide local mechanisms to suppress some advertisements (e.g.: don't advertise if bandwidth change is less than x %).

6. Security Considerations

This document raises no new security issues compared to [I-D.ietf-idr-link-bandwidth].

* The discovery of bandwidth by BGP from an interface is not a security concern. Fast interfaces flaps (that may be created by an attacker) may be damped by existing mechanisms so events are not reported to BGP. As mentioned in Section 5, implementations may also provide suppression mechanisms to avoid propagation of bandwidth updates churn throughout the network.

* Choice of source of information for getting the contributing bandwidth of a path is a local configuration and is not a vector of attack. An attacker having access to the device configuration could create far more damages than just manipulating this configuration.

* Cumulation of bandwidth during advertisement is nothing more than applying a new value to the LB extended community. This is not considered as a security threat.

7. Acknowledgements

Viral Patel did substantial work on an implementation along with the first author. The authors would like to thank Acee Lindem, Jakob Heitz, Swadesh Agrawal, Serge Krier 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.

8. References

8.1. Normative References

[I-D.ietf-idr-link-bandwidth] Mohapatra, P., Das, R., Satya, M. R., Krier, S., Szarecki, R. J., and A. Gattani, "BGP Link Bandwidth Extended Community", Work in Progress, Internet-Draft, draft-ietf-idr-link-bandwidth-17, 10 September 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-link-bandwidth-17>.

[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>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

8.2. Informative References

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[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>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[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

Stephane Litkowski (editor)
Cisco
Email: slitkows@cisco.com

Satya Ranjan Mohanty (editor)
Zscaler
120 Holgers Way
San Jose, CA 95134
United States of America
Email: smohanty@zscaler.com

Arie Vayner
Nvidia
Email: avayner@nvidia.com

Akshay Gattani
Arista Networks
5453 Great America Parkway
Santa Clara, CA 95054
United States of America
Email: akshay@arista.com

Ajay Kini
Arista Networks
5453 Great America Parkway
Santa Clara, CA 95054
United States of America
Email: ajkini@arista.com

Jeff Tantsura
Nvidia
Email: jefftant.ietf@gmail.com

Reshma Das
HPE
Email: dreshma@juniper.net