BESS WorkGroup
Internet-Draft
Intended status: Informational
Expires: 17 December 2024

S R. Mohanty
Cisco Systems
A. Vayner
Google
A. Gattani
A. Kini
Arista Networks
J. Tantsura
Nvidia
R. Das
Juniper Networks Inc.
15 June 2024

Cumulative DMZ Link Bandwidth and load-balancing
draft-ietf-bess-ebgp-dmz-05

Abstract

The DMZ Link Bandwidth draft provides a way to load-balance traffic to a destination which is reachable via more than one path according to the weight attahced. 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 that employs multi-path. The link-bandwidth value is then extracted from the extended community and is used as a weight in the RIB/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

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 17 December 2024.

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 IBGP peers. 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.

It is noted that there exist some vendors where the link bandwidth extended community is implemented as optional transitive. For interop purposes, it is recommended that in newer vendor implementations, vendors will accept both types i.e., transitive, and non-transitive (type/subtype) viz. 0x0004 and 0x4004. Vendors may need to implement an advertise knob to advertise the link-bandwidth community as transitive or non-transitive depending on requirement.

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.

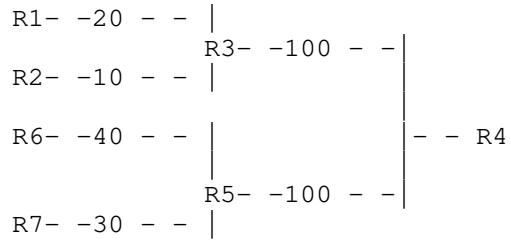


Figure 1

EBGP Network with cumulative DMZ requirement

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.

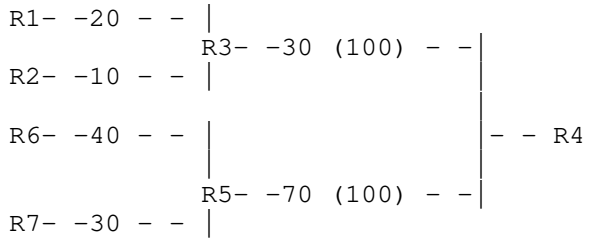


Figure 2

EBGP Network showing advertisement of cumulative link bandwidth

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 and to 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 Cases

The â\200\234Use 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/data centere stages. [RFC7938] section 6.3 (â\200\234Weighted 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.

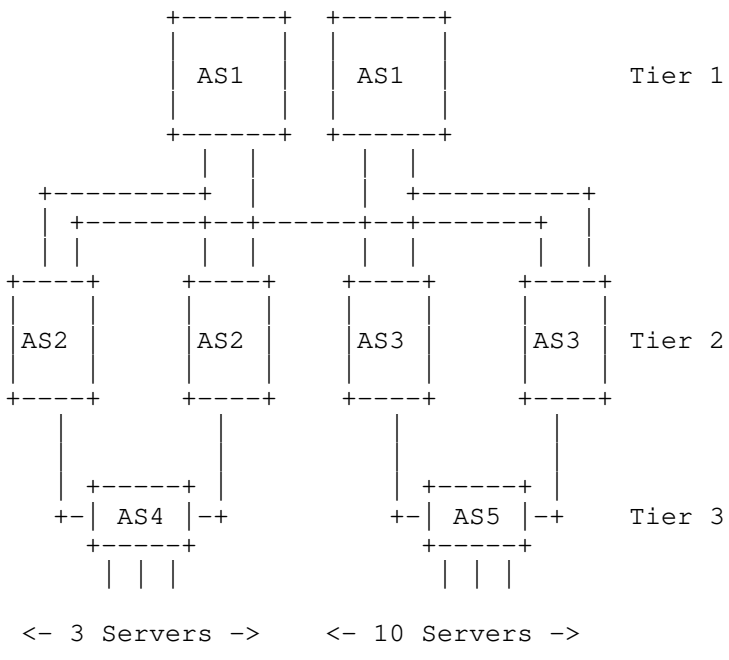


Figure 3

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.

Typical Data Center Topology (RFC7938)

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.

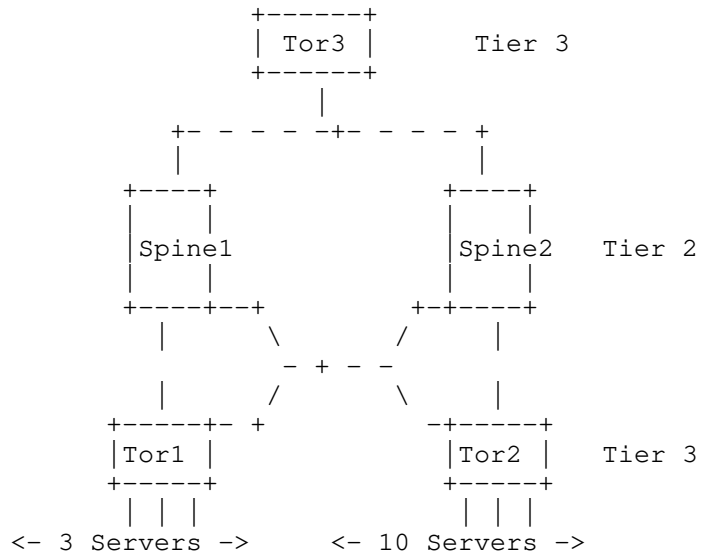


Figure 4

Two-tier Clos Data Center Topology

4.1. 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 the above examples, LB extended community propagation is bottom-up (W-ECMP towards a service/Tier 3). To address partial loss of external connectivity, LB extended community is propagated top-down, reflecting BW available towards regional aggregation blocks and the WAN. While, due to densely meshed connectivity and total BW available within the fabric and its ability to accommodate BW needed in case of partial loss of connectivity weighted ECMP is not mandatory, due to lower capacity, partial loss of connectivity towards regional aggregation blocks and the WAN can cause packet loss and/or increased latency and as the result, reduced availability. In order to be able to load-share traffic in accordance to the capacity available towards regional aggregation blocks and the WAN all routes that come from regional aggregation blocks (WAN routes + routes from other DCs) are to be tagged with the LB extended community and propagated to the bottom of the fabric. This allows to load-share the traffic between planes as well as within a plane in accordance with the associated weight.

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.

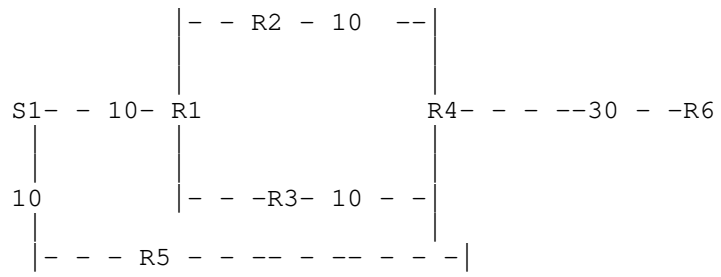


Figure 5

A non-conforming topology for the Cumulative DMZ

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 [RFC4364] , 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", Work in Progress, Internet-Draft, draft-ietf-idr-link-bandwidth-07, 5 March 2018, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-link-bandwidth-07>.

[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

[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>.

[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
United States of America
Email: satyamoh@cisco.com

Arie Vayner
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
United States of America
Email: avayner@google.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
Juniper Networks Inc.
Email: dreshma@juniper.net

BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: 10 March 2025

N. Malhotra, Ed.
A. Sajassi
Cisco Systems
J. Rabadan
Nokia
J. Drake
Juniper
A. Lingala
ATT
S. Thoria
Cisco Systems

6 September 2024

Weighted Multi-Path Procedures for EVPN Multi-Homing
draft-ietf-bess-evpn-unequal-lb-22

Abstract

EVPN enables all-active multi-homing for a CE (Customer Equipment) device connected to two or more PE (Provider Equipment) devices via a LAG (Link Aggregation), such that bridged and routed traffic from remote PEs to hosts attached to the Ethernet Segment can be equally load balanced (it uses Equal Cost Multi Path) across the multi-homing PEs. EVPN also enables multi-homing for IP subnets advertised in IP Prefix routes, so that routed traffic from remote PEs to those IP subnets can be load balanced. This document defines extensions to EVPN procedures to optimally handle unequal access bandwidth distribution across a set of multi-homing PEs in order to:

* provide greater flexibility, with respect to adding or removing individual multi-homed PE-CE links.

* handle multi-homed PE-CE link failures that can result in unequal PE-CE access bandwidth across a set of multi-homing PEs.

In order to achieve the above, it specifies signaling extensions and procedures to:

* Loadbalance bridged and routed traffic across egress PEs in proportion to PE-CE link bandwidth or a generalized weight distribution.

* Achieve BUM (Broadcast, UnknownUnicast, Multicast) DF (Designated Forwarder) election distribution for a given ES (Ethernet Segment) across the multi-homing PE set in proportion to PE-CE link bandwidth. Section 6 of this document further updates RFC 8584, draft-ietf-bess-evpn-per-mcast-flow-df-election and draft-ietf-bess-evpn-pref-df in order for the DF election extension defined in this document to work across different DF election algorithms.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

In an EVPN IRB (Integrated Routing and Bridging) overlay network as described in [RFC9135], with a CE multi-homed via a EVPN all-active multi-homing, bridged and routed traffic from ingress PEs can be equally load balanced (ECMPed) across the multi-homing egress PEs:

* ECMP Load-balancing for bridged unicast traffic is enabled via aliasing and mass-withdraw procedures detailed in [RFC7432].

* ECMP Load-balancing for routed unicast traffic is enabled via existing L3 ECMP mechanisms.

* Load-sharing of bridged BUM (Broadcast, UnknownUnicast, Multicast) traffic on local ports is enabled via EVPN DF election procedure detailed in [RFC7432].

All of the above load balancing and DF election procedures implicitly assume equal bandwidth distribution between the CE and the set of egress PEs. Essentially, with this assumption of equal "access" bandwidth distribution across all egress PEs, all remote traffic is equally load balanced across the egress PEs. This assumption of equal access bandwidth distribution can be restrictive with respect to adding / removing links in a multi-homed LAG interface and may also be easily broken on individual link failures. A solution to handle unequal access bandwidth distribution across a set of egress PEs is specified in this document. Primary motivation behind this proposal is to enable greater flexibility with respect to adding / removing member PE-CE links, as needed and to optimally handle PE-CE link failures.

1.1. PE-CE Link Provisioning

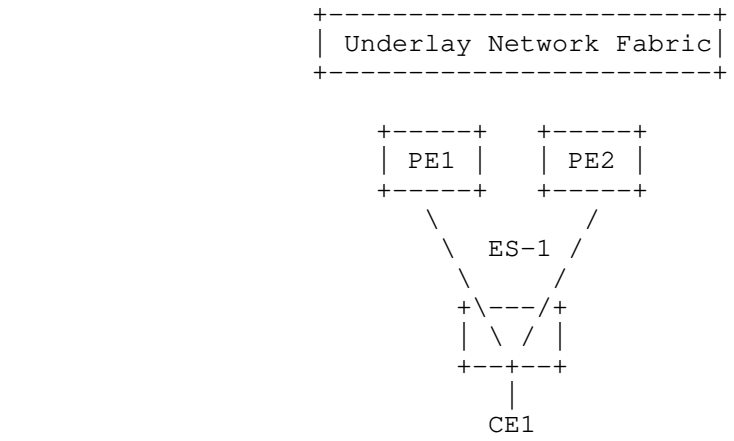


Figure 1

Consider CE1 that is dual-homed to egress PE1 and egress PE2 via EVPN all-active multi-homing with single member links of equal bandwidth to each PE (aka, equal access bandwidth distribution across PE1 and PE2). If the provider wants to increase link bandwidth to CE1, it must add a link to both PE1 and PE2 in order to maintain equal access bandwidth distribution and inter-work with EVPN ECMP load balancing. In other words, for a dual-homed CE, total number of CE links must be provisioned in multiples of 2 (2, 4, 6, and so on). For a triple-homed CE, number of CE links must be provisioned in multiples of three (3, 6, 9, and so on). To generalize, for a CE that is multi-

homed to "n" PEs, number of PE-CE physical links provisioned must be an integral multiple of "n". This is restrictive in case of dual-homing and very quickly becomes prohibitive in case of multi-homing.

Instead, a provider may wish to increase PE-CE bandwidth or number of links in any link increments. As an example, for CE1 dual-homed to egress PE1 and egress PE2 in all-active mode, provider may wish to add a third link to only PE1 to increase total bandwidth for this CE by 50%, rather than being required to increase access bandwidth by 100% by adding a link to each of the two PEs. While existing EVPN based all-active load balancing procedures do not necessarily preclude such asymmetric access bandwidth distribution among the PEs providing redundancy, it may result in unexpected traffic loss due to congestion in the access interface towards CE. This traffic loss is due to the fact that PE1 and PE2 will continue to be treated as equal cost paths at remote PEs, and as a result may attract approximately equal amount of CE1 destined traffic, even when PE2 only has half the bandwidth to CE1 as PE1. This may lead to congestion and traffic loss on the PE2-CE1 link. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts must also be load balanced across PE1 and PE2 in 2:1 manner.

1.2. PE-CE Link Failures

More importantly, unequal PE-CE bandwidth distribution described above may occur during regular operation following a link failure, even when PE-CE links were provisioned to provide equal bandwidth distribution across multi-homing PEs.

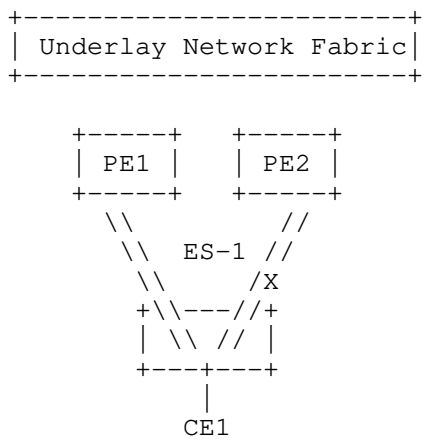


Figure 2

Consider a CE1 that is multi-homed to egress PE1 and egress PE2 via a LAG with two member links to each PE. On a PE2-CE1 physical link failure, LAG represented by an Ethernet Segment ES-1 on PE2 stays up, however, its bandwidth is cut in half. With existing ECMP procedures, both PE1 and PE2 may continue to attract equal amount of traffic from remote PEs, even when PE1 has double the bandwidth to CE1. If bandwidth distribution to CE1 across PE1 and PE2 is 2:1, traffic from remote hosts must also be load balanced across PE1 and PE2 in 2:1 manner to avoid unexpected congestion and traffic loss on PE2-CE1 links within the LAG. As an alternative, min-link on LAGs is sometimes used to bring down the LAG interface on member link failures. This however results in loss of available bandwidth in the network, and is not ideal.

1.3. Design Requirement

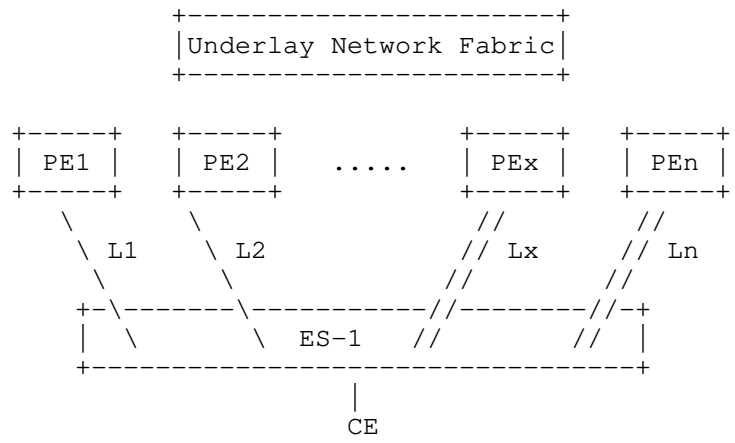


Figure 3

To generalize, if total link bandwidth to a CE is distributed across "n" egress PEs, with Lx being the total bandwidth to PEx across all links, traffic from ingress PEs to this CE must be load balanced unequally across egress PE set [PE1, PE2, ..., PEn] such that, fraction of total unicast and BUM flows destined for CE that are serviced by egress PEx is:

Lx / [L1+L2+.....+Ln]

Figure 3 illustrates a scenario where egress PE1..PEn are attached to a multi-homed Ethernet Segment, however this document generalizes this requirement so that the unequal load balancing can be applied to PEs attached to a vES or to a multi-homed subnet advertised by EVPN IP Prefix routes.

The solution specified below includes extensions to EVPN procedures to achieve the above. Following assumption apply to procedure described in this document:

* For procedures related to bridged unicast and BUM traffic, EVPN all active multi-homing is assumed.

* Procedures related to bridged unicast and BUM traffic are applicable to both aliasing and non-alaising mode as defined in [RFC7432].

2. Requirements Language and Terminology

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.

* BW: BandWidth

* LAG: Link Aggregation Group

* ES: Ethernet Segment

* ESI: Ethernet Segment ID

* vES: Virtual Ethernet Segment

* EVI: Ethernet virtual Instance, this is a mac-vrf

* RT-1: EVPN Route Type 1 as defined in [RFC7432]

* RT-2: EVPN Route Type 2 as defined in [RFC7432]

* RT-5: EVPN Route Type 5 as defined in [RFC7432]

* Path-List: A forwarding object used to load-balance routed or bridged traffic across multiple forwarding paths.

* Access Bandwidth: Bandwidth of PE-CE links in an Ethernet Segment

* Egress PE: In the context of an Ethernet Segment or a route, this is the PE that advertises a locally attached Ethernet Segment RT-1, or a locally attached host or prefix route (RT-2, RT-5)

* Ingress PE: In the context of an Ethernet Segment or a route, this is the receiving PE that learns remote Ethernet Segment RT-1 and/or host and prefix routes (RT-2, RT-5) from the Egress PE

* IMET: Inclusive Multicast Route

* DF: Designated Forwarder

* BDF: Backup Designated Forwarder

* DCI: Data Center Interconnect Router

3. Solution Overview

In order to achieve weighted load balancing to an ES or vES for overlay unicast traffic, Ethernet A-D per ES route (EVPN Route Type 1) is leveraged to signal the Ethernet Segment weight to ingress PEs. Using Ethernet A-D per ES route to signal the Ethernet Segment weight provides a mechanism that reacts to changes in access bandwidth or number of access links in a service and host independent manner. Ingress PEs computing the MAC path-lists based on global and aliasing Ethernet A-D routes now have the ability to setup weighted load balancing path-lists based on the ES access bandwidth or number of links received from each egress PE that the ES is multi-homed to.

In order to achieve weighted load balancing of overlay BUM traffic, EVPN ES route (Route Type 4) is leveraged to signal the ES weight to egress PEs within an ES's redundancy group to influence per-service DF election. Egress PEs in an ES redundancy group now have the ability to do service carving in proportion to each egress PE's relative ES weight.

Unequal load balancing to multi-homed subnets is achieved by signaling the weight along with the IP Prefix routes advertised for the subnet.

Procedures to accomplish this are described in greater detail next.

4. EVPN Link Bandwidth Extended Community

A new EVPN Link Bandwidth extended community is defined for the solution specified in this document:

* This extended community is defined of type 0x06 (EVPN Extended Community Sub-Types).

* IANA has assigned sub-type value of 0x10 for the EVPN Link bandwidth extended community, of type 0x06 (EVPN Extended Community Sub-Types).

* EVPN Link Bandwidth extended community is defined as transitive.

4.1. Encoding and Usage of EVPN Link Bandwidth Extended Community

EVPN Link Bandwidth Extended Community value field is used to carry total bandwidth of egress PE's all physical links in an ethernet segment, expressed in Mbits/sec (MegabitsPerSecond) represented as an unsigned integer. Note however that the load balancing algorithm defined in this document uses ratio of Link Bandwidths. Hence, the operator may choose a different unit or use the community as a generalized weight that may be set to link count, locally configured weight, or a value computed based on something other than link bandwidth. In such case, the operator MUST ensure consistent usage of the unit across all egress PEs in an ethernet segment. This may involve multiple routing domains/Autonomous Systems.

In order to facilitate this, as well as avoid interop issues because of provisioning error, one octet in the extended community's six octet 'value' field is used to explicitly signal if the weight encoded in the remaining five octets is link bandwidth expressed in Mbps or a generalized weight value. This results in the following encoding for EVPN link bandwidth extended community:

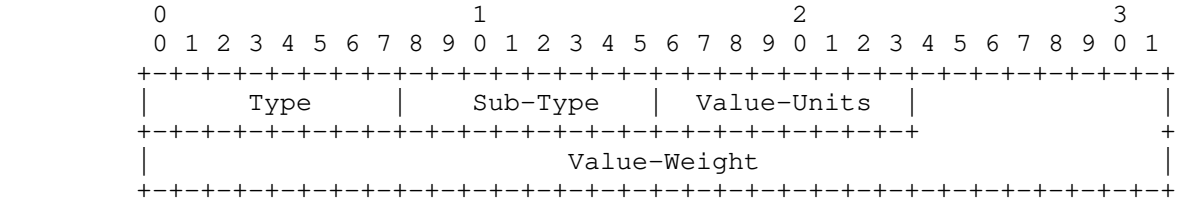


Figure 4

Value-Units is encoded as:

* 0x00: weight expressed using default units of Mbps

* 0x01: generalized weight expressed in something other than Mbps

Generalized weight units are intentionally left arbitrary to allow for flexibility in its usage for different applications without having to define new encoding for each non-default application. Implementations MUST support the default units of Mbps, while support of non-default generalized weight is considered optional.

Additionally, following considerations apply to handling of this extended community at the ingress PE:

* An ingress PE MUST check for consistent 'Value-Units' received in the EVPN link bandwidth exteneded community from each egress PE in an Ethernet Segment. In case of any inconsistency in 'Value-Units' across egress PEs in an Ethernet Segment, this EVPN Link Bandwidth extended community is to be ignored.

* An ingress PE MUST ensure that each route contains only a single instance of this extended community sub-type. In case of more than one instance, this EVPN Link Bandwidth extended community is to be ignored.

5. Weighted Unicast Traffic Load-balancing to an Ethernet Segment

5.1. Egress PE Behavior

A PE that is part of an Ethernet Segment's redundancy group MUST advertise an additional "EVPN link bandwidth" extended community with Ethernet A-D per ES route (EVPN Route Type 1), that carries total bandwidth of PE's physical links in an Ethernet Segment or a generalized weight. New EVPN link bandwidth extended community defined in this document is used for this purpose.

EVPN link bandwidth extended community MUST NOT be attached to per-EVI RT-1 or to EVPN RT-2 as it is a physical ESI property and hence advertised per-ESI.

5.2. Ingress PE Behavior

An ingress PE MUST ensure that the EVPN link bandwidth extended community is recevied from all the egress PEs in an Ethernet Segment and check for consistent 'Value-Units' received from each egress PE in an Ethernet Segment. In case of missing EVPN Link Bandwidth extended community or inconsistent 'Value-Units' from any of the egress PEs in an Ethernet Segment, this EVPN Link Bandwidth extended community is to be ignored by the ingress PE and ingress PE is to follow regular ECMP forwarding to that Ethernet Segment. Ingress PE MUST generate a syslog when the EVPN Link Bandwidth extended community is ignored.

Once consistency of 'Value-Units' is validated, ingress PE SHOULD use the 'Value-Weight' received from each egress PE to compute a relative (normalized) weight for each egress PE, per ES, and then use this relative weight to compute a weighted path-list to be used for load balancing, as opposed to using an ECMP path-list for load balancing across the egress PE paths. Egress PE Weight and resulting weighted path-list computation at ingress PEs is a local matter. An example computation algorithm is shown below to illustrate the idea:

if,

L(x,y) : link bandwidth advertised by egress PE-x for ES-y

W(x,y) : normalized weight assigned to egress PE-x for ES-y

H(y) : Highest Common Factor (HCF) of [L(1,y), L(2,y), ..., L(n,y)]

then, the normalized weight assigned to egress PE-x for ES-y may be computed as follows:

W(x,y) = L(x,y) / H(y)

For a MAC+IP route (EVPN Route Type 2) received with ES-y, ingress PE may compute MAC and IP forwarding path-list weighted by the above normalized weights.

As an example, for a CE multi-homed to PE-1, PE-2, PE-3 via 2, 1, and 1 GE physical links respectively, as part of a LAG represented by ES-10:

L(1, 10) = 2000 Mbps

L(2, 10) = 1000 Mbps

L(3, 10) = 1000 Mbps

H(10) = 1000

Normalized weights assigned to each egress PE for ES-10 are as follows:

W(1, 10) = 2000 / 1000 = 2.

W(2, 10) = 1000 / 1000 = 1.

W(3, 10) = 1000 / 1000 = 1.

For a remote MAC+IP host route received with ES-10, forwarding load balancing path-list may now be computed as: [PE-1, PE-1, PE-2, PE-3] instead of [PE-1, PE-2, PE-3]. This now results in load balancing of all traffic destined for ES-10 across the three egress PEs in proportion to ES-10 bandwidth at each egress PE.

Please note that the pathlist computation algorithm above is for illustration only. Weighted pathlist computation based on the received EVPN link bandwidth extended community is a local implementation choice. As an example, if the received link badwidth values do not result in a good HCF H(y) in the computation method above to be able to compute reasonable weights that can be programmed in hardware, implementation MAY choose another approximation to arrive at rounded integer weight values that can be programmed in hardware.

Weighted path-list computation must only be done for an ES if EVPN link bandwidth extended community is received from all of the egress PE's advertising reachability to that ES via Ethernet A-D per ES Route Type 1. In an unlikely event that EVPN link bandwidth extended community is not received from one or more egress PEs, forwarding path-list should be computed using regular ECMP semantics. Note that a default weight cannot be assumed for an egress PE that does not advertise its link bandwidth as the weight to be used in path-list computation is relative.

If per-ES RT-1 is not advertised or withdrawn from any of the egress PE(s), as per [RFC7432], egress PE is removed from the forwarding path-list for that [EVI, ES]. Hence, the weighted path-list MUST be re-computed.

In an unlikely scenario that per-[ES, EVI] RT-1 is not advertised from any of the egress PE(s), as per [RFC7432], egress PE is not included in the forwarding path-list for that [EVI, ES]. Hence, the weighted path-list for the [EVI, ES] MUST be computed based only on the weights received from egress PEs that advertised the per-[ES, EVI] RT-1.

6. Weighted BUM Traffic Load-Sharing across an Ethernet Segment

Optionally, load sharing of per-service DF role, weighted by individual egress PE's link-bandwidth share within a multi-homed ES may also be achieved.

In order to do that, a new DF Election Capability [RFC8584] called "BW" (Bandwidth Weighted DF Election) is defined. BW MAY be used along with some DF Election Types, as described in the following sections.

6.1. The BW Capability in the DF Election Extended Community

[RFC8584] defines a new extended community for PEs within a redundancy group to signal and agree on uniform DF Election Type and Capabilities for each ES. This document requests IANA to allocate a bit in the "DF Election capabilities" registry setup by [RFC8584]:

Bit 4: BW (Bandwidth Weighted DF Election)

ES routes advertised with the BW bit set will indicate the desire of the advertising egress PE to consider the link-bandwidth in the DF Election algorithm defined by the value in the "DF Type".

As per [RFC8584], all the egress PEs in the ES MUST advertise the same Capabilities and DF Type, otherwise the PEs will fall back to Default [RFC7432] DF Election procedure.

The BW Capability MAY be advertised with the following DF Types:

* Type 0: Default DF Election algorithm, as in [RFC7432]

* Type 1: HRW (Highest Random Weight) algorithm, as in [RFC8584]

* Type 2: Preference algorithm, as in [EVPN-DF-PREF]

* Type 4: HRW per-multicast flow DF Election, as in [EVPN-PER-MCAST-FLOW-DF]

The following sections describe how the DF Election procedures are modified for the above DF Types when the BW Capability is used.

6.2. BW Capability and Default DF Election algorithm

When all the PEs in the Ethernet Segment (ES) agree to use the BW Capability with DF Type 0, the Default DF Election procedure as defined in [RFC7432] is modified as follows:

* Each PE advertises a "EVPN Link Bandwidth" extended community along with the ES route to signal the PE-CE link bandwidth (LBW) for the ES.

* A receiving egress PE MUST use the ES link bandwidth extended community received from each egress PE to compute a relative weight for each egress PE in an Ethernet Segment.

* The DF Election procedure MUST now use this weighted list of egress PEs to compute the per-VLAN Designated Forwarder, such that the DF role is distributed in proportion to this normalized

weight. As a result, a single PE may have multiple ordinals in the DF candidate PE list and 'N' used in (V mod N) operation as defined in [RFC7432] is modified to be total number of ordinals instead of being total number of egress PEs in an Ethernet Segment.

Considering the same example as in Section 5.2, the candidate PE list for DF election is:

[PE-1, PE-1, PE-2, PE-3].

The DF for a given VLAN-a on ES-10 is now computed as (VLAN-a % 4). This would result in the DF role being distributed across PE1, PE2, and PE3 in portion to each PE's normalized weight for ES-10.

6.3. BW Capability and HRW DF Election algorithm (Type 1 and 4)

[RFC8584] introduces Highest Random Weight (HRW) algorithm (DF Type 1) for DF election in order to solve potential DF election skew depending on Ethernet tag space distribution. [EVPN-PER-MCAST-FLOW-DF] further extends HRW algorithm for per-multicast flow based hash computations (DF Type 4). This section describes extensions to HRW Algorithm for EVPN DF Election specified in [RFC8584] and in [EVPN-PER-MCAST-FLOW-DF] in order to achieve DF election distribution that is weighted by link bandwidth.

6.3.1. BW Increment

A new variable called "bandwidth increment" is computed for each [PE, ES] advertising the ES link bandwidth extended community as follows:

In the context of an ES,

L(i) = Link bandwidth advertised by PE(i) for this ES

Lmin = lowest link bandwidth advertised across all PEs for this ES

Bandwidth increment, "b(i)" for a given PE(i) advertising a link bandwidth of L(i) is defined as an integer value computed as:

b(i) = L(i) / Lmin

As an example,

with L(1) = 10, L(2) = 10, L(3) = 20

bandwidth increment for each PE would be computed as:

b(1) = 1, b(2) = 1, b(3) = 2

with L(1) = 10, L(2) = 10, L(3) = 10

bandwidth increment for each PE would be computed as:

b(1) = 1, b(2) = 1, b(3) = 1

Note that the bandwidth increment must always be an integer, including, in an unlikely scenario of a PE's link bandwidth not being an exact multiple of Lmin. If it computes to a non-integer value (including as a result of link failure), it MUST be rounded down to an integer.

6.3.2. HRW Hash Computations with BW Increment

HRW algorithm as described in [RFC8584] and in [EVPN-PER-MCAST-FLOW-DF] computes a random hash value for each PE(i), where, (0 < i <= N), PE(i) is the PE at ordinal i, and Address(i) is the IP address of PE(i).

For 'N' PEs sharing an Ethernet segment, this results in 'N' candidate hash computations. The PE that has the highest hash value is selected as the DF.

We refer to this hash value as "affinity" in this document. Hash or affinity computation for each PE(i) is extended to be computed one per bandwidth increment associated with PE(i) instead of a single affinity computation per PE(i).

PE(i) with b(i) = j, results in j affinity computations:

affinity(i, x), where 1 < x <= j

This essentially results in number of candidate HRW hash computations for each PE that is directly proportional to that PE's relative bandwidth within an ES and hence gives PE(i) a probability of being DF in proportion to it's relative bandwidth within an ES.

As an example, consider an ES that is multi-homed to two PEs, PE1 and PE2, with equal bandwidth distribution across PE1 and PE2. This would result in a total of two candidate hash computations:

affinity(PE1, 1)

affinity(PE2, 1)

Now, consider a scenario with PE1's link bandwidth as 2x that of PE2. This would result in a total of three candidate hash computations to be used for DF election:

affinity(PE1, 1)

affinity(PE1, 2)

affinity(PE2, 1)

which would give PE1 2/3 probability of getting elected as a DF, in proportion to its relative bandwidth in the ES.

Depending on the chosen HRW hash function, affinity function MUST be extended to include bandwidth increment in the computation.

For e.g.,

affinity function specified in [EVPN-PER-MCAST-FLOW-DF] MUST be extended as follows to incorporate bandwidth increment j:

affinity(S,G,V, ESI, Address(i,j)) =
(1103515245.((1103515245.Address(i).j + 12345) XOR
D(S,G,V,ESI))+12345) (mod 2^31)

affinity or random function specified in [RFC8584] MUST be extended as follows to incorporate bandwidth increment j:

affinity(v, Es, Address(i,j)) = (1103515245((1103515245.Address(i).j + 12345) XOR D(v,Es))+12345)(mod 2^31)

6.4. BW Capability and Preference DF Election algorithm

This section applies to ES'es where all the PEs in the ES agree use the BW Capability with DF Type 2. The BW Capability modifies the Preference DF Election procedure [EVPN-DF-PREF], by adding the LBW value as a tie-breaker as follows:

Section 4.1, bullet (f) in [EVPN-DF-PREF] is updated to now consider the LBW value as below:

f) In case of equal Preference in two or more PEs in the ES, the tie-breakers will be the DP (Don't Preempt me) bit, the LBW value and the lowest IP PE in that order. For instance:

* If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=1, LBW=2000] in PE2, PE2 would be elected due to the DP bit.

* If vES1 parameters were [Pref=500,DP=0,LBW=1000] in PE1 and [Pref=500,DP=0, LBW=2000] in PE2, PE2 would be elected due to a higher LBW, even if PE1's IP address is lower.

* The LBW exchanged value has no impact on the Non-Revertive option described in [EVPN-DF-PREF].

6.5. Cost-Benefit Tradeoff on Link Failures

While incorporating link bandwidth into the DF election process provides optimal BUM traffic distribution across the ES links, it also implies that DF elections are re-adjusted on link failures or bandwidth changes. If the operator does not wish to have this level of churn in their DF election, then they should not advertise the BW capability. Not advertising BW capability may result in less than optimal BUM traffic distribution while still retaining the ability to allow an ingress PE to do weighted ECMP for its unicast traffic to a set of egress PEs.

7. Additional Considerations

7.1. Real-time Available Bandwidth

PE-CE link bandwidth availability may sometimes vary in real-time disproportionately across PE-CE links within a multi-homed ES due to various factors such as flow based hashing combined with fat flows and unbalanced hashing. Reacting to real-time available bandwidth is at this time outside the scope of this document.

7.2. Weighted Load-balancing to Multi-homed Subnets

EVPN Link bandwidth extended community may also be used to achieve unequal load-balancing of prefix routed traffic by including this extended community in EVPN Route Type 5. When included in EVPN RT-5, its value is to be interpreted as egress PE's relative weight for the prefix included in this RT-5. Ingress PE will then compute the forwarding path-list for the prefix route using weighted paths received from each egress PE. EVPN Link bandwidth extended community MUST be encoded with "Value-Units = 0x01" to signal a generalized weight associated with the advertising PE.

7.3. Weighted Load-balancing without EVPN aliasing

[RFC7432] defines per-[ES, EVI] RT-1 based EVPN aliasing procedure as an optional propcedure. In an unlikely scenario where an EVPN implementation does not support EVPN aliasing procedures, MAC forwarding path-list at the ingress PE is computed based on per-ES RT-1 and RT-2 routes received from egress PEs instead of per-ES RT-1 and per-[ES, EVI] RT-1 from egress PEs. In such a case, only the weights received via per-ES RT-1 from the egress PEs included in the MAC path-list are to be considered for weighted path-list computation.

7.4. EVPN IRB Multi-homing With Non-EVPN routing

EVPN-LAG based multi-homing on an IRB gateway may also be deployed together with non-EVPN routing, such as global routing or an L3VPN routing control plane. Key property that differentiates this set of use cases from EVPN IRB use cases discussed earlier is that EVPN control plane is used only to enable LAG interface based multi-homing and not as an overlay VPN control plane. Applicability of weighted ECMP procedures specified in this document to these set of use cases is an area of further consideration beyond the scope of this document.

8. Operational Considerations

* In order for the solution specified in this document to function correctly, implementation SHOULD ensure that EVPN Link Bandwidth Extended Comminiuty is being advertised with same "Value-Units" across all PEs.

* Further, when a generalized weight option is used with "Value-Units = 0x1", implementation SHOULD ensure that the weights are assigned to each PE in a consistent manner.

* Implementation SHOULD alert the users via syslog when an inconsistency in "Value-Units" is detected across the PE set for a given ESI or prefix.

* Implementation SHOULD also alert users via syslog if an unreasonable discrepancy is detected across advertised BW or weights from different PEs, such that the implementation is unable to compute a weighted pathlist that can be programmed in hardware. This could likely result from inconsistent units of weight used by different PEs.

* Operators MAY monitor the traffic flow distribution and DF election distribution across the egress PE set to ensure that the implementation is working as expected.

9. Security Considerations

Security considerations discussed in [RFC7432] and [RFC8584] apply to this document. Methods described in this document further extend signaling of multi-homed devices using ESI LAG. They are hence subject to same considerations if the control plane or data plane was to be compromised. As an example, if control plane is compromised, signaling of heavily skewed Link Bandwidth Attributes could result in all traffic to be directed towards one PE resulting in its host facing links to be overloaded. Exposure to such an attack is limited by suggested syslogs discussed in Operational Consideration section. Considerations for protecting control and data plane described in [RFC7432] are equally applicable to signaling of Link Bandwidth Attribute defined in this document.

10. IANA Considerations

[RFC8584] defines a new extended community for PEs within a redundancy group to signal and agree on uniform DF Election Type and Capabilities for each ES. This document requests IANA to allocate a bit in the "DF Election capabilities" registry setup by [RFC8584] with the following suggested bit number:

Bit 4: BW (Bandwidth Weighted DF Election)

A new EVPN Link Bandwidth extended community is defined to signal local ES link bandwidth to ingress PEs. This extended community is defined of type 0x06 (EVPN Extended Community Sub-Types). IANA has assigned a sub-type value of 0x10 for the EVPN Link bandwidth extended community, of type 0x06 (EVPN Extended Community Sub-Types). EVPN Link Bandwidth extended community is defined as transitive.

IANA is requested to set up a registry called "Value-Units" for the 1-octet field in the EVPN Link Bandwidth Extended Community. New registrations will be made through the "RFC Required" procedure defined in [RFC8126]. The following are suggested initial values in that registry exist:

Value	Name	Reference
0	Weight in units of Mbps	This document
1	Generalized Weight	This document
2-255	Unassigned	

11. Acknowledgements

Authors would like to thank Satya Mohanty for valuable review and inputs with respect to HRW and weighted HRW algorithm refinements specified in this document. Authors would also like to thank Bruno Decraene and Sergey Fomin for valuable review and comments.

12. Contributors

Satya Ranjan Mohanty
Cisco Systems
US
Email: satyamoh@cisco.com

13. References

13.1. Normative References

[EVPN-DF-PREF]
Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., Mohanty, S., and , "Preference-based EVPN DF Election", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-pref-df-06, 19 June 2020, <https://tools.ietf.org/html/draft-ietf-bess-evpn-pref-df-06.txt>.

[EVPN-PER-MCAST-FLOW-DF]
Sajassi, A., mishra, m., Thoria, S., Rabadan, J., and J. Drake, "Per multicast flow Designated Forwarder Election for EVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-per-mcast-flow-df-election-04, 31 August 2020, <http://www.ietf.org/internet-drafts/draft-ietf-bess-evpn-per-mcast-flow-df-election-04.txt>.

[EVPN-VIRTUAL-ES]
Sajassi, A., Brissette, P., Schell, R., Drake, J., Rabadan, J., and , "EVPN Virtual Ethernet Segment", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-virtual-eth-segment-06, 9 March 2020, <https://tools.ietf.org/html/draft-ietf-bess-evpn-virtual-eth-segment-06.txt>.

[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>.

[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>.

[RFC7814] Xu, X., Jacquenet, C., Raszuk, R., Boyes, T., and B. Fee, "Virtual Subnet: A BGP/MPLS IP VPN-Based Subnet Extension Solution", RFC 7814, DOI 10.17487/RFC7814, March 2016, <https://tools.ietf.org/html/rfc7814>.

[RFC8584] Rabadan, J., Ed., Mohanty, R., Sajassi, N., Drake, A., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <https://www.rfc-editor.org/info/rfc8584>.

13.2. Informative References

[BGP-LINK-BW]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", Work in Progress, Internet-Draft, draft-ietf-idr-link-bandwidth-07, March 2019, <https://tools.ietf.org/html/draft-ietf-idr-link-bandwidth-07.txt>.

[RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in EVPN", RFC 9135, DOI 10.17487/RFC9135, October 2021, <https://www.rfc-editor.org/rfc/rfc9135>.

Appendix A. BGP-Link-Bandwidth-Extended-Community

Link bandwidth extended community described in [BGP-LINK-BW] for layer 3 VPNs was considered for re-use here. This Link bandwidth extended community is however defined in [BGP-LINK-BW] as optional non-transitive. Since it is not possible to change deployed behavior of extended community defined in [BGP-LINK-BW], it was decided to define a new one. In inter-AS scenarios, link-bandwidth needs to be signaled to eBGP neighbors. When signaled across AS boundary, this extended community can be used to achieve optimal load-balancing towards egress PEs in a different AS. This is applicable both when next-hop is changed or unchanged across AS boundaries.

Authors' Addresses

Neeraj Malhotra (editor)
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: nmalhotr@cisco.com

Ali Sajassi
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: sajassi@cisco.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043
United States of America
Email: jorge.rabadan@nokia.com

John Drake
Juniper
Email: jdrake@juniper.net

Avinash Lingala
ATT
200 S. Laurel Avenue
Middletown, CA 07748
United States of America
Email: ar977m@att.com

Samir Thoria
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: sthoria@cisco.com

Internet Engineering Task Force
Internet-Draft
Updates: 6790, 7447 (if approved)
Intended status: Standards Track
Expires: 2 September 2024

B. Decraene, Ed.
Orange
J. G. Scudder, Ed.
Juniper Networks
W. Henderickx
Nokia
K. Kompella
Juniper Networks
S. Mohanty
Cisco Systems
J. Uttaro
Independent Contributor
B. Wen
Comcast

1 March 2024

BGP Next Hop Dependent Capabilities Attribute
draft-ietf-idr-entropy-label-14

Abstract

RFC 5492 allows a BGP speaker to advertise its capabilities to its peer. When a route is propagated beyond the immediate peer, it is useful to allow certain capabilities, or other properties, to be conveyed further. In particular, it is useful to advertise forwarding plane features.

This specification defines a BGP transitive attribute to carry such capability information, the "Next Hop Dependent Capabilities Attribute," or NHC. Unlike the capabilities defined by RFC 5492, those conveyed in the NHC apply solely to the routes advertised by the BGP UPDATE that contains the particular NHC.

This specification also defines an NHC capability that can be used to advertise the ability to process the MPLS Entropy Label as an egress LSR for all NLRI advertised in the BGP UPDATE. It updates RFC 6790 and RFC 7447 concerning this BGP signaling.

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/.

Table of Contents

1. Introduction

[RFC5492] allows a Border Gateway Protocol (BGP) speaker to advertise its capabilities to its peer. When a route is propagated beyond the immediate peer, it is useful to allow certain capabilities, or other properties, to be conveyed further. In particular, it may be useful to advertise forwarding plane features.

This specification defines a BGP optional transitive attribute to carry such capability information, the "Next Hop Dependent Capabilities Attribute", or NHC. (Note that this specification should not be confused with RFC 5492 BGP Capabilities.)

Since the NHC is intended chiefly for conveying information about forwarding plane features, it needs to be regenerated whenever the BGP route's next hop is changed. Since owing to the properties of BGP transitive attributes this can't be guaranteed (an intermediate router that doesn't implement this specification would be expected to propagate the NHC as opaque data), the NHC encodes the next hop of its originator, or the router that most recently updated the attribute. If the NHC passes through a router that changes the next hop without regenerating the NHC, they will fail to match when later examined, and the recipient can act accordingly. This scheme allows NHC support to be introduced into a network incrementally. Informally, the intent is that,

* If a router is not changing the next hop, it can obliviously propagate the NHC just like any other optional transitive attribute.

* If a router is changing the next hop, then it has to be able to vouch for every capability it includes in the NHC.

Complete details are provided in Section 2.

An NHC carried in a given BGP UPDATE message conveys information that relates to all Network Layer Reachability Information (NLRI) advertised in that particular UPDATE, and only to those NLRI. A different UPDATE message originated by the same source might not include an NHC, and if so, NLRI carried in that UPDATE would not be affected by the NHC. By implication, if a router wishes to use NHC to describe all NLRI it originates, it needs to include an NHC with each UPDATE it sends. In this respect, despite its similar naming, the NHC is unlike RFC 5492 BGP Capabilities.

Informally, a capability included in a given NHC should not be thought of as a capability of the next hop, but rather a capability of the path, that depends on the ability of the next hop to support it. Hence it is said to be "dependent on" the next hop.

This specification also defines an NHC capability, called "ELCv3", to advertise the ability to process the Multiprotocol Label Switching (MPLS) Entropy Label as an egress Label Switching Router (LSR) for all NLRI advertised in the BGP UPDATE. It updates [RFC6790] and [RFC7447] with regard to this BGP signaling, this is further discussed in Section 3. Although ELCv3 is only relevant to NLRI of labeled address families, a future NHC capability might be applicable to non-labeled NLRI, or to both, irrespective of labels. (The term "labeled address family" is defined in the first paragraph of Section 3.5 of [RFC9012]. In this document, we use the term "labeled NLRI" as a short form of "NLRI of a labeled address family.")

1.1. 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.

2. BGP Next Hop Dependent Capabilities Attribute

2.1. Encoding

The BGP Next Hop Dependent Capabilities attribute (NHC attribute, or just NHC) is an optional, transitive BGP path attribute with type code 39. The NHC always includes a network layer address identifying the next hop of the route the NHC accompanies. The NHC signals potentially useful information related to the forwarding plane features, so it is desirable to make it transitive to ensure propagation across BGP speakers (e.g., route reflectors) that do not change the next hop and are therefore not in the forwarding path. The next hop data is to ensure correctness if it traverses BGP speakers that do not understand the NHC. This is further explained below.

The Attribute Data field of the NHC attribute is encoded as a header portion that identifies the router that created or most recently updated the attribute, followed by one or more Type-Length-Value (TLV) triples:

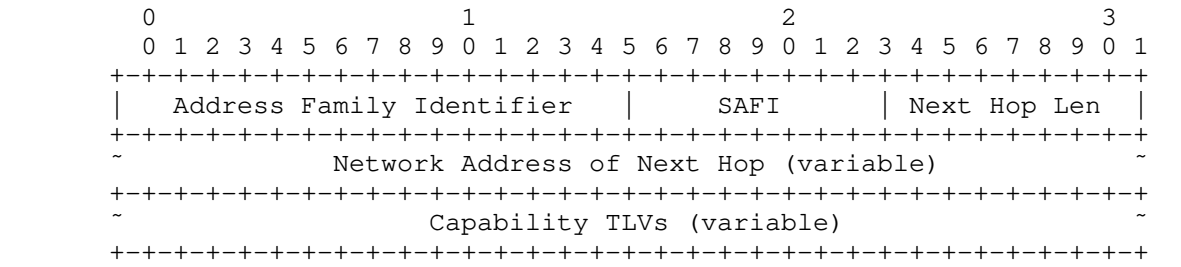


Figure 1: NHC Format

The meanings of the header fields (Address Family Identifier, SAFI or Subsequent Address Family Identifier, Length of Next Hop, and Network Address of Next Hop) are as given in Section 3 of [RFC4760].

In turn, each Capability is a TLV:

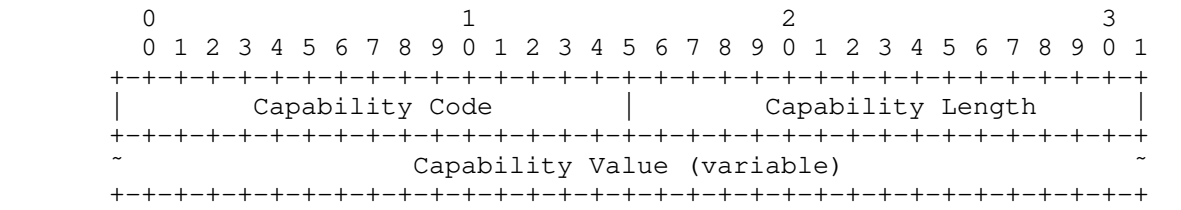


Figure 2: Capability TLV Format

Capability Code: a two-octet unsigned integer that indicates the type of capability advertised and unambiguously identifies an individual capability.

Capability Length: a two-octet unsigned integer that indicates the length, in octets, of the Capability Value field. A length of 0 indicates that the Capability Value field is zero-length, i.e. it has a null value.

Capability Value: a variable-length field. It is interpreted according to the value of the Capability Code.

A BGP speaker MUST NOT include more than one instance of a capability with the same Capability Code, Capability Length, and Capability Value. Note, however, that processing multiple instances of such a capability does not require special handling, as additional instances do not change the meaning of the announced capability; thus, a BGP speaker MUST be prepared to accept such multiple instances.

BGP speakers MAY include more than one instance of a capability (as identified by the Capability Code) with different Capability Value. Processing of these capability instances is specific to the Capability Code and MUST be described in the document introducing the new capability.

Capability TLVs MUST be placed in the NHC in increasing order of Capability Code. (In the event of multiple instances of a capability with the same Capability Code as discussed above, no further sorting order is defined here.) Although the major sorting order is mandated, an implementation MUST elect to be prepared to consume capabilities in any order, for robustness reasons.

2.2. Sending the NHC

Suppose a BGP speaker S has a route R it wishes to advertise with next hop N to its peer.

If S is originating R into BGP, it MAY include an NHC attribute with it, that carries capability TLVs that describe aspects of R. S MUST set the next hop depicted in the header portion of the NHC to be equal to N, using the encoding given above.

If S has received R from some other BGP speaker, two possibilities exist. First, S could be propagating R without changing N. In that case, S does not need to take any special action, it SHOULD simply propagate the NHC unchanged unless specifically configured otherwise. Indeed, we observe that this is no different from the default action a BGP speaker takes with an unrecognized optional transitive attribute -- it is treated as opaque data and propagated.

Second, S could be changing R in some way, and in particular, it could be changing N. If S has changed N it MUST NOT propagate the NHC unchanged. It SHOULD include a newly-constructed NHC attribute with R, constructed as described above in the "originating R into BGP" case. Any given capability TLV carried by the newly-constructed NHC attribute might use information from the received NHC attribute as input to its construction, possibly as straightforwardly as simply copying the TLV. The details of how the capabilities in the new NHC are constructed are specific to the definition of each capability. Any capability TLVs received by S that are for capabilities not supported by S will not be included in the newly-constructed NHC attribute S includes with R.

An implementation SHOULD propagate the NHC and its contained capabilities by default. An implementation SHOULD provide configuration control of whether any given capability is propagated. An implementation MAY provide finer-grained control on propagation based on attributes of the peering session, as discussed in Section 6.1.

Due to the nature of BGP optional transitive path attributes, any BGP speaker that does not implement this specification will propagate the NHC, the requirements of this section notwithstanding. Such a speaker will not update the NHC, however.

Certain NLRI formats do not include a next hop at all, one example being the Flow Specification NLRI [RFC8955]. The NHC MUST NOT be sent with such NLRI.

2.2.1. Aggregation

When aggregating routes, the above rules for constructing a new NHC MUST be followed. The decision of whether to include the NHC with the aggregate route and what its form will be, depends in turn on whether any capabilities are eligible to be included with the aggregate route. If there are no eligible capabilities, the NHC MUST NOT be included.

The specification for an individual capability must define how that capability is to be aggregated. If no rules are defined for a given capability, that capability MUST NOT be aggregated. Rules for aggregating the ELCv3 are found in Section 3.2.1.

(Route aggregation is described in [RFC4271]. Although prefix aggregation -- combining two or more more-specific prefixes to form one less-specific prefix -- is one application of aggregation, we note that another is when two or more routes for the same prefix are selected to be used for multipath forwarding.)

2.3. Receiving the NHC

An implementation receiving routes with a NHC SHOULD NOT discard the attribute or its contained capabilities by default. An implementation SHOULD provide configuration control of whether any given capability is processed. An implementation MAY provide finer-grained control on propagation based on attributes of the peering session, as discussed in Section 6.1.

When a BGP speaker receives a BGP route that includes the NHC, it MUST compare the address given in the header portion of the NHC and illustrated in Figure 1 to the next hop of the BGP route. If the two

match, the NHC may be further processed. If the two do not match, it means some intermediate BGP speaker that handled the route in transit both does not support NHC, and changed the next hop of the route. In this case, the contents of the NHC cannot be used, and the NHC MUST be discarded without further processing, except that the contents MAY be logged.

In considering whether the next hop "matches", a semantic match is sought. While bit-for-bit equality is a trivial test of matching, there may be certain cases where the two are not bit-for-bit equal, but still "match". An example is when an MP_REACH Next Hop encodes both a global and a link-local IPv6 address. In that case, the link-local address might be removed during Internal BGP (IBGP) propagation, the two would still be considered to match if they were equal on the global part. See Section 3 of [RFC2545].

A BGP speaker receiving a Capability Code that it supports behaves as defined in the document defining the Capability Code. A BGP speaker receiving a Capability Code that it does not support MUST ignore that Capability Code. In particular, the receipt of an unrecognized Capability Code MUST NOT be handled as an error.

The presence of a capability SHOULD NOT influence route selection or route preference, unless tunneling is used to reach the BGP next hop, the selected route has been learned from External BGP (that is, the next hop is in a different Autonomous System), or by configuration (see following). Indeed, it is in general impossible for a node to know that all BGP routers of the Autonomous System (AS) will understand a given capability, and if different routers within an AS were to use a different preference for a route, forwarding loops could result unless tunneling is used to reach the BGP next hop. Following this reasoning, if the administrator of the network is confident that all routers within the AS will interpret the presence of the capability in the same way, they could relax this restriction by configuration.

2.4. Attribute Error Handling

An NHC is considered malformed if the length of the attribute, encoded in the Attribute Length field of the BGP Path Attribute header (Section 4.3 of [RFC4271]), is inconsistent with the lengths of the contained capability TLVs. In other words, the sum of the sizes (Capability Length plus 4) of the contained capability TLVs, plus the length of the NHC header (Figure 1), must be equal to the overall Attribute Length.

A BGP UPDATE message with a malformed NHC SHALL be handled using the approach of "attribute discard" defined in [RFC7606].

Unknown Capability Codes MUST NOT be considered to be an error.

An NHC that contains no capability TLVs MAY be considered malformed, although it is observed that the prescribed behavior of "attribute discard" is semantically no different from that of having no TLVs to process. There is no reason to propagate an NHC that contains no capability TLVs.

A document that specifies a new NHC Capability should provide specifics regarding what constitutes an error for that NHC Capability.

If a capability TLV is malformed, that capability TLV SHOULD be ignored and removed. Other capability TLVs SHOULD be processed as usual. If a given capability TLV requires different error-handling treatment than described in the previous sentences, its specification should provide specifics.

2.5. Network Operation Considerations

In the corner case where multiple nodes use the same IP address as their BGP next hop, such as with anycast nodes as described in [RFC4786], a BGP speaker MUST NOT advertise a given capability unless all nodes sharing this same IP address support this capability. The network operator operating those anycast nodes is responsible for ensuring that an anycast node does not advertise a capability not supported by all nodes sharing this anycast address. The means for accomplishing this are beyond the scope of this document.

3. Entropy Label Capability (ELCv3)

The foregoing sections define the NHC as a container for capability TLVs. The Entropy Label Capability is one such capability.

When BGP [RFC4271] is used for distributing labeled NLRI as described in, for example, [RFC8277], the route may include the ELCv3 as part of the NHC. The inclusion of this capability with a route indicates that the egress of the associated Label Switched Path (LSP) can process entropy labels as an egress LSR for that route -- see Section 4.1 of [RFC6790]. Below, we refer to this for brevity as being "EL-capable."

For historical reasons, this capability is referred to as "ELCv3", to distinguish it from the prior Entropy Label Capability (ELC) defined in [RFC6790] and deprecated in [RFC7447], and the ELCv2 described in [I-D.scudder-bgp-entropy-label].

This section (and its subsections) replaces Section 5.2 of [RFC6790], which was previously deprecated by [RFC7447].

3.1. Encoding

The ELCv3 has capability code 1, capability length 0, and carries no value:

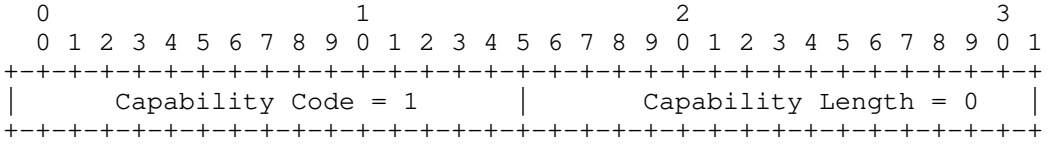


Figure 3: ELCv3 TLV Format

3.2. Sending the ELCv3

When a BGP speaker S has a route R it wishes to advertise with next hop N to its peer, it MAY include the ELCv3 capability if it knows that the egress of the associated LSP L is EL-capable, otherwise it MUST NOT include the ELCv3 capability. Specific conditions where S would know that the egress is EL-capable are if S:

* Is itself the egress, and knows itself to be EL-capable, or
* Is re-advertising a BGP route it received with a valid ELCv3 capability, and is preserving the value of N as received, or
* Is re-advertising a BGP route it received with a valid ELCv3 capability, and is changing the next hop that it has received to N, and knows that this new next hop (normally itself) is EL-capable, or
* Is re-advertising a BGP route it received with a valid ELCv3 capability, and is changing the next hop that it has received to N, and knows (for example, through configuration) that the new next hop (normally itself) even if not EL-capable will simply swap labels without popping the BGP-advertised label stack and processing the label below, as with a transit LSR, or
* Knows by implementation-specific means that the egress is EL-capable, or
* Is redistributing a route learned from another protocol, and that other protocol conveyed the knowledge that the egress of L was EL-capable. (For example, this might be known through the Label Distribution Protocol (LDP) ELC TLV, Section 5.1 of [RFC6790].)

The ELCv3 MAY be advertised with routes that are labeled, such as those using SAFI 4 [RFC8277]. It MUST NOT be advertised with unlabeled routes.

3.2.1. Aggregation

When forming an aggregate (see Section 2.2.1), the aggregate route thus formed MUST NOT include the ELCv3 unless each constituent route would be eligible to include the ELCv3 according to the criteria given above.

3.3. Receiving the ELCv3

(Below, we assume that "includes the ELCv3" implies that the containing NHC has passed the checks specified in Section 2.3. If it had not passed, then the NHC would have been discarded and the ELCv3 would be deemed not to have been included.)

When a BGP speaker receives an unlabeled route that includes the ELCv3, it MUST discard the ELCv3.

When a BGP speaker receives a labeled route that includes the ELCv3, it indicates that it can safely insert an entropy label into the label stack of the associated LSP. This implies that the receiving BGP speaker if acting as ingress, MAY insert an entropy label as per Section 4.2 of [RFC6790].

3.4. ELCv3 Error Handling

The ELCv3 is considered malformed and must be disregarded if its length is other than zero.

If more than one instance of the ELCv3 is included in an NHC, instances beyond the first MUST be disregarded.

4. Legacy ELC

The ELCv3 functionality introduced in this document replaces the "BGP Entropy Label Capability Attribute" (ELC attribute) that was introduced by [RFC6790], and deprecated by [RFC7447]. The latter RFC specifies that the ELC attribute, BGP path attribute 28, "MUST be treated as any other unrecognized optional, transitive attribute". This specification revises that requirement.

As the current specification was developed, it became clear that due to incompatibilities between how the ELC attribute is processed by different fielded implementations, the most prudent handling of attribute 28 is not to propagate it as an unrecognized optional,

transitive attribute, but to discard it. Therefore, this specification updates [RFC7447], by instead requiring that an implementation that receives the ELC attribute MUST discard any received ELC attribute.

5. IANA Considerations

IANA has made a temporary allocation in the BGP Path Attributes registry of the Border Gateway Protocol (BGP) Parameters group. IANA is requested to make this allocation permanent, and to update its name and reference as shown below.

Value	Code	Reference
39	BGP Next Hop Dependent Capabilities (NHC)	(this doc)

Table 1

IANA is requested to create a new registry called "BGP Next Hop Dependent Capability Codes" within the Border Gateway Protocol (BGP) Parameters group. The registry's allocation policy is First Come, First Served, except where designated otherwise in Table 2. It is seeded with the following values:

Value	Description	Reference	Change Controller
0	reserved	(this doc)	IETF
1	ELCv3	(this doc)	IETF
2	NNHN	draft-wang-idr-next-next-hop-nodes-00	kfwang@juniper.net
65400 - 65499	private use	(this doc)	IETF
65500 - 65534	reserved for experimental use	(this doc)	IETF
65535	reserved	(this doc)	IETF

Table 2

6. Security Considerations

6.1. Considerations for the NHC

The header portion of the NHC contains the next hop the attribute's originator included when sending it, or that an intermediate router included when updating the attribute (in the latter case, the "contract" with the intermediate router is that it performed the checks in Section 2.3 before propagating the attribute). This will typically be an IP address of the router in question. This may be an infrastructure address the network operator does not intend to announce beyond the border of its Autonomous System, and it may even be considered in some weak sense, confidential information.

A motivating application for this attribute is to convey information between Autonomous Systems that are under the control of the same administrator. In such a case, it would not need to be sent to other Autonomous Systems. At time of writing, work [I-D.uttaro-idr-bgp-oad] is underway to standardize a method of distinguishing between the two categories of external Autonomous Systems, and if such a distinction is available, an implementation can take advantage of it by constraining the NHC and its contained capabilities to only propagate by default to and from the former category of Autonomous Systems. If such a distinction is not available, a network operator may prefer to configure routers peering with Autonomous Systems not under their administrative control to not send or accept the NHC or its contained capabilities, unless there is an identified need to do so.

The foregoing notwithstanding, control of NHC propagation can't be guaranteed in all cases -- if a border router doesn't implement this specification, the attribute, like all BGP optional transitive attributes, will propagate to neighboring Autonomous Systems. (This can be seen as a specific case of the general "attribute escape" phenomenon discussed in [I-D.haas-idr-bgp-attribute-escape].) Similarly, if a border router receiving the attribute from an external Autonomous System doesn't implement this specification, it will store and propagate the attribute, the requirements of Section 2.3 notwithstanding. So, sometimes this information could leak beyond its intended scope. (Note that it will only propagate as far as the first router that does support this specification, at which point it will typically be discarded due to a non-matching next hop, per Section 2.3.)

If the attribute leaks beyond its intended scope, capabilities within it would potentially be exposed. Specifications for individual capabilities should consider the consequences of such unintended exposure, and should identify any necessary constraints on propagation.

6.2. Considerations for the ELCv3 Capability

Insertion of an ELCv3 by an attacker could cause forwarding to fail. Deletion of an ELCv3 by an attacker could cause one path in the network to be overutilized and another to be underutilized. However, we note that an attacker able to accomplish either of these (below, an "on-path attacker") could equally insert or remove any other BGP path attribute or message. The former attack described above denies service for a given route, which can be accomplished by an on-path attacker in any number of ways even absent ELCv3. The latter attack defeats an optimization but nothing more; it seems dubious that an attacker would go to the trouble of doing so rather than launching some more damaging attack.

7. References

7.1. 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/rfc/rfc2119>.

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, DOI 10.17487/RFC2545, March 1999, <https://www.rfc-editor.org/rfc/rfc2545>.

[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/rfc/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/rfc/rfc4760>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://www.rfc-editor.org/rfc/rfc6790>.

[RFC7447] Scudder, J. and K. Kompella, "Deprecation of BGP Entropy Label Capability Attribute", RFC 7447, DOI 10.17487/RFC7447, February 2015, <https://www.rfc-editor.org/rfc/rfc7447>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/rfc/rfc7606>.

[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/rfc/rfc8174>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/rfc/rfc9012>.

7.2. Informative References

[I-D.haas-idr-bgp-attribute-escape]
Haas, J., "BGP Attribute Escape", Work in Progress, Internet-Draft, draft-haas-idr-bgp-attribute-escape-01, 2 February 2024, <https://datatracker.ietf.org/doc/html/draft-haas-idr-bgp-attribute-escape-01>.

[I-D.ietf-idr-next-hop-capability]
Decraene, B., Kompella, K., and W. Henderickx, "BGP Next-Hop dependent capabilities", Work in Progress, Internet-Draft, draft-ietf-idr-next-hop-capability-08, 8 June 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-next-hop-capability-08>.

[I-D.scudder-bgp-entropy-label]
Scudder, J. and K. Kompella, "BGP Entropy Label Capability, Version 2", Work in Progress, Internet-Draft, draft-scudder-bgp-entropy-label-00, 28 April 2022, <https://datatracker.ietf.org/doc/html/draft-scudder-bgp-entropy-label-00>.

[I-D.uttaro-idr-bgp-oad]
Uttaro, J., Retana, A., Mohapatra, P., Patel, K., and B. Wen, "One Administrative Domain using BGP", Work in Progress, Internet-Draft, draft-uttaro-idr-bgp-oad-03, 10 January 2024, <https://datatracker.ietf.org/doc/html/draft-uttaro-idr-bgp-oad-03>.

[RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, DOI 10.17487/RFC4786, December 2006, <https://www.rfc-editor.org/rfc/rfc4786>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/rfc/rfc5492>.

[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/rfc/rfc8277>.

[RFC8955] Loibl, C., Hares, S., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", RFC 8955, DOI 10.17487/RFC8955, December 2020, <https://www.rfc-editor.org/rfc/rfc8955>.

Acknowledgements

The authors of this specification thank Randy Bush, Mach Chen, Wes Hardaker, Jeff Haas, Susan Hares, Ketan Talaulikar, and Gyan Mishra for their review and comments.

This specification derives from two earlier documents, [I-D.ietf-idr-next-hop-capability] and [I-D.scudder-bgp-entropy-label].

[I-D.ietf-idr-next-hop-capability] included the following acknowledgements:

> The Entropy Label Next-Hop Capability defined in this document is based on the ELC BGP attribute defined in section 5.2 of [RFC6790].
>
> The authors wish to thank John Scudder for the discussions on this topic and Eric Rosen for his in-depth review of this document.
>
> The authors wish to thank Jie Dong and Robert Raszuk for their review and comments.

[I-D.scudder-bgp-entropy-label] included the following acknowledgements:

> Thanks to Swadesh Agrawal, Alia Atlas, Bruno Decraene, Martin Djernaes, John Drake, Adrian Farrell, Keyur Patel, Toby Rees, and Ravi Singh, for their discussion of this issue.

Contributors

Serge Krier
Cisco Systems
Email: sekrier@cisco.com

Kevin Wang
Juniper Networks
Email: kfwang@juniper.net

Authors' Addresses

Bruno Decraene (editor)
Orange
Email: bruno.decraene@orange.com

John G. Scudder (editor)
Juniper Networks
Email: jgs@juniper.net

Wim Henderickx
Nokia
Email: wim.henderickx@nokia.com

Kireeti Kompella
Juniper Networks
Email: kireeti@juniper.net

Satya Mohanty
Cisco Systems
Email: satyamoh@cisco.com

James Uttaro
Independent Contributor
Email: juttaro@ieee.org

Bin Wen
Comcast
Email: Bin_Wen@comcast.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 12 February 2025

K. Vairavakkalai, Ed.
M. Jeyananth
Juniper Networks, Inc.
M. Nanduri
Microsoft
Lingala
AT&T

11 August 2024

BGP MultiNexthop Attribute
draft-ietf-idr-multinexthop-attribute-01

Abstract

Today, a BGP speaker can advertise one nexthop for a set of NLRIs in an Update message. This nexthop can be encoded in either the top-level BGP-Nexthop attribute (code 3), or inside the MP_REACH_NLRI attribute (code 14). Forwarding information related to the nexthop is scattered across various attributes, extended communities or the NLRI field.

This document defines a new optional non-transitive BGP attribute called "MultiNexthop (MNH)" with IANA BGP attribute type code TBD, that can be used to carry an ordered set of one or more Nexthops in the same Update message, with all forwarding information being carried in one attribute, scoped on a per nexthop basis.

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

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 12 February 2025.

Table of Contents

1. Introduction

Today, a BGP speaker can advertise one nexthop for a set of NLRIs in an Update message. This nexthop can be encoded in either the top-level BGP-Nexthop attribute (code 3), or inside the MP_REACH_NLRI attribute (code 14). Forwarding information related to the nexthop is scattered across various attributes, extended communities or the NLRI field.

This document defines a new optional non-transitive BGP attribute called "MultiNexthop (MNH)" with IANA BGP attribute type code TBD, that can be used to carry an ordered set of one or more Nexthops in the same route, with all forwarding information being carried in one attribute, scoped on a per nexthop basis.

2. Terminology

iSN: Ingress Service Node

eSN: Egress Service Node

NLRI: Network Layer Reachability Information

AFI: Address Family Identifier

SAFI: Subsequent Address Family Identifier

PE : Provider Edge

RT : Route-Target extended community

RD : Route-Distinguisher

MPLS: Multi Protocol Label Switching

ECMP: Equal Cost Multi Path

WECMP: Weighted Equal Cost Multi Path

FRR: Fast Re Route

PNH : Protocol Next hop address carried in a BGP Update message

MNH: BGP MultiNextHop attribute

NFI: Nexthop Forwarding Information

FI: Forwarding Instruction

FA: Forwarding Argument

2.1. Definitions

MULTI_NEXT_HOP (aka MNH): BGP MultiNexthop attribute. The new attribute defined by this document.

MNH TLV: Top level TLV contained in a MULTI_NEXT_HOP.

NFI TLV: Nexthop Forwarding Information TLV, contained in a MNH TLV.

FI TLV: Forwarding Instruction TLV, contained in a NFI TLV.

FA TLV: Forwarding Argument TLV, contained as an argument to a FI in the FI TLV.

3. Motivation

Today, in a BGP Update, forwarding information related to the BGP nexthop is scattered across various attributes, extended communities or the NLRI field. On some other address families like Flowspec, nexthop address is carried without using the nexthop attribute, in one or more extended communities of specific type. It may be desirable to carry them scoped in a single attribute.

It may be desirable to carry the forwarding information for a nexthop scoped in a single attribute, and uniformly for all address families.

For cases where multiple nexthops need to be advertised, BGP Addpath [RFC7911] is used with some address families. Though Addpath allows basic ability to advertise multiple routes, it does not allow the sender to express the desired relationship between the multiple nexthops being advertised e.g., relative ordering, type of load balancing, fast reroute. These are local decisions based on configuration and path selection at the receiving node. Also, Addpath does not consider things like Link-bandwidth community when selecting add-path routes. Some scenarios (explained in Appendix A) may benefit from having a mechanism, where egress node can signal multiple nexthops along with their relationship to ingress nodes.

It would be desirable to have a common way to carry more than one nexthop on a BGP route of any family, and express relationship between them.

This document defines a new optional non-transitive BGP attribute "MultiNexthop (MNH)" that can be used for these purposes. The MNH attribute can be used in any BGP family that wants to carry one or more nexthops, with all forwarding information being carried in one attribute, scoped on a per nexthop basis.

E.g. The MNH can be used to advertise MPLS label along with nexthop for labeled and unlabeled families (e.g. Inet Unicast, Inet6 Unicast, Flowspec) alike. Such that, mechanisms at the transport layer can work uniformly on labeled and unlabled BGP families to realize various usecases.

4. Base Encoding And Protocol Procedures

"MultiNexthop (MNH)" is a new BGP optional non-transitive attribute (code TBD), that can be used to carry an ordered set of one or more Nexthops in the same route, with all forwarding information being carried in one attribute, scoped on a per nexthop basis. This attribute describes forwarding instructions using TLVs as shown below.

This section describes the organization and encoding of the MNH attribute.

```
MNH Attribute: {
      PrimaryPath {
          [Forwarding Instruction 1],
           ..
          [Forwarding Instruction n]
      }
      BackupPath {
          [Forwarding Instruction 1],
           ..
          [Forwarding Instruction n]
      }
}

Forwarding Instruction: {
    {FwdAction, Forwarding Arguments}
}
```

Figure 1: Overview of MNH Attribute Layout - Eye candy summary

A MNH attribute consists of a Header and one or more "MNH TLVs".

A MNH TLV contains a Type and one unit of "Nexthop Forwarding Information" (NFI TLV).

A NFI TLV contains one or more "Forwarding Instructions" (FI TLVs).

A FI TLV contains a "Forwarding Action" code and one more "Forwarding Arguments" (FA TLVs).

The FA TLVs describe the parameters required to complete a "Forwarding Action".

4.1. MultiNexthop Attribute

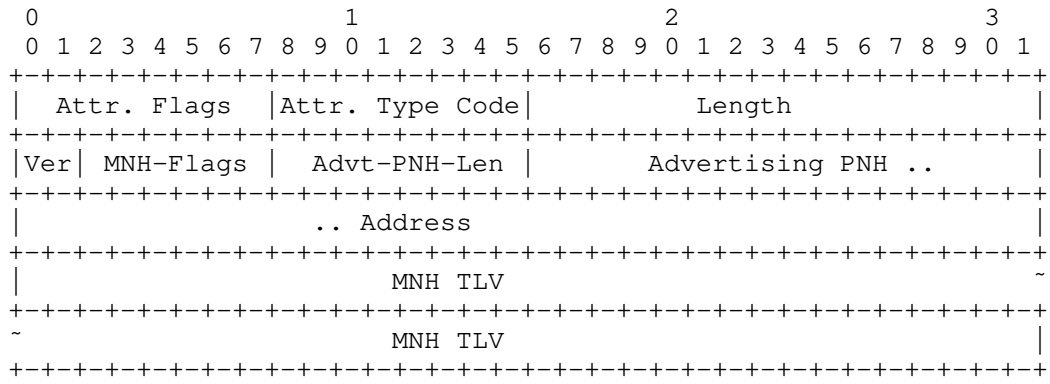


Figure 2: MultiNexthop - BGP Attribute

MNH Header:

- Attr. Flags (1 octet)
  BGP Path-attribute flags. indicating an Optional Non-Transitive attribute. i.e. Optional bit set, Transitive bit reset.

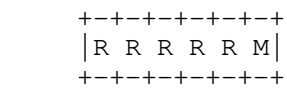
- Attr. Type Code (1 octet)
  Type code allotted by IANA. TBD.

- Length (1 or 2 octets)
  One or Two bytes field stating length of attribute value in bytes.

- Version (2 bits)
  MNH Version - indicates layout of the MNH header.
  Set to 0x0 indicating "MNH v0", which is defined in this document.

  If there is any significant changes to the skeletal layout of MNH attribute in future, this Version field will be useful.

- MNH Flags (6 bits)



  6 bits following the Version bits are MNH Flags.

  M: "Mandatory".
  Value 1 indicates that this MNH attribute is mandatory.
  If this MNH attr is invalid, the route is Unusable Hidden.

  R: Reserved. MUST be set to zero, SHOULD be ignored by receiver.

- Advt-PNH-Len (1 octet)
  Length in octets (4 for IPv4, 16 for IPv6, 12 for VPN-IPv4, 24 for VPN-IPv6) of Advertising PNH Address.

- Advertising PNH Address (Advt-PNH-Len octets)
  BGP Protocol Nexthop address advertised in NEXT_HOP or MP_REACH_NLRI attr. Used to sanity-check the MNH attribute. In case of RFC-2545, this will be the global (non link-local) IPv6 address.

MNH TLVs: One or more MNH TLVs are carried in a MNH attr.
MNH TLV is described in subsequent sections.

4.1.1. Processing the MNH Header

A BGP speaker MUST fill MNH Version field to 0.

If a MNH is received with a Version other than 0, the MNH attribute MUST be considered invalid, and be treated as Unrecognized Non-transitive attribute.

The "Advertising PNH" field is validated as described in Section 4.1.2

4.1.2. Validation of MNH against Nexthop

When adding a MultiNexthop attribute to an advertised BGP route, the speaker MUST put the same next-hop address in the Advertising PNH field as it put in the Nexthop field inside MP_REACH_NLRI attribute (code 14) if one exists, or the NEXT_HOP attribute (code 3).

A speaker that adds a new MNH attribute to the advertised BGP route, it MUST record in the "Advertising PNH" field the same next-hop address as used in MP_REACH_NLRI attribute if one exists, or the NEXT_HOP attribute.

A speaker receiving a MNH attribute SHOULD ignore it if the next-hop address contained in 'Advertising PNH' field is not the same as the nexthop address contained in MP_REACH_NLRI attribute if one exists, or the NEXT_HOP attribute. [RFC7606] 'Attribute Discard' approach is used.

In case of [RFC2545], the global (non link-local) IPv6 address should be used for this purpose.

As specified in [RFC7606]BGP update message can contain no more than one instance of MP_REACH attribute or NEXT_HOP attribute. Similarly, a BGP update MUST contain no more than one instance of MNH attribute. If the MNH attribute (whether recognized or unrecognized) appears more than once in an UPDATE message, then all the occurrences of the attribute other than the first one SHALL be discarded and the UPDATE message will continue to be processed. The anamoly MAY be logged for diagnosis.

4.1.3. Scope of Use, Origination and Propagation

The MNH attribute is intended to be used in a BGP free core, between egress and ingress BGP speakers that understand this attribute. These BGP speakers may have an intra-AS or inter-AS BGP session between them. On propagating the route with nexthop altered, a new MNH attribute MAY be added by the advertising speaker.

To avoid un-intentionally leaking the MNH to another AS, via a BGP speaker that does not understand MNH attribute, it is defined as "optional non-transitive". But this also means that a RR needs to be upgraded to support this attribute before any PEs in the network can make use of it.

Use of MNH on a BGP session is disabled by default. An implementation MUST provide configuration control on a per BGP neighbor address family basis, to enable MNH support.

A BGP speaker MUST NOT advertise MNH on a BGP route if MNH support is not enabled for the corresponding address family on the advertising BGP session.

If the MNH attribute is received on a BGP session where MNH support is not enabled, the attribute MUST be treated as Unrecognized non-transitive. This rule provides additional protection against unintended propagation of this attribute, when both BGP speakers understand MNH but receiver has not enabled the support. A RFC3392 Capability is not used for this purpose, because it would cause session reset whenever 'MNH support' config is changed.

Remaining text in this section apply when both receiving and advertising BGP sessions are enabled with MNH support.

When a BGP speaker receives the MNH attribute on a BGP route, and re-advertises it with the nexthop unchanged, it MUST propagate the attribute unchanged. E.g. a Route Reflector.

When a BGP speaker receives the MNH attribute on a BGP route, and re-advertises it with the nexthop altered, it processes the attribute but MUST NOT propagate it as-is. The BGP speaker MAY however attach a new instance of MNH attribute on the re-advertised route, and MAY derive its value from the received MNH.

A BGP speaker re-advertising a BGP route with nexthop unchanged MAY add the MNH attribute on the reflected BGP route, on behalf of the originating BGP speaker. The "Advertising PNH field" in the MNH is set to the Nexthop field in BGP route being re-advertised.

4.1.4. Error Handling

A TLV or sub-TLV of a certain Type in a MNH attribute can occur only once, unless specified otherwise by that type value. If multiple instances of such TLV or sub-TLV is received, the instances other than the first occurance are ignored.

If processing of a received MNH attribute resulted in an error, then the "M bit" is used to decide the action. If the M bit is 0, then the MNH attribute is ignored, [RFC7606] 'Attribute Discard' approach MUST be used, and continue to process rest of the update. If M bit is 1, then the BGP Route containing the MNH MUST be considered Unusable.

MNH employs a hierarchical error detection mechanism, where an error in lower layer TLVs is percolated upwards to the MNH attribute, based on the M bit.

Implementations MAY provide policy configuration to set M bit to 0 on a MNH attribute being added, this helps with testing impact of the MNH on receiving nodes. Once confident, the MNH attribute can be re-advertised with M bit set. This helps in graceful incremental deployment.

The definition of a certain type of TLV or Sub-TLV in the MNH should specify in it's procedures, the value of M bit to be used. An implementation MAY provide configuration to set or reset the M bit.

4.2. MNH TLV

The type of MNH TLV describes how the forwarding information carried in the MNH TLV is used.

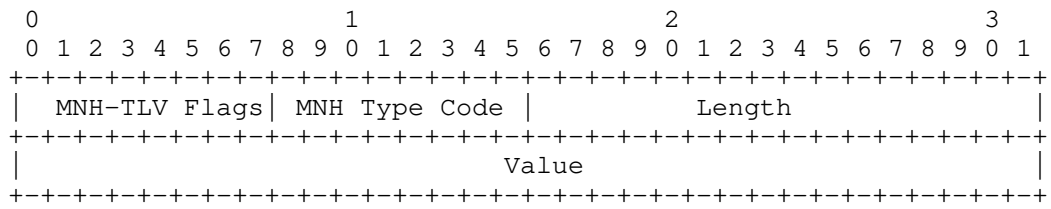
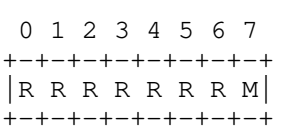


Figure 3: MNH TLV

- MNH-TLV Flags (1 octet)



All bits are reserved.

M: "Mandatory".
Value 1 indicates that this MNH TLV is mandatory.
If this MNH TLV is not understood, the MNH attribute containing it is considered invalid.

R: "Reserved".
MUST be set to zero, SHOULD be ignored by receiver.

This document defines the following MNH TLV types:

- MNH Type Code (1 octet)
Type of MNH TLV. 0 is Reserved.

- Length
Length of Value portion in octects.

- Value
Value portion contains the NFI TLV.

A sending BGP speaker advertises the information for one ore more nexthops in a MNH TLV.

Information received in MNH TLV is used to create the Forwarding state at receiving BGP speaker.

The MNH Type code indicates how the information carried in the TLV is used at the receiving node.

4.2.1. Error Handling

If invalid Type Code 0 is received, the TLV is ignored irrespective of "M bit", and continue to process rest of the update.

If the received Type Code is incompatible for the prefix in BGP NLRI, the TLV is considered invalid.

If an unrecognized Type Code is received, or processing of a recognized MNH TLV type results in an error, the TLV is considered invalid.

Invalid TLV is handled based on the "M bit" on the TLV.

If the M bit is 0, then the TLV is ignored and continue to process rest of the update. If M bit is 1, then the MNH attribute is considered invalid, triggering the procedures in Section 4.1.4.

4.3. Nexthop Forwarding Information TLV

A Nexthop Forwarding Information TLV describes a MNH TLV. It contains one or more Forwarding Instruction TLVs. These Forwarding Instructions are the Forwarding Legs of the MNH.

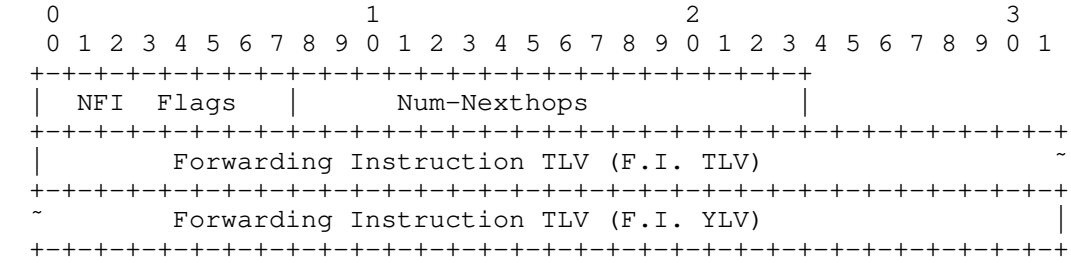
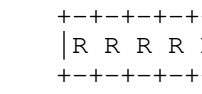


Figure 4: Nexthop Forwarding Information TLV

- NFI Flags (1 octet)



M: "Mandatory".
Value 1 indicates that this NFI TLV (Nexthop Leg) is mandatory. If this Nexthop Leg is not understood, the MNH TLV containing it is considered invalid.

R: "Reserved".
MUST be set to zero, SHOULD be ignored by receiver.

- Num-Nexthops
  Number of F.I. TLVs.

- Forwarding Instruction TLV
  Each F.I. TLV describes a Nexthop Leg.
  Layout of Forwarding Instruction TLV is described in next section.

M bit on a NFI TLV SHOULD be set to 1.

4.3.1. Error Handling

If Num-Nexthops in a received NFI is 0, it is considered invalid. Irrespective of M bit value, the NFI TLV is ignored and remaining update is processed.

The receiving BGP speaker MAY consider the "Num-Nexthops" value in a Nexthop Forwarding Information TLV not acceptable, based on it's forwarding capabilities or local policy. In such cases, the NFI TLV is considered Invalid.

An Invalid NFI TLV is handled based on value of M bit on it. If the M bit is 0, the NFI TLV is ignored, and remaining update continue to be processed. If M bit is 1, the MNH TLV carrying this NFI is considered Invalid, triggering the procedures in Section 4.2.1.

4.4. Forwarding Instruction TLV

Each Forwarding Instruction TLV describes a Nexthop Leg. It expresses a "Forwarding Action" (FwdAction) along with arguments required to complete the action. The type of actions defined by this TLV are given below. The arguments are denoted by "Forwarding Argument TLVs". The Forwarding Argument TLVs takes appropriate values based on the FwdAction.

Each FwdAction should note the Arguments needed to complete the action. Any extranous arguments should be ignored. If the minimum set of arguments required to complete an action is not received, the Forwarding Instruction TLV should be ignored. Appropriate logging and diagnostic info MAY be provided by an implementation to help troubleshoot such scenarios.

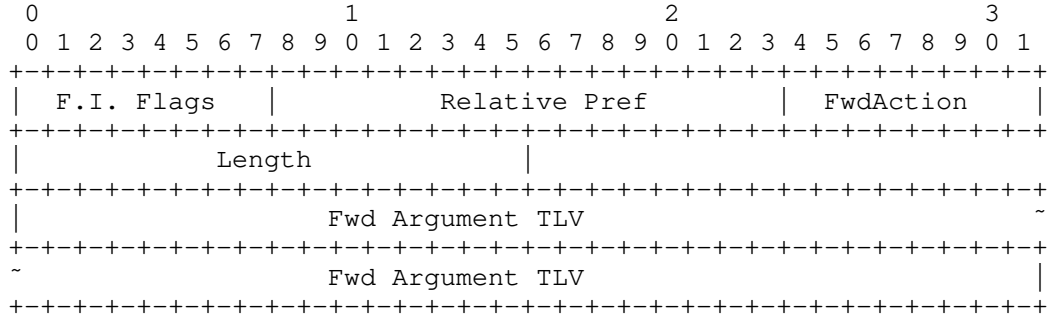
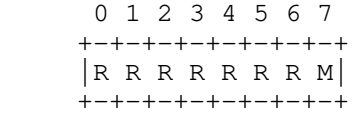


Figure 5: Forwarding Instruction TLV

- F.I. Flags (1 octet)



M: "Mandatory".
Value 1 indicates that this Forwarding Instruction is mandatory. If this instruction is not understood, the NFI TLV containing it is considered invalid.

R: "Reserved".
MUST be set to zero, SHOULD be ignored by receiver.

- Relative Pref (2 octets)

Unsigned 2 octet integer specifying relative order or preference, among the many forwarding instructions, to use in FIB. All usable nexthop legs with lowest relative-pref are installed in FIB as primary-path. Thus if multiple legs exist with that lowest relative-pref, ECMP is formed.

- FwdAction (1 octet)

Type Code denoting the Forwarding action to be performed by receiving node. 0 is Reserved.

- Length (2 octets)

Length in octets, of all Forwarding Argument TLVs.

Definition of a Forwarding Action should specify the set of forwarding arguments required to execute the action, and value of M bit.

4.4.1. Error Handling

If an Invalid value of 0 is received as FwdAction, the TLV is ignored irrespective of "M bit", and continue to process rest of the update..

If an unrecognized or unsupported FwdAction is received, the FI TLV is considered Invalid.

If a certain Forwarding Action is unable to be executed because the set of required arguments are not available, the FI TLV is considered Invalid. If a certain Forwarding Action is applied to an incompatible NLRI, the FI TLV is considered Invalid.

An Invalid FI TLV is handled based on value of M bit on it. If the M bit is 0, the FI TLV is ignored, and remaining update continue to be processed. If M bit is 1, the NFI TLV carrying this NFI is considered Invalid, triggering the procedures in Section 4.3.1.

4.5. Forwarding Argument TLV

The Forwarding Argument TLV describes various parameters required to execute a FwdAction.

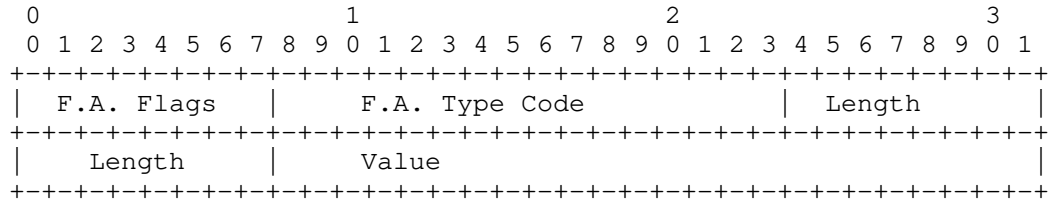
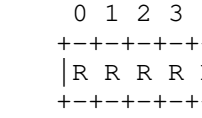


Figure 6: Forwarding Argument TLV

- F.A. Flags (1 octet)



M: "Mandatory".
Value 1 indicates that this argument is mandatory for the Forwardin
g Action.
If this argument is not understood, the FI TLV containing it
is considered invalid.

C: "Cummulative".
Request nodes to accumulate value in re-advertised MNH.
By default Forwarding Arguments are not commulative, so C bit is 0
unless otherwise specified by the forwarding argument type.

E: "Egress Attached".
This bit is maintained when C bit is set to 1.
E bit is set to 1 if a cummulative argument is being added to a rou
te
with empty AS-path.

R: "Reserved".
MUST be set to zero, SHOULD be ignored by receiver.

- F.A. Type Code (2 octets)

Type Code of Forwarding Argument. 0 is Reserved.

- Length (2 octets)

Length in bytes of Value field.

The C bit is set to 1 on attributes that need to be accumulated across BGP nexthop-self propagation hops. If a received MNH has a FA with C bit 1, it MUST be set to 1 on the FA inserted in any advertised MNH also. The value of the FA in the advertised MNH MAY be derived from the value of the FA in the received MNH. The specific FA SHOULD define the procedure on how the accumulation of value happens for the specific type of FA.

If a received MNH has a FA with C bit 1, and receiving speaker is unable to perform the accumulation of FA, it MUST NOT include the FA type in any advertised MNH.

A FA that need to be accumulated end-to-end may want to know if the cummulative value denotes the path until the Egress node. The E bit denotes that the FA was originated by the Egress node that originated this BGP route. The E bit is set to 1 by the node adding the FA, if the AS-path on the route is empty. The E bit value received on a MNH MUST be propagated on the MNH added to the re-advertisement. This allows the Ingress node to see the E bit value set by the Egress node.

4.5.1. Error Handling

If an Invalid F.A. Type Code value of 0 is received, the TLV is ignored irrespective of "M bit", and continue to process rest of the update..

If an unrecognized F.A. Type Code is received, the FA TLV is considered Invalid.

An Invalid FA TLV is handled based on value of M bit on it. If the M bit is 0, the FA TLV is ignored, and remaining update continue to be processed. If M bit is 1, the FI TLV carrying this FA is considered Invalid, triggering the procedures in Section 4.4.1.

4.6. Interaction with Addpath

[ADDPATH-GUIDELINES] suggests the following:

"Diverse path: A BGP path associated with a different BGP next-hop and BGP router than some other set of paths. The BGP router associated with a path is inferred from the ORIGINATOR_ID attribute or, if there is none, the BGP Identifier of the peer that advertised the path."

When selecting "diverse paths" for ADD_PATH as specified above, the MNH attribute should also be compared if it exists, to determine if two routes have "different BGP next-hop".

4.7. Path Selection Considerations

4.7.1. Determining IGP Cost

While tie breaking in the path-selection as described in [RFC4271], 9.1.2.2. step (e) viz. the "IGP cost to nexthop", consider the highest cost among the nexthop-legs present in this attribute.

The IGP cost thus calculated is also used when constructing AIGP TLV ([RFC7311])

5. TLVs Defined In This Document

This section describes the initial set of MNH TLVs, Forwarding Instructions and Arguments that this document defines.

5.1. MNH TLVs

The type of MNH TLV describes how the forwarding information carried in the MNH TLV is used.

This document defines the following MNH TLV types:

MNH Type Code	Meaning
0	Reserved
1	Primary forwarding path
2	Backup forwarding path

- Length
  Length of Value portion in octects.

- Value
  Value portion contains the NFI TLV.

Type codes 1 and 2 are applicable for upstream allocated prefixes, example IP, Upstream MPLS labels, Flowspec routes.

Note that usage of Type code 1 in a BGP route containing IP prefix gives similar result as advertising the route with nexthop contained in BGP path-attributes: Nexthop (code 3) or MP_REACH_NLRI (code 14).

Upstream allocation for MPLS routes is achieved by using mechanisms explained in [MPLS-NAMESPACES].

If an invalid Type Code 0 is received, the TLV is ignored, and continue to process rest of the update.

If the received Type Code is incompatible for the prefix in BGP NLRI, the TLV should be ignored.

5.1.1. Primary Forwarding Path

This is a MNH TLV (Section 4.2) with MNH Type Code = 1, called "Primary Forwarding Path"

This TLV describes forwarding state to be programmed at receiving speaker as Primary Path nexthop leg. This TLV is used with Upstream allocated or global scope prefixes carried in BGP NLRI. Value part of this TLV contains Nexthop Forwarding Information TLV.

A BGP speaker uses the nexthop forwarding information received in this TLV as a primary path nexthop leg when programming the route for the NLRI prefix in its Forwarding table.

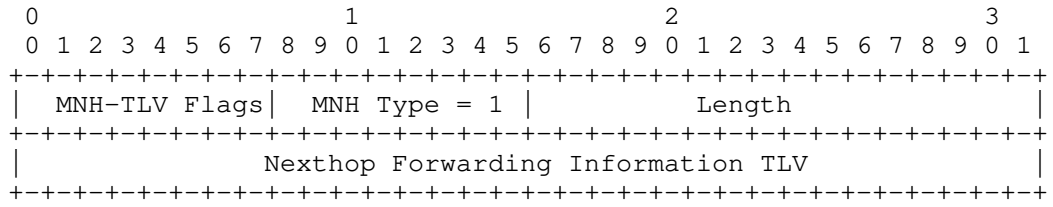


Figure 7: Primary forwarding path TLV

5.1.2. Repair Forwarding Path

This is a MNH TLV (Section 4.2) with MNH Type Code = 2, called "Repair Forwarding Path"

This TLV describes forwarding state to be programmed during traffic repair at receiving speaker. i.e. This TLV is used to program a backup path. This TLV is used with Upstream allocated prefixes or global scoped prefixes. Value part contains Nexthop Forwarding Information TLV.

Signaling a different nexthop for use as backup path is desirable in some labeled forwarding scenarios, where two multihomed edge devices use each other as backup path to protect traffic when primary path fails.

This is required to avoid label advertisement oscillation between the multihomed PEs when they implement per-nexthop label allocation mode.

The label advertised by a PE1 for primary path advertisement is allocated/forwarded using external paths as primary leg and backup-path label from other multihomed PE2 as backup-path label. Such that primary-path label allocation at PE1 is not a function of the primary-path label advertised by PE2. Thus the primary path label remains stable at a PE and does not change when a new primary path label is received from the other multihomed PE. This prevents the label oscillation problem.

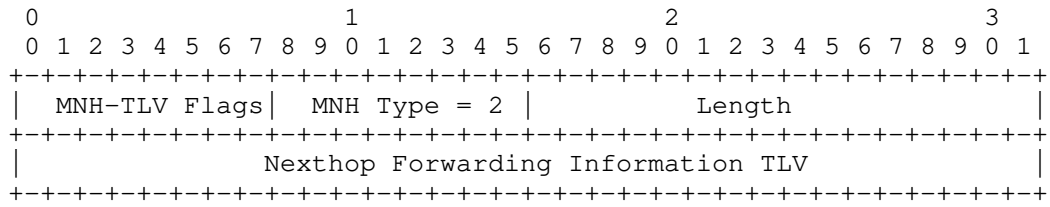


Figure 8: Repair forwarding path TLV

The backup path label allocated and advertised by a PE is a function of only the primary path. E.g. path to the CE device. So this label value does not change when a new label is received from the other multihomed PE

5.2. Forwarding Actions in FI TLV

Each Forwarding Instruction TLV describes a Nexthop Leg. It expresses a "Forwarding Action" (FwdAction) along with arguments required to complete the action. The type of actions defined by this TLV are given below. The arguments are denoted by "Forwarding Argument TLVs". The Forwarding Argument TLVs takes appropriate values based on the FwdAction.

Each FwdAction should note the Arguments needed to complete the action. Any extranous arguments should be ignored. If the minimum set of arguments required to complete an action is not received, the Forwarding Instruction TLV should be ignored. Appropriate logging and diagnostic info MAY be provided by an implementation to help troubleshoot such scenarios.

Following Forwarding Actions are defined by this document.

FwdAction	Meaning
0	Reserved
1	Forward
2	Pop-And-Forward
3	Swap
4	Push
5	Pop-And-Lookup
6	Replicate

Forwarding Instruction TLV with unknown FwdAction should be ignored, skipped and rest of the attribute processed; gracefully handling the error. The event may be appropriately logged for diagnosis.

- Length (2 octets)

Length in octets, of all Forwarding Argument TLVs.

Meaning of most of the above FwdAction semantics is well understood. FwdAction 1 is applicable for both IP and MPLS routes. FwdActions 2-5 are applicable for encapsulated payloads (like MPLS) only. FwdActions 1, 6 are applicable for Flowspec routes for Redirect and Mirror actions. FwdAction 6 can also be used to indicate multicast replication like functionality.

The "Forward" action means forward the IP/MPLS packet with the destination prefix (IP-dest-addr/MPLS-label) value unchanged. For IP routes, this is the forwarding-action given for next-hop addresses contained in BGP path-attributes: Nexthop (code 3) or MP_REACH_NLRI (code 14). For MPLS routes, usage of this action is equivalent to SWAP with same label-value; one such usage is explained in [MPLS-NAMESPACES] when Upstream-label-allocation is in use.

The "Pop-And-Forward" action means Pop the payload header (e.g. MPLS-label) and forward the payload towards the Nexthop IP-address specified in the Endpoint Id TLV, using appropriate encapsulation to reach the Nexthop.

When applied to MPLS packet, the "Pop-And-Lookup" action may result in a MPLS-lookup or an upper-layer header (like IPv4, IPv6) lookup, depending on whether the label that was popped was the bottom of stack label.

If an incompatible FwdAction is received for a prefix-type, or an unsupported FwdAction is received, it is considered a semantic-error and MUST be dealt with as explained in "Error handling procedures" section.

5.3. Forwarding Argument TLVs

The Forwarding Argument TLV describes various parameters required to execute a FwdAction.

Following types of Forwarding Argument are defined by this document.

F.A. Type Code	Meaning
0	Reserved
1	Endpoint Identifier
2	Path Constraints
3	Payload encapsulation info signaling
4	Endpoint attributes advertisement

- Length (2 octets)

  Length in bytes of Value field.

5.3.1. Endpoint Identifier

This is a Forwarding Argument (Section 4.5) with F.A. Type Code = 1. It identifies an Endpoint of certain type.

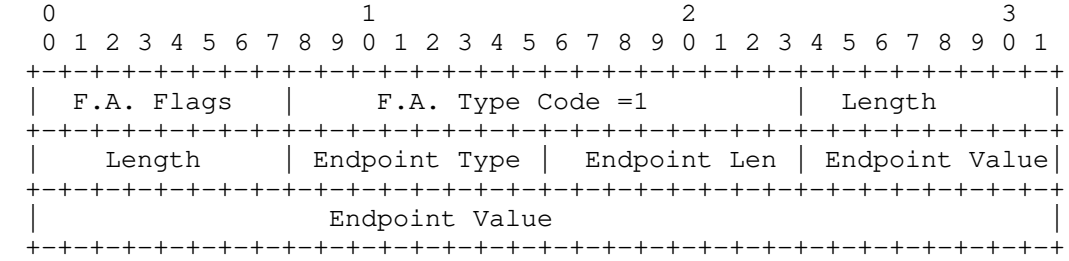


Figure 9: Endpoint Identifier

- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

Endpoint Type	Value	Len (octets)
0	Reserved	
1	IPv4 Address	4
2	IPv6 Address	16
3	MPLS Label (Upstream allocated or Global scope)	4
4	Fwd Context RD	8
5	Fwd Context RT	8

- Endpoint Len (1 octet)

  Length in bytes of Endpoint Value field.

5.3.2. Path Constraints

This is a Forwarding Argument (Section 4.5) with F.A. Type Code = 2. It defines Constraints for Path to the Endpoint..

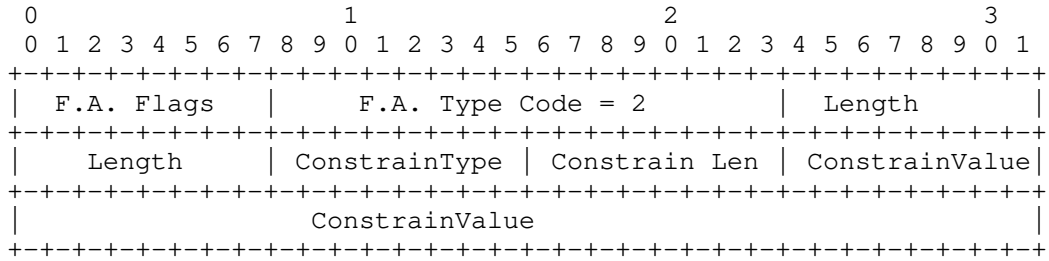


Figure 10: Path Constraints

- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

ConstrainType	Value	Len (octets)
0	Reserved	
1	Proximity check	2
2	Transport Class ID (Color)	4
3	Load balance factor	2

- Constrain Len (1 octet)

  Length in bytes of Constrain Value field.

- Proximity check Flags (2 octets)
  Flags describing whether the nexthop endpoint is expected to be single hop away, or multihop away. Format of flags is described in next section.

- Transport Class ID (Color):

  This is a 32 bit identifier, associated with the Nexthop address.
  The Nexthop IP-address specified in "Endpoint Identifier" TLVs
  are resolved over tunnels of this color.
  Defined in [BGP-CT] [draft-kaliraj-idr-bgp-classful-transport-planes]

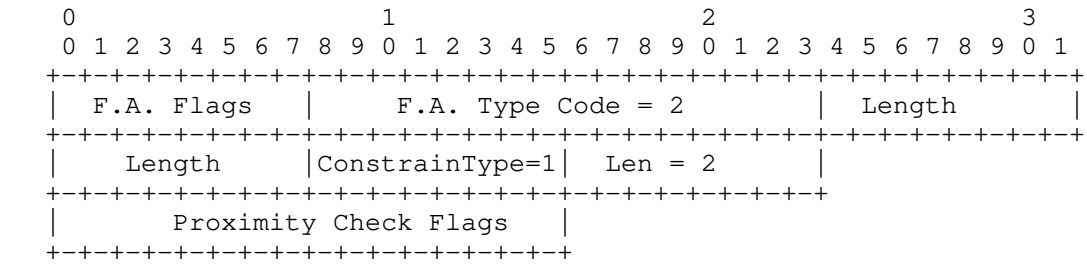
- Load balance factor (2 octets)
  Balance Percentage

5.3.2.1. Proximity Check

Usually EBGP singlehop received routes are expected to be one hop away, directly connected. And IBGP received routes are expected to be multihop away. Implementations today provide configuring exceptions to this rule.

The 'expected proximity' of the Nexthop can be signaled to the receiver using the Proximity check flags. Such that irrespective of whether the route is received from IBGP/EBGP peer, it can be treated as a single-hop away or multihop away nexthop.

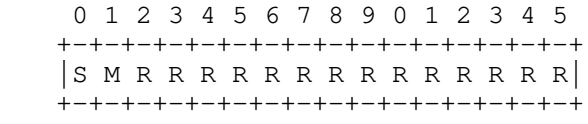
The format of the Proximity check Sub-TLV is as follows:



- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

- Proximity check Flags (2 octets)



S: Restrict to Singlehop path.
M: Expect Multihop path.
R: Reserved. MUST be set to zero, SHOULD be ignored by receiver.

Figure 11: Proximity check constrain

This TLV would be valid with Forwarding Instructions TLV with FwdAction of Forward, Pop-And-Forward, Swap or Push.

When S bit is set, receiver considers the nexthop valid only if it is directly connected to the receiver.

When M bit is set, receiver assumes that the nexthop can be multiple hops away, and resolves the path to the nexthop via another route.

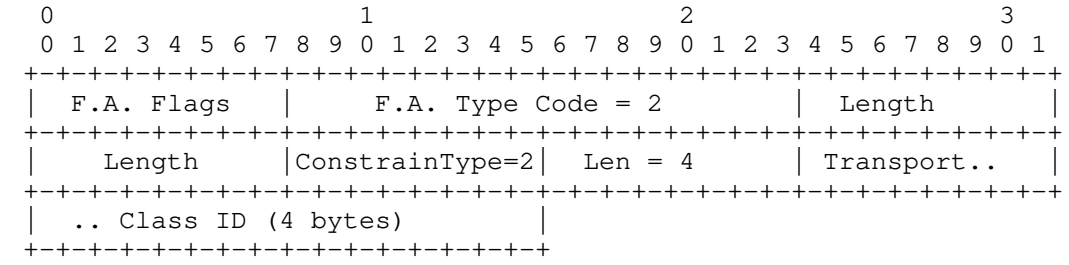
When both S and M bits are set, M bit behavior takes precedence. When both S and M bits are Clear, the current behavior of deriving proximity from peer type (EBGP is singlehop, IBGP is multihop) is followed.

5.3.2.2. Transport Class ID (Color)

The Nexthop can be associated with a Transport Class, so as to resolve a path that satisfies required Transport tunnel characteristics. Transport Class is defined in [BGP-CT]

Transport Class is a per-nexthop scoped attribute. Without MNH, the Transport class is applied to the nexthop IP-address encoded in the BGP-Nexthop attribute (code 3), or inside the MP_REACH_NLRI attribute (code 14). With MNH, the Transport Class can be specified per Nexthop-Leg (Forwarding Instruction TLV). It is applied to the IP-address encoded in the Endpoint Identifier TLV of type "IPv4 Address", "IPv6 Address" , "MPLS Label (Upstream allocated or Global scope)".

The format of the Transport Class ID Sub-TLV is as follows:



- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

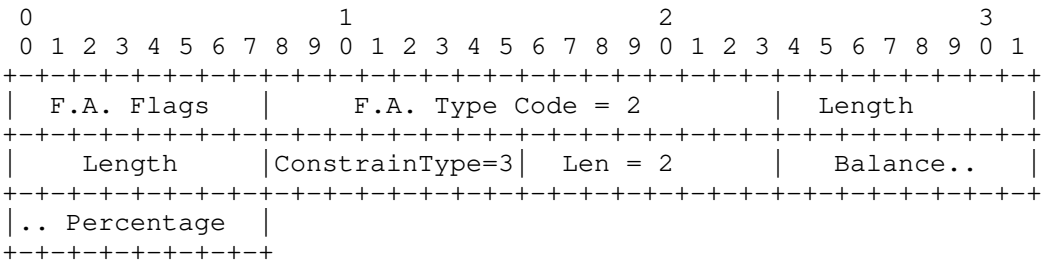
- Transport Class ID (Color):
  This is a 32 bit identifier, associated with the Nexthop address. The Nexthop specified in Endpoint Identifier TLVs are resolved over tunnels of this color.

Defined in [BGP-CT] [draft-kaliraj-idr-bgp-classful-transport-planes]

Figure 12: Transport Class ID (Color)

This TLV would be valid with Forwarding Instructions TLV with FwdAction of Forward, Swap or Push.

5.3.2.3. Load Balance Factor



- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

- Len (1 octet)
  Length of the Constrain Value field.

- Balance Percentage:
  This is the explicit "balance percentage" requested by the sender, for unequal load-balancing over these Nexthop-Descriptor-TLV legs. This balance percentage would override the implicit balance-percentage calculated using "Bandwidth" attribute sub-TLV.

Figure 13: Load Balance Factor

This sub-TLV would be valid with Forwarding Instructions TLV with FwdAction of Forward, Swap or Push.

This is the explicit "balance percentage" requested by the sender, for unequal load-balancing over these Nexthop-Descriptor-TLV legs. This balance percentage would override the implicit balance-percentage calculated using "Bandwidth" attribute sub-TLV

When the sum of "balance percentage" on the nexthop legs does not equal 100, it is scaled up or down to match 100. The individual balance percentages in each nexthop leg are also scaled up or down proportionally to determine the effective balance percentage per nexthop leg.

5.3.3. Payload Encapsulation Info

This is a Forwarding Argument (Section 4.5) with F.A. Type Code = 3. It defines Payload Encapsulation Information.

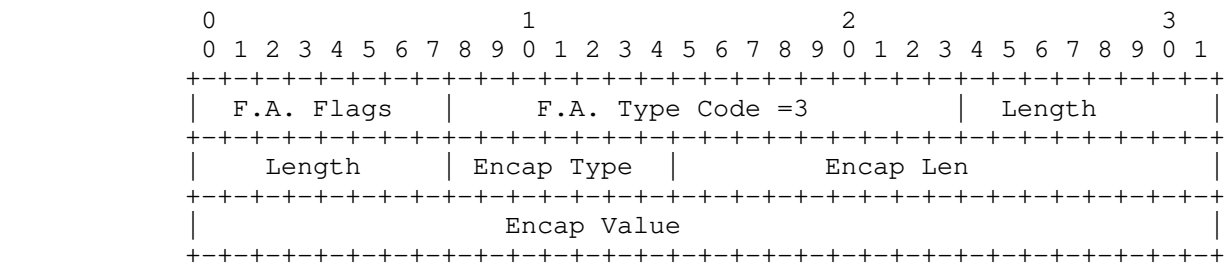


Figure 14: Payload Encapsulation Info

- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

Endcap Type	Value
0	Reserved
1	MPLS Label Info
2	SR MPLS label Index Info
3	SRv6 SID info
4	DSCP code point

- Encap Len (2 octets)

  Length in octets of Encap Value field.

5.3.3.1. MPLS Label Info

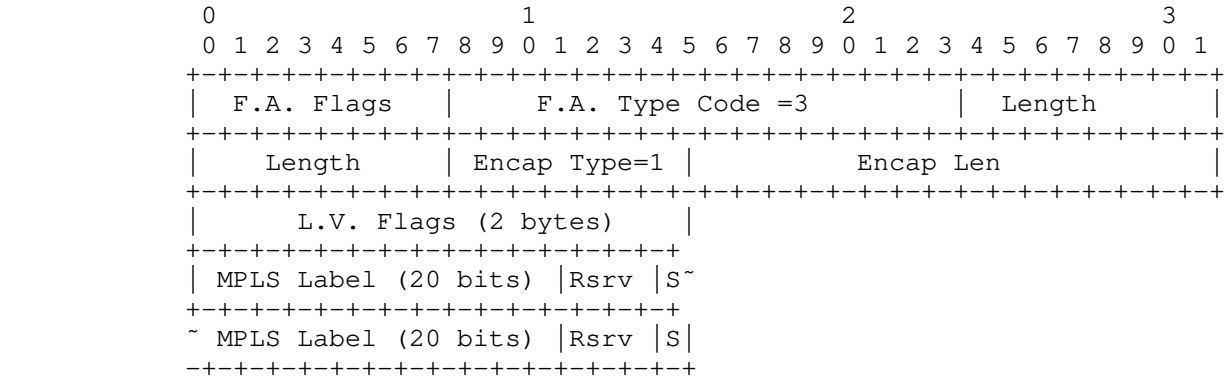


Figure 15: MPLS Label Info

- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

- Encap Type
  = 1, to signify MPLS Label Info.

- Encap Len (2 octets)
  Length in bytes of following Encap Value field.

- L.V. Flags (2 octets):

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|E R R R R R R R R R R R R R R R|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

  E: ELC bit. Indicates if this egress NH is Entropy Label Capable.
  1 means the Entropy Label capable.
  0 means not capable to handle Entropy Label.

  R: Reserved. MUST be set to zero, SHOULD be ignored by receiver.

- MPLS Label, Rsrv, S bit.
  20 bit MPLS Label stack encoded as in RFC 8277.
  S bit set on last label in label stack.

5.3.3.2. SR MPLS Label Index Info

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  F.A. Flags   |     F.A. Type Code =3         |  Length       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Length     | Encap Type=2 |            Encap Len           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   RESERVED    |       LI Flags                |    Label ..   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                ..Index                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 16: SR MPLS Label Index Info

- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

- Encap Type
  = 2, to signify SR MPLS SID Info.

- Encap Len (2 octets)
  Length in bytes of following Encap Value field.

Rest of the value portion is encoded as specified in RFC-8669 sec 3.1.

- RESERVED: 8-bit field. MUST be set to zero, SHOULD be ignored by receiver.

- LI Flags: 16 bits of flags. None defined. MUST be set to zero, SHOULD be ign
ored by receiver.

- Label Index:
  32-bit value representing the index value in the SRGB space.

5.3.3.3. SRv6 SID Info

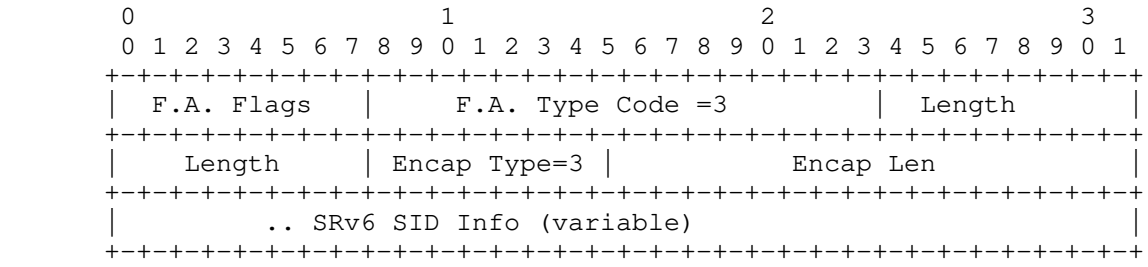


Figure 17: SRv6 SID Info

- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

- Encap Type
  = 3, to signify SRv6 SID Info.

- Encap Len (2 octets)
  Length in bytes of following Encap Value field.

- SRv6 SID Info:
  SRv6 SID Information, as specified in RFC-9252 sec 3.1.

5.3.3.4. DSCP

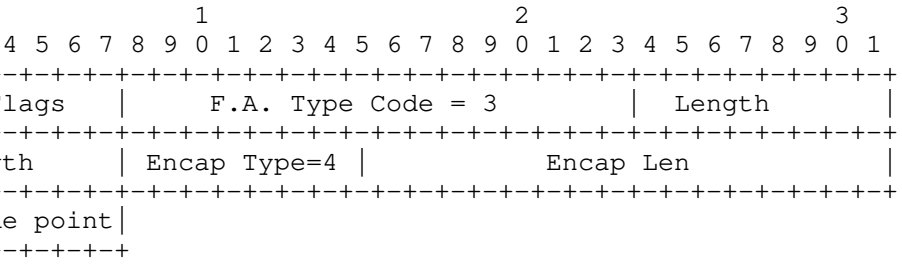


Figure 18: DSCP

- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

- Length (2 octets)
  Length in bytes of Value field.

- Encap Type
  = 4, to signify DSCP code point.

- Encap Len (2 octets)
  = 1, Length in bytes of following Encap Value field.

- DSCP code point:
  DS Field, as specified in RFC-2474 sec 3.

5.3.4. Endpoint Attributes

This is a Forwarding Argument (Section 4.5) with F.A. Type Code = 4. It defines Attributes of an Endpoint.

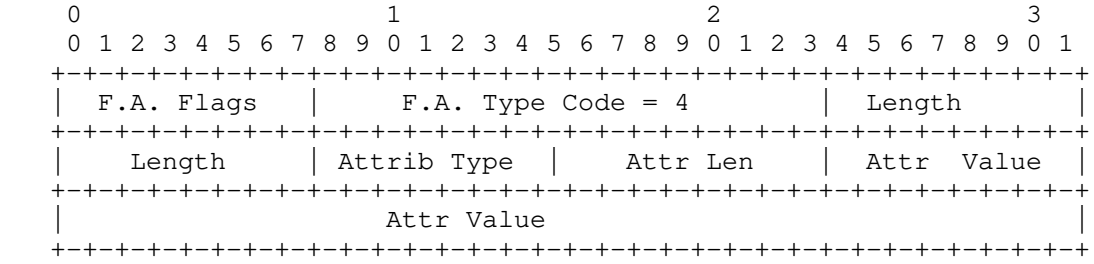
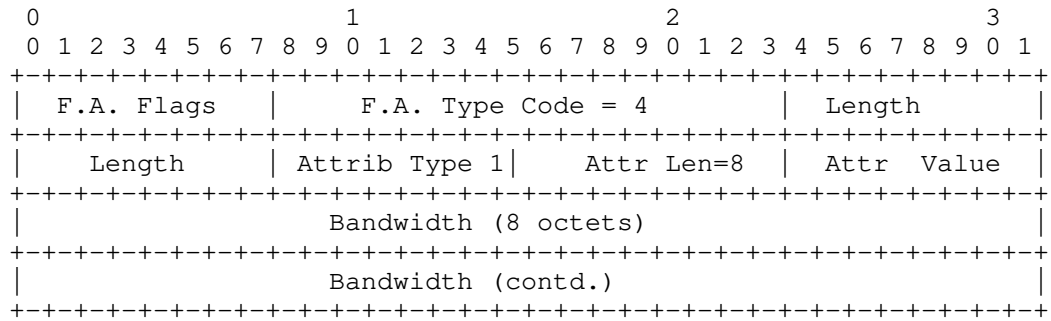


Figure 19: Endpoint attributes

EP Attrib Type	Attrib Value	Attrib Len (octets)
0	None	
1	Endpoint Bandwidth	8
2	Accumuated Metric	Variable

5.3.4.1. Endpoint Bandwidth



- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

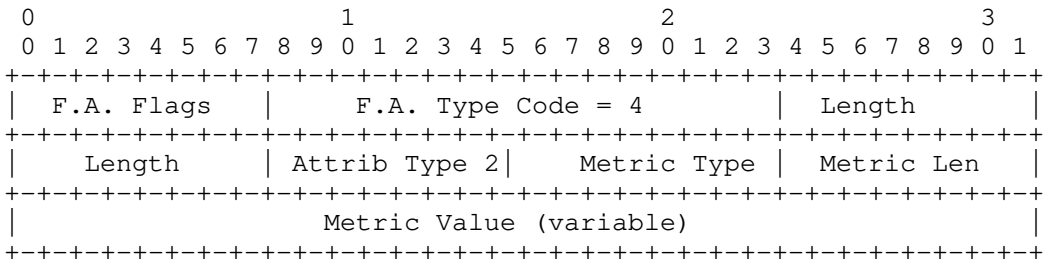
- Len (2 octets)
  Length in bytes of remaining portion of SubTLV.

- Bandwidth
  The bandwidth to the endpoint expressed as 8 octets, units being bits per second.

Figure 20: Endpoint Bandwidth

This sub-TLV would be valid with Forwarding Instruction TLV with FwdAction of Forward, Swap or Push.

5.3.4.2. Accumulated Metric to Endpoint



- F.A. Flags (1 octet)
  As defined in Forwarding Argument TLV.

    C: Cummulative bit is set to 1 by originator of this argument.

- Len (2 octets)
  Length in bytes of remaining portion of SubTLV.

- Metric Type: Type from "IGP Metric-Type" IANA registry under IGP Parameters
  Following types are defined by this document to be accumulated:
    0 IGP Metric
    1 Min Unidirectional Link Delay as defined in [RFC8570, Section 4.2]

- Metric Len: Length in octets of Metric Value field.
    IGP Metric: 4
    Min Unidirectional Link Delay: 4

- Metric Value:
    IGP Metric: 4 octet Accumulated IGP cost
    Min Unidirectional Link Delay: 4 octet Accumulated min delay in microseconds.

Figure 21: Accumulated Metric to Endpoint

This sub-TLV would be valid with Forwarding Instruction TLV with FwdAction of Forward, Swap or Push.

6. Scaling Considerations

The MNH attribute allows receiving multiple nexthops on the same BGP session. This flexibility also opens up the possibility that a peer can send large number of multipath (ECMP/UCMP/FRR) nexthops that may overwhelm the local system's forwarding plane. Prefix-limit based checks will not avoid this situation.

To keep the scaling limits under check, a BGP speaker MAY keep account of number of unique multipath nexthops that are received from a BGP peer, and impose a configurable max-limit on that. This is especially useful for EBGP peers.

A good scaling property of conveying multipath nexthops using the MNH attribute with N nexthop legs on one BGP session, as against BGP routes on N BGP sessions is that, it limits the amount of transitionary multipath combinatorial state in the latter model. Because the final multipath state is conveyed by one route update in deterministic manner, there is no transitionary multipath combinatorial explosion created during establishment of N sessions.

7. IANA Considerations

This document makes request to IANA to allocate the following codes in BGP attributes registry.

7.1. BGP Path Attributes

A new BGP attribute code TBD for "BGP MultiNexthop Attribute (MULTI_NEXT_HOP)", in "BGP Path Attributes" registry.

7.2. Capability Codes

This document makes request to IANA to allocate a BGP capability code TBD for "BGP MultiNexthop Attribute (MULTI_NEXT_HOP), in "Capability Codes" registry.

7.3. BGP MultiNextHop Attribute

This document requests IANA to create a new registry group for MultiNextHop attribute, and the following registries in it.

7.3.1. MultiNextHop (MNH) TLV Types

This is a Registry for Type codes in Section 4.2 "MULTI_NEXT_HOP TLV"

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: MultiNexthop (MNH) TLV Types

MNH Type Code	Meaning
0	Reserved
1	Primary forwarding path
2	Backup forwarding path
3-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards Action process defined in [RFC2434], or the Early IANA Allocation process defined in [RFC4020].

7.3.2. Forwarding Action Types

This is a Registry for Type codes in Section 4.4 "Forwarding Instruction TLV"

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Forwarding Action Types

FwdAction	Meaning
0	Reserved
1	Forward
2	Pop-And-Forward
3	Swap
4	Push
5	Pop-And-Lookup
6	Replicate
7-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards Action process defined in [RFC2434], or the Early IANA Allocation process defined in [RFC4020].

7.3.3. Forwarding Argument Types

This is a Registry for Type codes in Section 4.5 "Forwarding Arguments TLV"

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Forwarding Argument Types

F.A. Type Code	Meaning
0	Reserved
1	Endpoint Identifier
2	Path Constraints
3	Payload encapsulation info signaling
4	Endpoint attributes advertisement
5-65534	Unassigned
65535	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards Action process defined in [RFC2434], or the Early IANA Allocation process defined in [RFC4020].

7.3.4. Endpoint Types

This is a Registry for Type codes in Section 5.3.1 "Endpoint Identifier" Forwarding Argument.

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Endpoint Types

Endpoint Type	Value
0	Reserved
1	IPv4 Address
2	IPv6 Address
3	MPLS Label
4	Fwd Context RD
5	Fwd Context RT
6-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards Action process defined in [RFC2434], or the Early IANA Allocation process defined in [RFC4020].

7.3.5. Path Constrain Types

This is a Registry for Type codes in Section 5.3.2 "Path Constrain" Forwarding Argument.

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Path Constrain Types

ConstrainType	Value
0	Reserved
1	Proximity check
2	Transport Class ID (Color)
3	Load balance factor
4-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards Action process defined in [RFC2434], or the Early IANA Allocation process defined in [RFC4020].

7.3.6. Encapsulation Types

This is a Registry for Type codes in Section 5.3.3 "Payload Encapsulation Info" Forwarding Argument.

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Encapsulation Types

Encap Type	Value
0	Reserved
1	MPLS Label Info
2	SR MPLS label Index Info
3	SRv6 SID info
4	DSCP code point
5-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards Action process defined in [RFC2434], or the Early IANA Allocation process defined in [RFC4020].

7.3.7. Endpoint Attribute Types

This is a Registry for Type codes in Section 5.3.4 "Endpoint attributes" Forwarding Argument.

Under "Border Gateway Protocol (BGP) Parameters",

Registry Group: BGP MultiNextHop Attribute

Registry Name: Endpoint Attribute Types

EP Attrib Type	Attrib Value
0	Reserved
1	Bandwidth
2-254	Unassigned
255	Reserved

Reference: This document.

Registration Procedure(s)
Future assignments are to be made using either the Standards Action process defined in [RFC2434], or the Early IANA Allocation process defined in [RFC4020].

Note to RFC Editor: this section may be removed on publication as an RFC.

8. Security Considerations

The MNH attribute is defined as optional non-transitive BGP attribute, such that it does not accidentally get propagated or leaked via BGP speakers that dont support this feature, especially does not unintentionally leak across EBGP boundaries.

MNH may be used to advertise nexthop with MPLS label in various BGP families. In scenarios where MPLS is enabled on link to a device in an untrusted domain, e.g. a PE-CE link or ASBR-ASBR inter-AS link, security can be provided against MPLS label spoofing by using MPLS context tables as described in MPLS enabled CE (Appendix A.9.2). Such that only MPLS traffic with labels advertised to the BGP speaker are allowed to forward. However, the PE may not be able to perform any checks based on inner payload in the MPLS packet since it performs label swap forwarding. Such 'inner payload' based checks may be offloaded to a downstream node that forwards and processes inner payload, e.g., an IP router having full FIB. These security aspects should be considered when using MPLS enabled CE devices.

Contributors

Reshma Das
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: dreshma@juniper.net

Natrajan Venkataraman
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: natv@juniper.net

Acknowledgements

Thanks to Jeff Haas, Robert Raszuk, Ron Bonica for the review, discussions and input to the draft.

Thanks to Blaine Williams and Satya Mohanty for the discussions on some usecases.

References

Normative References

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, DOI 10.17487/RFC2545, March 1999, <https://www.rfc-editor.org/info/rfc2545>.

[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>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[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>.

Informative References

[ADDPATH-GUIDELINES] Uttaro, Ed., "BGP Flow-Spec Redirect to IP Action", 25 April 2016, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-add-paths-guidelines-08#section-2>.

[BGP-CT] Vairavakkalai, Ed. and Venkataraman, Ed., "BGP Classful Transport Planes", 17 March 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ct-28>.

[FLWSPC-REDIR-IP] Simpson, Ed., "BGP Flow-Spec Redirect to IP Action", 2 February 2015, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-flowspec-redirect-ip#section-3>.

[MPLS-NAMESPACES] Vairavakkalai, Ed., "BGP Signaled MPLS Namespaces", 10 July 2023, <https://datatracker.ietf.org/doc/html/draft-kaliraj-bess-bgp-sig-private-mpls-labels-06>.

[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>.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, DOI 10.17487/RFC2474, December 1998, <https://www.rfc-editor.org/info/rfc2474>.

[SRTE-COLOR-ONLY] Filsfils, Ed., "BGP Flow-Spec Redirect to IP Action", 21 February 2018, <https://tools.ietf.org/html/draft-filsfils-spring-segment-routing-policy-06#section-8.8.1>.

Appendix A. Example of Usecases

This section describes various example usecases of the MNH attribute.

A.1. Signaling WECMP to Ingress Node

This section describes how MNH can be used to provide weighted equal cost multipath in a network fabric, while not increasing RIB scale.

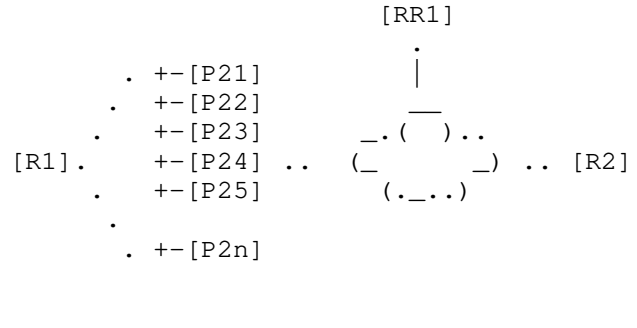


Figure 22: WECMP without increasing RIB scale

Figure 22 shows a network with BGP speaker R1 connected to a number of routers P21 .. P2n in its region. R1 is eSN and R2 is iSN for the IP traffic in consideration. BGP service families IPv4 Unicast (AFI/SAFI: 1/1) and IPv6 Unicast (AFI/SAFI: 2/1) are negotiated on the BGP sessions between RR1 - R1 and RR1 - R2. RR1 reflects the BGP routes between R1 and R2 with next hop unchanged.

When MNH is not in use, R1 advertises "n" BGP Addpath routes for a service prefix Pfx1, each having a distinct next hop, P21 .. P2n, and desired Link Bandwidth Extended Community. These Addpath routes will be received by R2, which can do WECMP based on the Link Bandwidth Extended Communities attached on the routes. This model increases RIB scale by "n" times, so that WECMP can be achieved.

When MNH is used in this network, R1 advertises a single BGP route for prefix Pfx1, which contains a MNH attribute with "n" next hops, each carrying the desired link bandwidth using Section 5.3.2.3 or Section 5.3.4.1

This allows achieving WECMP in the network without increasing RIB scale.

A.2. Signaling Optimal Forwarding Exitpoints to Ingress Node

In a BGP free core, one can dynamically signal to the ingress-node, how traffic should be load-balanced towards a set of exit nodes, in one BGP-route containing this attribute.

Example, for prefix1, perform equal load balancing towards exit nodes A, B; where as for prefix2, perform weighted load balancing (40%, 30%, 30%) towards exit nodes A, B, C.

Example, for prefix1, use PE1 as primary-nexthop and use PE2 as a backup-nexthop.

A.3. Load balancing to multiple CEs in a VRF

This section describes how MNH can be used to provide load balancing and entropy in a provider network for traffic destined to multiple CEs in a VRF, without increasing RIB scale.

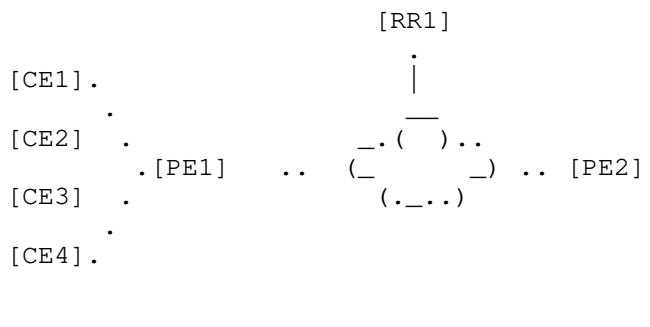


Figure 23: Load balancing to multiple CEs in a VRF

Figure 23 shows a L3VPN network with multiple CE devices connected to the same VRF at PE1. The VRF is configured with a RD: RD1, and uses "per next hop" label allocation mode to advertise the CE routes to L3VPN core. PE1 is eSN and PE2 is iSN for the IP traffic in consideration. CE1..CE4 advertise route for same prefix Pfx1 in BGP service families IPv4 Unicast (AFI/SAFI: 1/1) negotiated on the BGP sessions between the CEs and PE1. BGP L3VPN address family (AFI/SAFI: 1/128) is negotiated between PE1 - RR1, and RR1 - PE2. RR1 reflects the BGP routes between PE1 and PE2 with next hop unchanged.

PE1 would typically advertise to RR1 only the best path for prefix Pfx1 out of routes received from CE1..CE4. Using per CE RD or Addpath for L3VPN family may allow PE1 to advertise all CE routes to the RR, with an increase in RIB scale. This model increases RIB scale by "n" times, where 'n' is the number of CEs.

When MNH is used in this network, PE1 advertises a single BGP L3VPN route for prefix Pfx1, which contains a MNH attribute with "n" next hops, each carrying the label pointing towards a particular CE, using Section 5.3.3 along with the Section 5.3.1

This allows the network to direct traffic to a specific CE, and better loadbalance traffic in the provider network, with entropy provided by the per CE VPN labels, without increasing RIB scale.

A.4. Signaling Desired Forwarding Behavior for MPLS Upstream labels at Receiving Node

In Upstream label allocation case, the receiving speaker's forwarding-state can be controlled by the advertising speaker, thus enabling a standardized API to program desired MPLS forwarding-state at the receiving node. This is described in the [MPLS-NAMESPACES]

A.5. Load Balancing over EBGP Parallel Links

Consider N parallel links between two EBGP speakers. There are different models possible to do load balancing over these links:

N single-hop EBGP sessions over the N links. Interface addresses are used as next-hops. N copies of the RIB are exchanged to form N-way ECMP paths. The routes advertised on the N sessions can be attached with Link bandwidth comunity to perform weighted ECMP.

1 multi-hop EBGP session between loopback addresses, reachable via static route over the N links. Loopback addresses are used as next-hops. 1 copy of the RIB is exchanged with loopback address as nexthop. And a static route can be configured to the loopback address to perform desired N-way ECMP path. M loopbacks are configured in this model, to achieve M different load balancing schemes: ECMP, weighted ECMP, Fast-reroute enabled paths etc.

1 multi-hop EBGP session between loopback addresses, reachable via static route over the N links. Interface addresses are used as next-hops, without using additional loopbacks. 1 copy of the RIB is exchanged with MNH attribute to form N-way ECMP paths, weighted ECMP, Fast-reroute backup paths etc. BFD may be used to these directly connected BGP nexthops to detect liveness.

A.6. Flowspec Routes with Multiple "Redirect IP" next hops

There are existing protocol machinery which can benefit from the ability of MNH to clearly specify fallback behavior when multiple nexthops are involved. One example is the scenario described in [FLWSPC-REDIR-IP] where multiple Redirect-to-IP nexthop addresses exist for a Flowspec prefix. In such a scenario, the receiving speakers may redirect the traffic to different nexthops, based on variables like IGP-cost. If instead, the MNH was used to specify the redirect-to-IP nexthop, then the order of preference between the different nexthops can be clearly specified using one flowspec route carrying a MNH containing those different nexthop-addresses specifying the desired preference-order. Such that, irrespective of IGP-cost, the receiving speakers will redirect the flow towards the same traffic collector device.

A.7. Color-Only Resolution next hop

Another existing protocol machinery that manufactures nexthop addresses from overloaded extended color community is specified in [SRTE-COLOR-ONLY]. In a way, the color field is overloaded to carry one anycast BGP next-hop with pre-specified fallback options. This approach gives us only two next-hops to play with. The 'BGP nexthop address' and the 'Color-only nexthop'

Instead, the MNH could be used to achieve the same result with more flexibility. Multiple BGP nexthops can be carried, each resolving over a desired Transport class (Color), and with customizable fallback order. And the solution will work for non-SRTE networks as-well.

A.8. Avoid Label Advertisement Oscillation Between Multihomed PEs.

In a MPLS network, a router may be multihomed to two PEs. The PEs may re-advertise routes received from the router to the IBGP core with self as nexthop and a "per nexthop" label. The PEs may also protect failure of primary path to the router by using the IBGP path via the other multihomed PE as a backup path.

In this scenario, label allocation oscillation may occur when one PE advertises a new label to the other PE. Reception of a new label results in change of nexthop, as the label is used as back nexthop leg, and per-nexthop label allocation is in use. Thus a new label is allocated and advertised. And when this new label is received by the first PE, it allocates a new label in turn. This process repeats.

This oscillation can be stopped only if the primary path label allocated by a PE does not depend on the primary path label advertised by other PE. A PE needs to be able to advertise multiple labels, one for use as primary path and another to be used as bacakup path by the receiver.

MNH attribute allows to advertise a Backup forwarding path label using Section 5.1.2 in addition to Primary forwarding path label using Section 5.1.1

A.9. Signaling Intent over PE-CE Attachment Circuit

BGP CT specifies procedures for Intent Driven Service Mapping in a service provider network, and defines 'Transport Class' construct to represent an Intent.

It may be desirable to allow a CE device to indicate in the data packet it sends what treatment it desires (the Intent) when the packet is forwarded within the provider network.

This section describes the mechanisms that enable such signaling. These procedures use existing AFIs 1 or 2, and service families (SAFI 1) on the PE-CE attachment circuit, with a new BGP attribute.

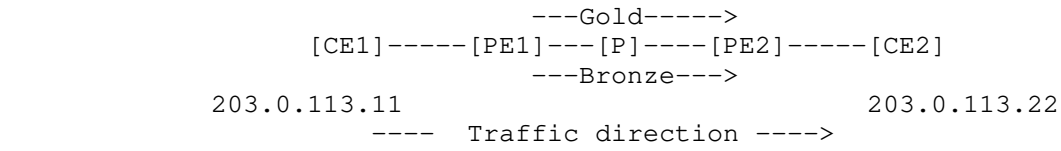


Figure 24: Example Topology with PE-CE Links

A.9.1. Using DSCP in MultiNexthop Attribute

Such an indication can be in form of DSCP code point ([RFC2474]) in the IP header.

In RFC2474, a Forwarding Class Selector maps to a PHB (Per-hop Behavior). The Transport Class construct is a PHB at transport layer.

Let PE1 be configured to map DSCP1 to Gold Transport class, and DSCP2 to Bronze Transport class. Based on the DSCP code point received on the IP traffic from CE1, PE1 forwards the IP packet over a Gold or Bronze tunnel. Thus, the forwarding is not based on just the destination IP address, but also the DSCP code point. This is known as Class Based Forwarding (CBF). Today CBF is configured at the PE1 device roles and CE1 doesn't receive any indication in BGP signaling regarding what DSCP code points are being offered by the provider network.

With a BGP MultiNexthop Attribute attached to a AFI/SAFI 1/1 service route, it is possible to extend the PE-CE BGP signaling (if used) to communicate such information to the CE1. In the preceding example, the MNH contains two Next hop Legs, described by two Forwarding Instruction TLVs. Each Next hop Leg contains PE1's peering self address in Endpoint Identifier TLV (Section 5.3.1), the color Gold or Bronze encoded in the Transport class ID TLV (Section 5.4.2.2, Figure 12), and associated DSCP code point indicating Gold or Bronze transport class encoded in the Payload Encapsulation Info TLV (Section 5.4.3.4, Section 5.3.3). This allows the CE to discover what transport classes exist in the provider network, and which DSCP codepoint to encode so that traffic is forwarded using the desired transport class in the provided network.

A.9.2. MPLS-enabled CE

If the PE-CE link is MPLS enabled, a distinct MPLS label can also be used to express Intent in data packets from CE. Enabling MPLS forwarding on PE-CE links comes with some security implications. This section gives details on these aspects.

Consider the ingress PE1 receiving a VPN prefix RD:Pfx1 received with VPN label VL1, next hop as PE2 and a mapping community containing TC1 as 'Transport class ID'. PE1 can allocate a MPLS Label PVL1 for the tuple "VPN Label, PNH Address, Transport class ID" and advertise to CE1.

Label PVL1 may identifies a service function at any node in the network, e.g. a Firewall device or egress node PE2. And, for the same service prefix, a distinct label may be advertised to different CEs, such that incoming traffic from different CEs to the same service prefix can be diverted to a distinct devices in the network for further processing. This provides Ingress Peer Engineering control to the network.

PE1 installs a MPLS FIB route for PVL1 with next hop as "Swap VL1, Push TL1 towards PE2". TL1 is the BGP CT label received for the tuple 'PE2, TC1'. In forwarding, when MPLS packet with label PVL1 is

received from CE1, PVL1 Swaps to label VL1 and pushes the BGP CT label TL1. PE1 advertises the label "PVL1" in the MNH to CE1. PE1 forwards based on MPLS label without performing any IP lookup. This allows for PE1 to be a low IP FIB device and still support CBF by using MPLS Label inferred PHB. The number of MPLS Labels consumed at PE1 for this approach will be proportional to the number of Service functions and Intents that are exposed to CE1.

A BGP MultiNexthop Attribute is attached to a AFI/SAFI 1/1 service route to convey the MPLS Label information to CE1. In the preceding example, the MNH contains two Next hop Legs, described by two Forwarding Instruction TLVs. Each Next hop Leg contains PE1's peering self address in Endpoint Identifier TLV (Section 5.3.1), the color Gold or Bronze encoded in the Transport class ID TLV (Figure 12), and associated MPLS Label "PVL1" or "PVL2" encoded in the Payload Encapsulation Info TLV (Section 5.4.3.1, Section 5.3.3). This allows the CE to discover what transport classes exist in the provider network, and which MPLS Label to encode so that traffic is forwarded using the desired transport class.

A.9.2.1. Secure MPLS Forwarding on Inter-AS Link

The MPLS enabled PE-CE attachment circuit is considered connecting to an untrusted domain. Such interfaces can be secured against MPLS label spoofing by a walled garden approach using "MPLS context tables".

The PE1-CE1 interface can be confined to a specific MPLS context table "A" corresponding to the BGP peer. Such that only the routes for labels advertised to CE1 are installed in MPLS context table "A".

This ensures that if CE1 sends MPLS packet with a label that was not advertised to the CE1, the packet will be dropped.

Furthermore, the routes for labels PVL1, PVL2 installed in MPLS context table "A" can match on 'Bottom of stack' bit being 'one', ensuring a MPLS packet is accepted from CE1 only if it has no more than one label in the label stack.

However, the PE itself may not be able to perform any checks based on inner payload in the MPLS packet since it performs label swap forwarding. Such inner payload based checks may be offloaded to a downstream node that forwards and processes inner payload, e.g. a IP FIB router. These security aspects should be considered when using MPLS enabled CE devices.

A.10. 4PE - Signal MPLS Label for IPv4 Unicast routes

This section describes how MNH can be used to signal MPLS explicit null label in AFI/SAFI: 1/1 routes in a pure IPv6 core environment, to achieve 4PE.

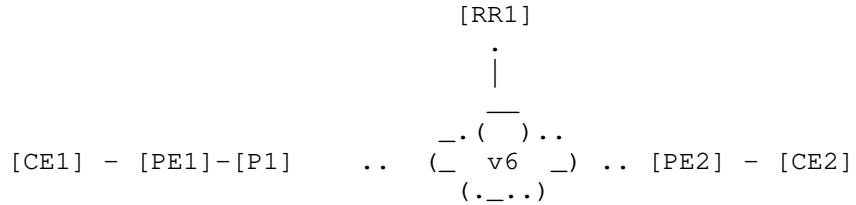


Figure 25: 4PE Network with Pure IPv6 Core

Figure 25 shows a 4PE network with pure IPv6 core, PE1 is the egress PE connected to penultimate hop node P1. PE1 to PE2 have some IPv6 core tunneling protocol like LDPv6. When PE1 has advertised Implicit Null label in LDPv6, some implementations of P1 may not be able to forward the inner IPv4 payload to PE1.

To solve this problem, PE1 needs to signal IPv4 Explicit NULL Label (Special Label 0) to PE2. PE2 will push this IPv4 Explicit NULL Label received in the MNH on the AFI/SAFI:1/1 route. Such that P1 does a MPLS Label swap operation and does not need to look into inner payload.

MNH can be used by PE1 on a AFI/SAFI: 1/1 route, to advertise the IPv4 Explicit Null label for the IPv4 Unicast service route. MPLS Label is encoded in the Payload Encapsulation Info TLV (Section 5.4.3.1, Section 5.3.3).

This allows the network to provide clear separation of service and transport routes, and not overloading AFI/SAFI: 1/4 to carry the IPv4 service routes. Not mixing service and transport routes improves security and manageability aspects of the network.

An egress PE may not need to advertise IPv4 Explicit Null label for the IPv4 service route, if it does UHP label in LDPv6. This model using MNH provides a homogenous service layer (AFI/SAFI: 1/1) that accomodates differences in requirement of different PE and P routers. Only the PEs which are connected to P nodes that cannot handle the

PHP situation need to advertise Label using MNH. The service layer is kept consistent in the network, and can seamlessly extend to multiple domains without needing redistribution between AFI/SAFIs.

Not mixing service and transport routes improves security and manageability aspects of the network.

Authors' Addresses

Kaliraj Vairavakkalai (editor)
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: kaliraj@juniper.net

Minto Jeyananth
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: minto@juniper.net

Mohan Nanduri
Microsoft
1 Microsoft Way,
Redmond, WA 98052
United States of America
Email: mohannanduri@microsoft.com

Avinash Reddy
AT&T
3400 W Plano Pkwy,
Plano, TX 75075
United States of America
Email: ar977m@att.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 9 January 2025

W. Li
H. Wang
J. Dong
Huawei Technologies
8 July 2024

Extension of Link Bandwidth Extended Community
draft-li-idr-link-bandwidth-ext-02

Abstract

[I-D.ietf-idr-link-bandwidth] defines a BGP link bandwidth extended community attribute, which can enable devices to implement unequal-cost load-balancing. However, the bandwidth value encapsulated by the extended community attribute is of the floating-point type, which is inconvenient to use. In this document, a set of new types of link bandwidth extended community are introduced to facilitate the configuration and calculation of link bandwidth.

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

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 9 January 2025.

Copyright Notice

Table of Contents

1. Introduction

In [I-D.ietf-idr-link-bandwidth], the link bandwidth extended community attribute is added to implement unequal-cost load balancing based on the bandwidth on a path. As defined in the draft, the bandwidth of a link is expressed in 4-octets in IEEE floating-point format.

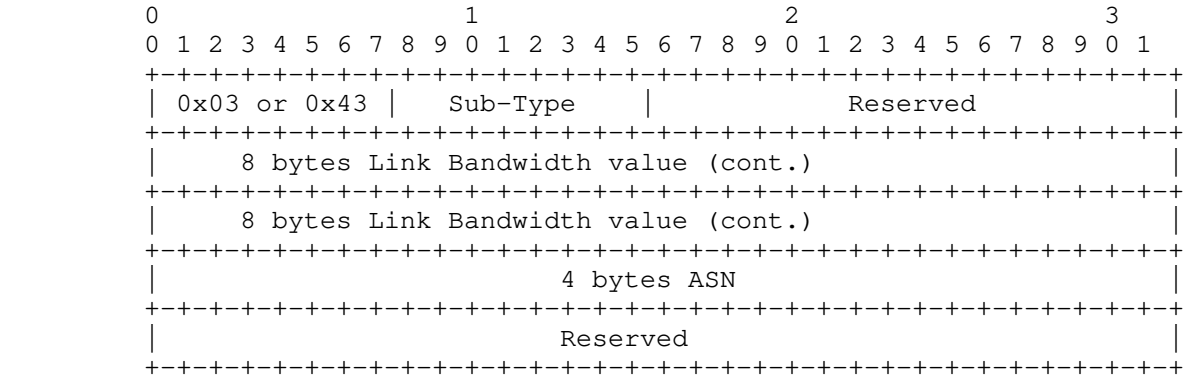
In practice, the use of this floating-point format may result errors in configuration and computation. When an operator needs to manually specify the bandwidth, you also need to consider the conversion from the bandwidth value to the floating-point number. This mode is not user-friendly, especially when the routing policy is used for bandwidth matching.

This document introduce a more intuitive expression of link bandwidth in BGP. It uses an unsigned long integer value to describe the link bandwidth value. This is easier for operators to use and understand, and can avoid configuration and computation errors.

2. Link Bandwidth Extended Community

The type of Link Bandwidth Extended Community is 0x40, and the subtype is 0x04. In the attribute value, the global administrator subfield is set to the AS number of the route to which the Link Bandwidth attribute is added. In the local administrator subfield, the link bandwidth value [I-D.ietf-idr-link-bandwidth] is set to the IEEE floating-point type.

A new type of IPv6 Address Specific Extended Community[RFC5701] is added in this document. The ASN field of this attribute is set to the AS number of the route to which the link bandwidth attribute is added. The Link Bandwidth value field (8 bytes) is set to the link bandwidth. The following extended contents are added:



* The value of the high-order octet of the extended Type, refer to [RFC4360], It is recommended that 0x03 and 0x43 be used.

* New Link Bandwidth, subtype is TBD. The value of the Link Bandwidth subfield is an unsigned long integer, in bytes per second.

The subtypes defined here can be used for both optional transitive and non-transitive extended community attributes.

3. Deployment Considerations

The extended link bandwidth extended community attribute in this document should not be used together with the standard link bandwidth extended community attribute. If a route carries both the standard link bandwidth extended community attribute and the unit link bandwidth extended community attribute, the standard link bandwidth extended community attribute is ignored.

In actual deployment, if a routing policy is used to match link bandwidth attributes, you can directly perform exact value matching.

4. IANA Considerations

This document defines a specific application of the two-octet AS specific extended community. IANA is requested to assign new sub-types for both non-transitive and transitive extended communities.

SubType	Description
TBD	Link Bandwidth EC in bytes per second

5. Security Considerations

There are no additional security risks introduced by this design.

6. Acknowledgements

7. References

7.1. 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>.

7.2. References

[I-D.ietf-idr-link-bandwidth] Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", Work in Progress, Internet-Draft, draft-ietf-idr-link-bandwidth-07, 5 March 2018, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-link-bandwidth-07>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, DOI 10.17487/RFC5701, November 2009, <https://www.rfc-editor.org/info/rfc5701>.

Authors' Addresses

Wenyan Li
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing
100095
China
Email: liwenyan@huawei.com

Haibo Wang
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing
100095
China
Email: rainsword.wang@huawei.com

Jie Dong
Huawei Technologies
Huawei Campus, No. 156 Beiqing Road
Beijing
100095
China
Email: jie.dong@huawei.com

IDR
Internet-Draft
Intended status: Standards Track
Expires: 6 January 2025

S. Sangli
S. Hegde
R. Das
Juniper Networks Inc.
B. Decraene
Orange
B. Wen
M. Kozak
Comcast
J. Dong
Huawei
L. Jalil
Verizon
K. Talaulikar
Cisco

5 July 2024

Accumulated Metric in NHC attribute
draft-ssangli-idr-bgp-generic-metric-00

Abstract

RFC7311 describes mechanism for carrying accumulated IGP cost across BGP domains however it limits to IGP-metric only. There is a need to accumulate and propagate different types of metrics as it will aid in intent-based end-to-end path across BGP domains. This document defines BGP extensions for Generic Metric sub-types that enable different types of metrics to be accumulated and carried in BGP. This is applicable when multiple domains exchange BGP routing information.

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 6 January 2025.

Table of Contents

1. Introduction

Large Networks belonging to an enterprise may consist of nodes in the order of thousands and may span across multiple IGP domains where each domain can run separate IGPs or levels/areas. BGP may be used to interconnect such IGP domains, with one or more IGP domains within an Autonomous System. The enterprise network can have multiple Autonomous Systems and BGP may be employed to provide connectivity between these domains. Furthermore, BGP can be used to provide routing over many such independent administrative domains.

The traffic types have evolved over years and operators have resorted to defining different types of metrics within a IGP domain (ISIS or OSPF) for IGP path computation. An operator may want to create an end-to-end path that satisfy certain intent. The intent could be to create end-to-end path that minimizes one of the metric-types. These metrics can be assigned administratively by an operator. While some are described in the base ISIS, OSPF specifications, other metrics are the Traffic Engineering Default Metric defined in [RFC5305] and [RFC3630], Min Unidirectional delay metric defined in [RFC8570] and [RFC7471]. There may be other parameters such as jitter, reliability, fiscal cost, etc. that an operator may incorporate while computing the cost of a link. The procedures mentioned in the above specifications describe the IGP path computation within IGP domains.

For computing the best path for a BGP route in such a domain, the step(e) of the section 9.1.2.2 of [RFC4271] specifies that the interior cost of a route as determined via the IGP metric value is to be used to break the tie among multiple paths. When multiple domains are interconnected via BGP, protocol extensions for advertising best-external path and/or ADDPATH as described in [RFC7911] are employed to take advantage of network connectivity thus providing alternate paths. For each route that is advertised, the IGP cost of the end-to-end path is accumulated and encoded in the AIGP attribute as described in [RFC7311]. This can be used to compute the AIGP-enhanced interior cost and it will be used in the decision process for selecting the best path as documented in section 2 of [RFC7311] . The [RFC7311] specifies how AIGP attribute can carry the accumulated IGP metric value. However, [RFC7311] describes only one TLV (AIGP TLV) in the AIGP attribute to carry the IGP cost. Most of the implementations available today encode IGP-metric metric type in the AIGP TLV. See section Section 15 for different interpretations of [RFC7311] that are deployed today.

With the advent of 5G applications and Network Slicing applications, for catering to the various traffic constraints, an operator may wish to provision end-to-end paths across multiple domains satisfying required intents. This is also known as intent-based inter-domain routing. The description of the problem space and requirements can be found in [I-D.draft-hr-spring-intentaware-routing-using-color]. The Clasful Transport Planes as described in [I-D.draft-ietf-idr-bgp-classful-transport-planes] and and Color-Based Routing as described in [I-D.draft-ietf-idr-bgp-car] describe how intent-based end-to-end paths can be established. The proposal described in this document can be used in conjunction with such architectures.

If the type of metric used in a IGP domain differs from the accumulated metric type carried in BGP, the metric values should be synchronized and translated between IGP domain and BGP. The metric-type and metric-value in the AIGP TLV does not support different IGP metric-types defined in the IGP-Protocol registry for metric-types. Hence there is a need to provide a generic metric TLV template to embed the different types of IGP metrics and their values in BGP.

This document proposes "Accumulated Metric" TLV in the Next-Hop Dependent Capability (NHC) attribute described in [I-D.ietf-idr-entropy-label] to carry the accumulated metric value for end-to-end path, hereby referred as AMetric. The AMetric supports all the metric types defined in the IGP-Parameters metric-type registry. Additionally this document provides procedures for computation and usage of accumulated generic metric value during the BGP best path computation.

[RFC7311] introduces the notion that a set of ASes can be under a common administrative domain. This document borrows the same concept, "AMetric administrative domain" to refer to ASes under a common administration within which an operator wishes to establish any intent-based end-to-end path.

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. Multiple types of metrics in a network

Consider the network as shown in Figure 1. The network has multiple domains. Each domain runs a separate IGP instance. Within each domain iBGP sessions are established between the PE routers. The eBGP sessions are established between the Border Routers across domains.

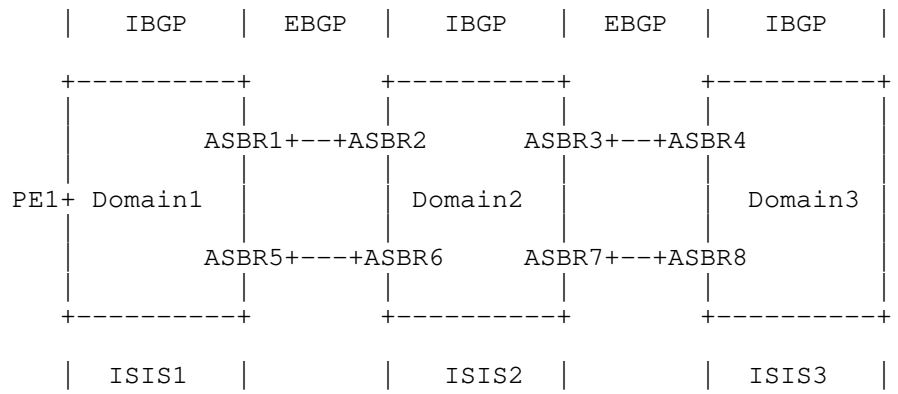


Figure 1: WAN Network

An operator wishes to compute end-to-end path optimized for "delay" metric-type. Each domain will be enabled for computation of the IGP paths based on delay metric type. As a result, the intra-domain reachability will be based on delay metric. Such values should also be propagated to the adjacent domains for effective end-to-end path computation. However, the AIGP TLV in the AIGP attribute as specified in [RFC7311] supports only the default IGP-metric. As a result, with AIGP TLV, only default IGP-metric based end-to-end path can be computed and this will not address the operator requirements.

The [I-D.ietf-lsr-flex-algo-bw-con] proposes extension in ISIS and OSPF, a generic metric type that can embed multiple metric types within it. It supports both standard metric-types and user-defined metric-types. This document describes extensions in BGP to support such metric types. To compute the end-to-end path with metric other than default IGP-metric cost, the new metric TLV for NHC that this document proposes will enable different types of metrics and their values to be accumulated and propagated across the domains.

4. Discontinuity in end-to-end intent

For determining the end-to-end path for an intent and to derive the accurate cumulative cost, all routers along the path that modify the next hop should participate in cumulating the cost. New applications may require new intents that may result in new metric types to be used in the network. It is quite possible that certain domains may have specific metric types. For example, core network may have latency metric while metro may just have IGP-cost because delay in metro regions may be insignificant. It is quite possible that one or more routers might not understand the new intent and the metric.

If one or more routers in a domain either border routers or any intra-domain routers that modify the next hop, do not participate in accumulating the cost when they propagate a route, it is impossible for the ingress router to determine the cost of the end-to-end path accurately. This will result in sub-optimal best path selection. Such an end-to-end path is called discontinuous path for that intent. The discontinuity can manifest in different dimensions and Section 12 provides detailed explanation.

5. Generic Metric Encoding

This document proposes "Accumulated Metric" TLV (AMetric), in the Next Hop Dependent Capability (NHC) attribute. The format is shown below.

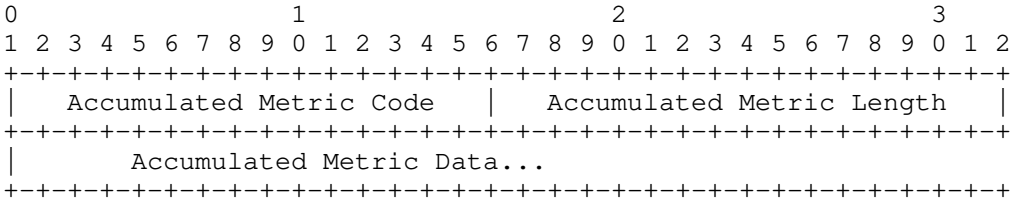


Figure 2: AMetric

Accumulated Metric Code (2 octet): Code point to be assigned by IANA

Accumulated Metric Length (2 octets): Value 10

Accumulated Metric Data (10 octets): 3 sub-fields as shown below:

The AMetric data carries the metric type and metric value as defined in the IGP-Protocol registry for metric-types. The format is shown below.

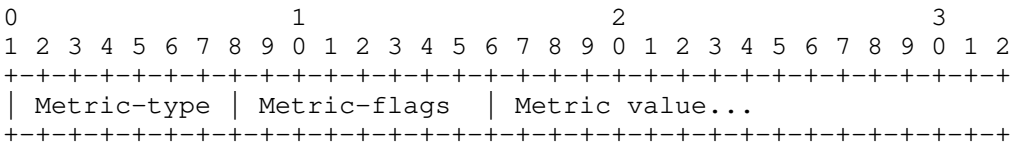


Figure 3: AMetric Data

1. Metric-type (1 octet): Value of metric-type from IGP-Protocol registry for metric-types.

2. Metric-flags (1 octet): Bits defined below.

3. metric-value (8 octets): Value range (0 - 0xffffffffffffffff)

The metric-flags carry additional information about the accumulated generic metric.

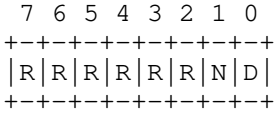


Figure 4: Accumulated Metric flags

Bit D : Represents discontinuity in metric accumulation for the end-to-end path. 1 indicates discontinuous, 0 indicates continuous.

Bit N : Represents normalization of metric in the local domain. 1 indicates metric normalization has been applied. 0 indicates no normalization has been applied.

Bit R : Reserved for future use. MUST be set to zero when originated, and MUST be ignored on receipt.

6. Leverage Next Hop Dependent Capability (NHC) Attribute

The Next Hop Dependent Capability attribute has two important characteristics that are relevant for establishing the end-to-end intent path: Transitivity and Scoping.

Transitivity: The NHC is an optional and transitive attribute hence according to [RFC4271] if this attribute is present in an update message, it will be propagated to all neighbors. Via the AMetric in NHC attribute, the accumulated metric value is propagated to all the routers in the network, thus making the cost computation for the end-to-end intent path possible.

Scoping: The NHC provides an ability to perform Next Hop scoping. The originating router encodes the next hop address in the "Network Address of Next Hop" field in the NHC attribute before advertising the route. If the non-originator router upon learning such a route, does not modify the next hop, it will advertise the route along with capabilities in NHC without modification. On the other hand, if the non-originator router upon learning such a route, modifies the next hop, it will perform two actions: Encode the next hop address (it is going to advertise in the update), in "Network Address of Next Hop" field of the NHC attribute; and recompute the capabilities and encode them in the NHC attribute, before advertising the route. With this support, the receiving BGP speaker can make specific decisions for the route by comparing the advertised next hop present in the update message with the "Network Address of Next Hop" field of the NHC attribute. If the next hop in the update message does not match with the "Network Address of Next Hop" field in the NHC attribute, it can be concluded that the advertising router did not understand the NHC attribute and as a result, the capabilities have not been updated at the advertising router.

The above mechanism is leveraged to determine if the end-to-end path for an intent (represented by a metric in AMetric) is discontinuous or not. The document [I-D.ietf-idr-entropy-label] provides detailed explanation on the NHC procedures.

7. Comparison between AIGP and AMetric

The AIGP TLV described in [RFC7311] carries IGP metric type only. The AIGP TLV is encoded in AIGP attribute, an optional and non-transitive attribute. The BGP speaker S may attach AIGP attribute with AIGP TLV in it, to a route R and advertise to its peers. When such a route propagates across domains, the metric value in AIGP TLV is accumulated thus providing end-to-end IGP cost.

The AMetric is like AIGP TLV. The AMetric can carry not just the IGP metric type but also other types of metrics. The AMetric will be encoded in the NHC attribute. The BGP speaker S may attach NHC attribute with AMetric in it to a route R and advertise R to its peers. When such a route propagates across domain, the metric value will be accumulated. This provides the end-to-end cost for desired intent as represented by one or more metric types.

The section 3.4 of [RFC7311] describe procedures for creating and modifying the AIGP TLV in the AIGP attribute and these are also applicable for creating and modifying the AMetric in the NHC attribute.

8. Usage of Accumulated Metric (AMetric)

8.1. Generation of Accumulated Metric

It is recommended that an implementation that supports AMetric MUST support a configuration item AMetric_ORIGINATE, that enables or disables its creation and inclusion into NHC. The default value of AMetric_ORIGINATE MUST be "disabled". If the BGP speaker S is originating route R into BGP, it MAY include NHC attribute to it. When a BGP speaker wishes to generate the AMetric and add it to the NHC attribute, it MUST perform the following procedures:

- The procedures for the BGP speaker S to send NHC attribute for a route R with next hop N as described in section 2.2 of [I-D.ietf-idr-entropy-label] MUST be followed while encoding AMetric.

- A BGP speaker S MUST NOT add the AMetric to NHC attribute for a route R whose path leads outside the AMetric administrative domain. When S as an ASBR, advertises the route to peers inside the AS by setting itself as next hop, it MUST add AMetric to NHC. S MUST NOT add AMetric to NHC for route advertisements to neighboring AS that are not part of AMetric administrative domain.

- The BGP speaker S MUST not encode the AMetric in the NHC attribute for a route R for which S does not set itself as the next hop.

- In section 3.4 of [RFC7311] whenever AIGP TLV is referred to, it MUST be treated as AMetric. Whenever AIGP attribute is referred to, it MUST be treated as NHC attribute. The procedures outlined in section 3.4 of [RFC7311] MUST be followed for creation of AMetric that is encoded in NHC attribute. Similarly, the procedures outlined in 3.4 of [RFC7311] MUST be followed for the modifications of AMetric in NHC attribute by the originator and non-originator of the route.

- Repeated metric changes may cause large number of BGP updates to be generated and propagated throughout the network. To avoid this, a configurable threshold for the metric is defined. If the difference between the new metric-value and the advertised metric-value is less than the configured threshold, the update MAY be suppressed. For each of type of metric used in the domain, if the new metric-value encoded in AMetric is above the configured threshold, a new BGP update containing the new set of metric-values SHOULD be advertised.

- Procedures for defining the cost to reach a next hop for various metric-types is outside the scope of this document.

8.1.1. Originator of route into BGP

In addition to the above, the BGP speaker S MUST perform the following procedures when it wishes to add AMetric to NHC.

- The BGP Speaker S MUST encode the type of metric as specified in the IGP Protocol registry in the metric-type sub-field. This metric type should represent the intent required for establishing the end-to-end path.

- The BGP Speaker S MUST encode the value of the metric or cost to reach the next hop N in the metric-value sub-field. The cost may be normalized if required.

- If a domain adopts more than one metric type to represent an intent, the BGP speaker S may encode more than one AMetric, each encoding different type of metric as defined in the IGP Protocol Registry.

- The "D" bit of the metric-flags MUST be set to zero. If the domain internal cost to reach the next hop is normalized to the type of metric in the metric-type sub-field, the "N" bit of the metric-flags MUST be set to 1, else it MUST be set to zero.

8.1.2. Non-Originator of route into BGP

The BGP speaker S has received the route R that has NHC and when it advertises R to its peers, it recreates NHC. Here, S MUST perform the following procedures.

- The BGP speaker S MUST retain all the received AMetric and for each of the AMetric received, it MUST perform following procedures.

- If the type of the metric used in local domain is same as the metric-type of the AMetric, the BGP speaker S MUST add the metric or cost to reach the next hop N to the metric-value sub-field of the AMetric.

- If the type of the metric used in the local domain is different from the metric-type of the AMetric, the BGP speaker S MUST normalize the metric or cost to reach the next hop N before adding to the metric-value sub-field of the AMetric. The metric-value sub-field MUST be increased by a non-zero amount.

- If the local domain's internal cost to reach the next hop is normalized to the type of metric in the metric-type sub-field, then "N" bit MUST be set to 1. If the BGP speaker S does not understand the type of metric, then "D" bit MUST be set to 1.

8.2. Reception of Accumulated Metric

When a BGP speaker S receives a BGP update that has a route R to destination prefix P with next hop N, and has the NHC attribute with AMetric, it MUST perform the following procedures:

- A BGP speaker S MUST perform the procedures described in section 2.3 of [I-D.ietf-idr-entropy-label] for processing the NHC attribute.

- For a given metric-type, if there are more than one AMetric in the NHC attribute, first occurrence MUST be processed, and the other occurrences MUST be ignored.

- There may be more than one AMetric, each carrying different types of metrics. For each AMetric, if the BGP speaker S recognizes the type of the metric encoded in the metric-type sub-field, it MUST process the metric-value and metric-flags sub-fields as per following.

- If the type of the metric used in the local domain (for resolving the next hop N) matches with the metric-type of AMetric of NHC attribute, then the metric-value sub-field MUST be used in the AIGP-enhanced interior cost computation as specified in Section 9.

- If the type of the metric used in the local domain (for resolving the next hop N) does not match with the metric-type of AMetric of the NHC attribute, then the BGP speaker S may normalize the cost of the path used for resolving the next hop before using it in the AIGP-enhanced cost computation. A policy may be used to provide the metric normalization.

9. Updates to Decision Procedure

This section follows the approach as laid out in [RFC7311] to select the best path when the route has NHC attribute with AMetric. The domain that the BGP speaker S belongs to, may support different intent-based paths represented via different types of metric to reach next hop N. The following procedures MUST be followed in addition to the general procedure described in section 4 of [RFC7311].

Consider a BGP speaker S receiving a route R with next hop N. The route R is attached with NHC attribute having AMetric. For each AMetric, the BGP speaker S MUST perform following procedures.

If the metric-type sub-field of AMetric matches with the type of the metric used in the local domain for resolving the next hop N, the AIGP-enhanced interior cost should be computed as below.

> Let 'm' be the cost to reach the next hop N that is used in the local domain for the path computation as described in [RFC7311] .

If the metric-type sub-field of the AMetric does not match with the type of the metric used in the local domain for resolving the next hop N, the cost of the path to reach next hop N may be normalized. The normalized metric value can be zero, maximum metric value or scaled up (multiple of a positive number).

> Let 'm' be the normalized value of the cost to reach the next hop N that is used in the local domain for path computation as described in [RFC7311].

The AIGP-enhanced interior cost computation as described below will be used in the decision process as described in [RFC7311].

> Let 'A' be the value of the value of the metric-value sub-field of the AMetric.
>
> The AIGP-enhanced interior cost will be 'A+m'.

A path with IGP metric-type in AMetric of NHC attribute and a path with IGP metric-type in AIGP TLV of AIGP attribute can be compared. However, a path with AMetric carrying different metric-type and a path with AIGP TLV carrying IGP metric-type cannot be compared. To enable end-to-end path selection based on intent, the path with AMetric in NHC attribute may be chosen over path with AIGP TLV in AIGP attribute. The implementation should allow a local policy to specify these preferences.

A path with AMetric of metric-type 'a' cannot be compared with a path with AMetric of metric-type 'b'. The path with lower metric-type MAY be chosen as best between two such paths and implemented consistently across AIGP domain.

10. Use-case: Different Metrics across Domains

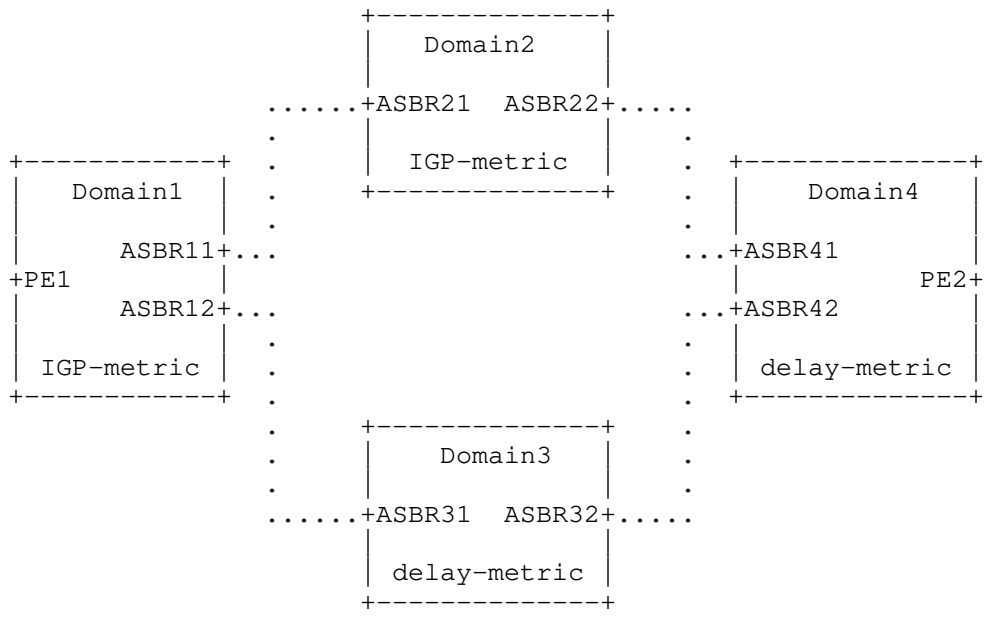


Figure 5: Different metric across network

Each domain is a separate Autonomous System. Within each domain, ASBR and PE form iBGP peering and they may employ Route Reflectors. The IGP within each domain uses domain specific metric. Domain3 and Domain4 use delay as the metric while Domain1 and Domain2 use default IGP-metric cost. ASBRs across domains form eBGP peering.

10.1. Scenario 1: Find delay-based end-to-end path.

This is about finding the delay-based end-to-end path from Domain1 to Domain4. This can be achieved as follows. The advertising router adds the AMetric with metric type 1 that represents delay metric, encoded in NHC attribute. In the above network diagram, ASBR41 (and ASBR42) will advertise prefix PE2-loopback with AMetric with delay as metric-type. The metric-value sub-field of the AMetric will represent the delay cost to reach PE2's loopback end-point from the advertising router as they will do next hop self.

In Domain3, when ASRB32 advertises the prefix PE2-loopback within the local domain, it may add cost to the metric-value, the value representing the delay introduced by the DMZ link between ASRB32 to ASBR42. When ASRBR31 advertises the prefix PE2-loopback, it will perform the following procedures.

1. Compute the delay d of the path to reach ASBR32 from which it has chosen the best path.

2. Add the above d value to the metric-value sub-field of the AMetric.

In Domain2 however, the local metric type is default IGP-metric. The ASBR22 may follow the procedure similar to ASBR32 and add the delay value corresponding to the DMZ link between ASBR22 and ASBR41 before advertising the path internally in Domain2. When ASBR21 computes the AIGP-enhanced interior cost, as mentioned before, it may normalize the internal cost to reach ASBR22 and may add the normalized value to the metric-value of AMetric representing delay metric-type. The ASBR21 will also update metric-flags sub-field to indicate that metric value has been normalized. In the above network example, the delay cost from ASBR21 to ASBR22 is negligible and hence delay-metric value will be increased nominally with a non-zero value.

The procedures for AIGP-enhanced interior cost computation at ASBR11 (and ASBR12) will follow DMZ delay computation procedure described above. PE1 will have two paths to reach PE2-loopback: P1 via ASBR11 (and domain2) and P2 via ASBR12 (and domain3), each having respective AIGP-enhanced interior cost representing end-to-end delay. The local metric type is default IGP-metric and hence PE1 may normalize the internal cost for the AIGP-enhanced interior cost computation. The BGP decision process described in Section 9 will result in delay optimized end-to-end path for PE2-loopback on PE1 that can be used to resolve the service prefixes.

10.2. Scenario 2: Find IGP-metric based end-to-end path leveraging domain-specific path.

This is about providing best-effort or default IGP-metric based end-to-end path while leveraging the domain-specific delay-based metric for intra-domain path selection. All the ASBR routers will update the AMetric for NHC attribute for the default IGP-metric metric-type, accumulating the cost for end-to-end path. The PE1 router will have two paths (from ASBR11 and ASBR12) decorated with different best-effort default IGP-metric cost. The intra-domain path to reach the domain exit can be based on domain-specific metric-type. For example, in Domain3, ASBR31 can select lowest delay path to reach ASBR32. The ASBR and the PE routers may be configured to prefer one metric-type for end-to-end path while another metric-type for intra-domain and such configuration mechanism is outside the scope of this document.

10.3. Scenario 3: Path selection when a router does not understand the new metric-type.

This is about selecting a path which a router does not support the new type of metric. The Domain2 implements only default IGP-metric and does not support delay-metric. When ASBR21 receives the route with NHC attribute and AMetric, the metric-type delay-metric is unrecognized. The ASBR21 will update the metric-flags, setting the "D" bit to 1 indicating that path is discontinuous and accumulation is incomplete. When such a route reaches PE1, the PE1 router will have two paths, one via ASBR11 with "D" bit set and another path from ASBR12 with "D" bit set to zero. The local policy on PE1 can provide guidance on the preference between these two paths.

11. Deployment Considerations

It can be noted that for a path, a domain may normalize the metric-value used to resolve next hop to match with the metric-type present in the AMetric. The idea is to propagate the cost of reaching the prefix through the domain while maintaining the metric-type chosen by the originating router and domain, thereby providing an end-to-end path for the desired intent. Such normalization of metric values to the match with the metric type present in the AMetric(s) can be done via policy. The definition of such policies and how they can be enforced is outside the scope of this document. In topologies where there is a common router between adjacent domains that do iBGP peering, the Border router can provide the normalization.

In a domain, the cost of a path is derived from the metric of its links, summed up typically. It is important to maintain the same behavior for inter-domain path. The AIGP-enhanced interior cost should not be allowed to decrease through the metric normalization. When adjacent domains use different metric types, the ASBR that connects two domains is better suited to pass on the metric values by setting itself as next hop.

All routers of a domain MUST compute the AIGP-enhanced interior cost as described above to be used during decision process. Within a domain, if one router R1 applies AIGP-enhanced interior cost while R2 does not, it may lead to routing loop unless some sort of tunneling technology viz. MPLS, SRv6, IP, etc. is adopted to reach the next hop. In a network where any tunneling technology is used, one can incrementally deploy the generic metric functionality. In a network without any tunneling technology, it is recommended that all routers MUST support Accumulated Generic-Metric based AIGP-enhanced interior cost computation. Additionally, to have consistent BGP best path in the network, all routers should use the accumulated cost during the

best path computation. To ease the deployment of this generic metric based end-to-end path selection, it is recommended to enable AMetric via configuration and should be disabled by default.

In certain networks, routes may be redistributed between BGP and IGP, usually controlled via a policy. When a route is propagated across domains, a router should use the metric-value in the AMetric of the NHC attribute, optionally modified via the local policy as the IGP cost during route redistribution into IGP. The local policy should apply metric normalization or translation based on metric-type of AMetric and the metric-type adopted in the local domain.

12. Manifestations of Discontinuity in end-to-end path

The network operators would like to avoid the scenarios when the entire network has to be upgraded before enabling the new functionality. New functionality across a network is typically deployed incrementally and one cannot expect that all routers shall be capable of handling new functionality on day-one. However, for determining the end-to-end path for an intent and to derive the accurate cumulative cost, all routers along the path that modify the next hop should participate in accumulating the cost. The discontinuity can manifest in three different forms.

* Type-A discontinuity: The advertising router does not support new attribute. However, the route is propagated to the ingress router and one cannot determine if the accumulation done end-to-end or not.

* Type-B discontinuity: The advertising router supports the new attribute but does not support TLV that accumulates the end-to-end cost. Similar to the above, the route is propagated to the ingress router and one cannot distinguish if the cumulated value represents the end-to-end cost.

* Type-C discontinuity: The advertising router supports the new attribute and the TLV that carries the accumulated cost, however as new metric-types and intent are introduced, the advertising router might not support them. Similar to the above, the route is propagated to the ingress router and one cannot distinguish if the cumulated value represents the end-to-end cost for the intended intent.

12.1. Handling discontinuous path with AIGP attribute as per [RFC7311]

The AIGP TLV is used for accumulating the metric across domains. The AIGP attribute is optional and non-transitive. The accumulated metric is used in best path computation. The AIGP mechanism requires all the routers MUST understand the new attribute so that accumulated metric will reach all the routers for consistent best path computation. Hence, if an advertising router along the path does not understand the AIGP attribute, the accumulation will not be complete making the path discontinuous. The receiving (or ingress) router will not be able to compute the end-to-end cost and so the path will not be considered for providing end-to-end intent.

On the other hand, when the ingress router receives a route with AIGP attribute, the value accumulated in the TLV is guaranteed to reflect the end-to-end cost. Therefore First-Order discontinuity does not exist. The [RFC7311] introduced both AIGP attribute and AIGP TLV together, most of BGP routers support AIGP TLV along with the attribute. Hence Type-B discontinuity is less likely to happen. The [RFC7311] specifies only default IGP-metric in AIGP TLV, hence Type-C discontinuity is not applicable. If [RFC7311] was extended to support generic metric types, it will suffer from Type-C discontinuity.

12.2. Handling discontinuous path with AMetric of NHC attribute

The AMetric is part of NHC which is an optional and transitive attribute. The routes with NHC attribute and accumulated metric reaches all the routers across the domains and it will result in consistent best path computation. If any advertising router along the path does not understand the NHC attribute, it fails to update the next hop field in NHC attribute. This is the Type-A discontinuity. However, with the NHC procedures, the receiving router will detect this through next hop validation thus providing mechanism to detect the Type-A discontinuity deterministically.

If the advertising router along the path supports NHC attribute, but does not support AMetric, following NHC procedures, the router will not propagate the accumulated metric. This is the Type-B discontinuity but this will not result non-deterministic behaviour the receiving BGP router will not select the path for desired intent given the lack of AMetric. If the advertising router understands NHC attribute and AMetric, but does not understand the specific metric-type, by following the NHC procedures, the router will still propagate NHC attribute and AIGPv2 even though unrecognized metric is not accumulated. This is Type-C discontinuity. The procedures in this document specifies the "D" bit of the unrecognized AIGPv2 to be set to 1 by the advertising router and hence the receiving router deterministically identifies the discontinuity.

12.3. Contiguity Compliance

Even though NHC attribute is transitive, the AMetric might not be interpreted and/or updated by routers along the path. If all BGP routers that modify the next hop accumulate the cost and propagate the metric, the receiving BGP router will be assured of a correct end-to-end path for the intent and the metric. Although the three types of discontinuity can be addressed using a local policy, it is recommended that operators identify such routers and upgrade them to achieve intent-based end-to-end path for optimal results.

13. Security Considerations

This document does not introduce any new security considerations beyond those already specified in [RFC4271], [RFC7311] and [I-D.ietf-idr-entropy-label].

14. IANA Considerations

IANA is requested to assign a code point for AMetric. The metric-type field refers to the IGP-Protocols registry for metric-type defined in [I-D.ietf-lsr-flex-algo-bw-con]

15. Limitations of [RFC7311]

This section provides an overview of limitations and different interpretations of [RFC7311]. Various vendors have interpreted [RFC7311] differently and encode AIGP TLV or treat AIGP attribute differently. The following lists some of them.

- The [RFC7311] introduces only one TLV: AIGP TLV. Some vendors propagate the only AIGP TLV and drop other unrecognized TLV if any.

- The [RFC7311] specifies only one type of metric: IGP-metric. However, some vendors provide option to encode different types of metrics in AIGP TLV other than default IGP-metric type.

- Some vendors do not propagate AIGP attribute if AIGP TLV is not present in it.

16. Acknowledgements

The authors would like to thank John Scudder, Jeff Haas, Robert Raszuk, Kaliraj Vairavakkalai, and Peng Shaofu for careful review and suggestions.

17. References

17.1. Normative References

[I-D.ietf-idr-entropy-label]
Decraene, B., Scudder, J., Henderickx, W., Kompella, K., Mohanty, M., Uttaro, J., and B. Wen, "BGP Next Hop Dependent Capabilities Attribute", Work in Progress, Internet-Draft, draft-ietf-idr-entropy-label-14, 1 March 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-entropy-label-14>.

[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>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

[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>.

17.2. Informative References

[I-D.hr-spring-intentaware-routing-using-color]
Hegde, S., Rao, D., Uttaro, J., Bogdanov, A., and L. Jalil, "Problem statement for Inter-domain Intent-aware Routing using Color", Work in Progress, Internet-Draft, draft-hr-spring-intentaware-routing-using-color-03, 23 October 2023, <https://datatracker.ietf.org/doc/html/draft-hr-spring-intentaware-routing-using-color-03>.

[I-D.ietf-idr-bgp-car]
Rao, D., Agrawal, S., and Co-authors, "BGP Color-Aware Routing (CAR)", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-car-10, 26 April 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-car-10>.

[I-D.ietf-idr-bgp-ct]
Vairavakkalai, K. and N. Venkataraman, "BGP Classful Transport Planes", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-ct-33, 26 April 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ct-33>.

[I-D.ietf-lsr-flex-algo-bw-con]
Hegde, S., Britto, W., Shetty, R., Decraene, B., Psenak, P., and T. Li, "Flexible Algorithms: Bandwidth, Delay, Metrics and Constraints", Work in Progress, Internet-Draft, draft-ietf-lsr-flex-algo-bw-con-12, 19 May 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-lsr-flex-algo-bw-con-12>.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <https://www.rfc-editor.org/info/rfc3630>.

[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>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC7471] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", RFC 7471, DOI 10.17487/RFC7471, March 2015, <https://www.rfc-editor.org/info/rfc7471>.

[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>.

[RFC8570] Ginsberg, L., Ed., Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 8570, DOI 10.17487/RFC8570, March 2019, <https://www.rfc-editor.org/info/rfc8570>.

Authors' Addresses

Srihari Sangli
Juniper Networks Inc.
Exora Business Park
Bangalore, KA 560103
India
Email: ssangli@juniper.net

Shraddha Hegde
Juniper Networks Inc.
Exora Business Park
Bangalore, KA 560103
India
Email: shraddha@juniper.net

Reshma Das
Juniper Networks Inc.
1133 Innovation Way
Sunnyvale, CA 94089
USA
Email: dreshma@juniper.net

Bruno Decraene
Orange
France
Email: bruno.decraene@orange.com

Bin Wen
Comcast
USA
Email: bin_wen@comcast.com

Marcin Kozak
Comcast
USA
Email: marcin_kozak@comcast.com

Jie Dong
Huawei
China
Email: jie_dong@huawei.com

Luay Jalil
Verizon
USA
Email: luay.jalil@verizon.com

Ketan Talaulikar
Cisco
India
Email: ketant.ietf@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 22 January 2025

X. Xu
China Mobile
S. Hegde
Juniper
Z. He
Broadcom
J. Wang
Centec
H. Huang
Huawei
Q. Zhang
H3C
H. Wu
Ruijie Networks
Y. Liu
Y. Xia
Tencent
P. Wang
Baidu
H. Li
IEIT SYSTEMS

21 July 2024

Fully Adaptive Routing Ethernet using BGP
draft-xu-idr-fare-01

Abstract

Large language models (LLMs) like ChatGPT have become increasingly popular in recent years due to their impressive performance in various natural language processing tasks. These models are built by training deep neural networks on massive amounts of text data, often consisting of billions or even trillions of parameters. However, the training process for these models can be extremely resource-intensive, requiring the deployment of thousands or even tens of thousands of GPUs in a single AI training cluster. Therefore, three-stage or even five-stage CLOS networks are commonly adopted for AI networks. The non-blocking nature of the network become increasingly critical for large-scale AI models. Therefore, adaptive routing is necessary to dynamically load balance traffic to the same destination over multiple ECMP paths, based on network capacity and even congestion information along those paths.

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

Large language models (LLMs) like ChatGPT have become increasingly popular in recent years due to their impressive performance in various natural language processing tasks. These models are built by training deep neural networks on massive amounts of text data, often consisting of billions or even trillions of parameters. However, the training process for these models can be extremely resource-intensive, requiring the deployment of thousands or even tens of thousands of GPUs in a single AI training cluster. Therefore, three-stage or even five-stage CLOS networks are commonly adopted for AI networks. Furthermore, In rail-optimized CLOS topologies with standard GPU servers (HB domain of eight GPUs), the Nth GPUs of each server in a group of servers are connected to the Nth leaf switch, which provides higher bandwidth and non-blocking connectivity between the GPUs in the same rail. In rail-optimized topology, most traffic between GPU servers would traverse the intra-rail networks rather than the inter-rail networks.

The non-blocking nature of the network, especially the network for intra-rail communication, become increasingly critical for large-scale AI models. AI workloads tend to be extremely bandwidth-hungry and they usually generate a few elephant flows simultaneAously. If the traditional hash-based ECMP load-balancing was used without any optimization, it's highly possible to cause serious congestion and high latency in the network once multiple elephant flows are routed to the same link. Since the job completion time depends on worst-case performance, serious congestion will result in model training time longer than expected. Therefore, adaptive routing is necessary to dynamically load balance traffic to the same destination over multiple ECMP paths, based on network capacity and even congestion information along those paths. In other words, adaptive routing is a capacity-aware and even congestion-aware path selection algorithm.

Furthermore, to reduce the congestion risk to the maximum extent, the routing should be more granular if possible. Flow-granular adaptive routing still has a certain statistical possibility of congestion. Therefore, packet-granular adaptive routing is more desirable although packet spray would cause out-of-order delivery issue. A flexible reordering mechanism must be put in placeï¼\210e.g., egress ToRs or the receiving servers). Recent optimizations for RoCE and newly invented transport protocols as alternatives to RoCE no longer require handling out-of-order delivery at the network layer. Instead, the message processing layer is used to address it.

To enable adaptive routing, no matter whether flow-granular or packet-granular adaptive routing, it is necessary to propagate network topology information, including link capacity and/or even available link capacity (i.e., link capacity minus link load) across the CLOS network. Therefore, it seems straightforward to use link-state protocols such as OSPF or ISIS as the underlay routing protocol in the CLOS network, instead of BGP, for propagating link capacity information and/or even available link capacity information. How to leverage OSPF or ISIS to achieve adaptive routing has been described in [I-D.xu-lsr-fare]. However, some data center network operators have been used to the use of BGP as the underlay routing protocol of data center networks [RFC7938]. Therefore, there is a need to leverage BGP to achieve adaptive routing as well.

[I-D.ietf-idr-link-bandwidth] has specified a way to perform weighted ECMP based on link bandwidths conveyed in the non-transitive link bandwith extended community. However, it is impractical to enable adaptive routing by directly using the non-transitive link bandwidth extended community due to the following constraints as mentioned in [I-D.ietf-idr-link-bandwidth].

"No more than one link bandwidth extended community SHALL be attached to a route. Additionally, if a route is received with link bandwidth extended community and the BGP speaker sets itself as next-hop while announcing that route to other peers, the link bandwidth extended community should be removed. The extended community is optional non-transitive."

Hence, this document defines a new extended community referred to as Path Bandwidth Extended Community and describes how to use this newly defined path bandwidth extended community to achieve adaptive routing.

Note that while adaptive routing especially at the packet-granular level can help reduce congestion between switches in the network, thereby achieving a non-blocking fabric, it does not address the incast congestion issue which is commonly experienced in last-hop switches that are connected to the receivers in many-to-one communication patterns. Therefore, a congestion control mechanism is always necessary between the sending and receiving servers to mitigate such congestion.

2. Terminology

This memo makes use of the terms defined in [RFC4360].

3. Path Bandwidth Extended Community

The Path Bandwidth Extended Community is used to indicate the minimum bandwith of the path towards the destination. It is an new IPv4 Address Specific Extended Community that can be transitive or non-transitive.

The value of the high-order octet of this extended type is either 0x01 or 0x41. The low-order octet of this extended type is TBD.

The Value field consists of two sub-fields:

Global Administrator sub-field: This sub-field contains the router ID of the advertising router that appends the path bandwidth extended community or updates the path bandwidth value of the existing path bandwidth extended community.

Local Administrator sub-field: This sub-field contains the path bandwidth value in IEEE floating point format with units of Gigabytes per second (GB/s).

4. Solution Description

4.1. Adaptive Routing in 3-stage CLOS

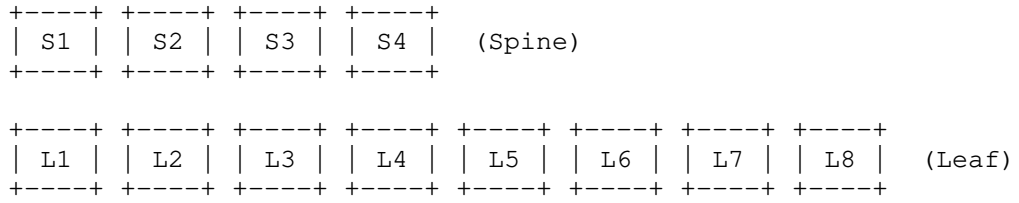


Figure 1

(Note that the diagram above does not include the connections between nodes. However, it can be assumed that leaf nodes are connected to every spine node in their CLOS topology.)

In a three-stage CLOS network as shown in Figure 1, also known as a leaf-spine network, each leaf node would establish eBGP sessions with all spine nodes.

All nodes are enabled for adaptive routing.

When a leaf node, such as L1, advertises the route to a specific IP prefix that it originates, it will attach a transitive path bandwidth extended community filled with a maximum bandwidth value.

Upon receiving the above advertisement, a spine node, such as S1, SHOULD determine the minimum value between the bandwidth of the link towards the advertising node (e.g., L1) and the value of the path bandwidth extended community carried in the received route, and then update the path bandwidth extended community with the above minimum value before readvertising that route to remote eBGP peers. Once S1 receives multiple equal-cost routes for a given prefix from multiple leaf nodes (e.g., L1 and L2 in the server multi-homing scenario), for each route, it SHOULD determine the minimum value between the bandwidth of the link towards the advertising node and the value of the path bandwidth extended community carried in the received route, and then use that minimum bandwidth value as a weight value for that route when performing weighted ECMP. When readvertising the route for that prefix to remote eBGP peers further, the path bandwidth extended community would be updated with the sum of the minimum bandwidth value of each route.

When a leaf node, such as L8, receives multiple equal-cost routes for that prefix from spine nodes (e.g., S1, S2, S3 and S4), for each route, it will determine the minimum value between the bandwidth of the link towards the advertising node and the value of the path bandwidth extended community carried in the received route, and then use that minimum bandwidth value as a weight value for that route when performing weighted ECMP.

Note that the weighted ECMP according to path bandwidth SHOULD NOT be performed unless all equal-cost routes for a given prefix carry the path bandwidth extended community.

4.2. Adaptive Routing in 5-stage CLOS

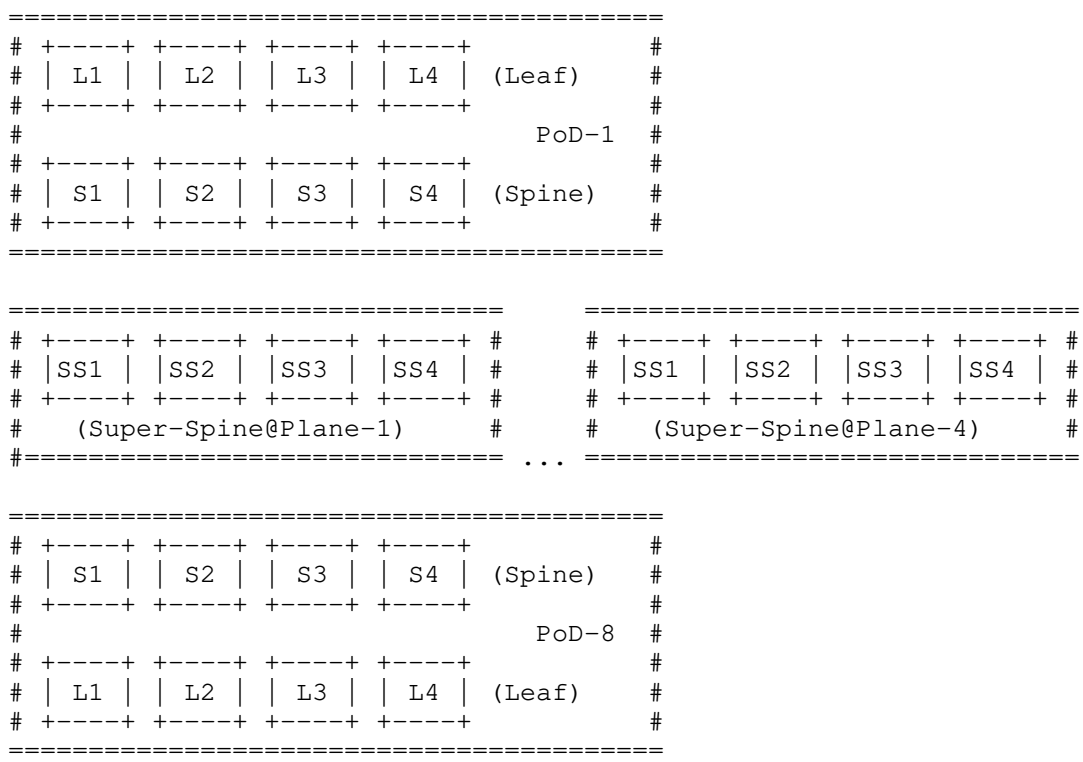


Figure 2

(Note that the diagram above does not include the connections between nodes. However, it can be assumed that the leaf nodes in a given PoD are connected to every spine node in that PoD. Similarly, each spine node (e.g., S1) is connected to all super-spine nodes in the corresponding PoD-interconnect plane (e.g., Plane-1).)

For a five-stage CLOS network as illustrated in Figure 2, each leaf node would establish eBGP sessions with all spine nodes of the same PoD while each spine node would establish eBGP sessions with all super-spine nodes in the corresponding PoD-interconnect plane.

In rail-optimized topology, Intra-rail communication with high bandwidth requirements would be restricted to a single PoD. Inter-rail communication with relatively lower bandwidth requirements need to travel across PoDs through PoD-interconnect planes. Therefore, enabling adaptive routing only in PoD networks is sufficient. It's optional to perform adaptive routing for cross-PoD traffic.

When a leaf node, such as L1 in PoD-1, advertises the route for a specific IP prefix that it originates, it will attach a transitive path bandwidth extended community filled with a maximum bandwidth value.

Upon receiving the above route advertisement, a spine node, such as S1 in PoD-1, will determine the minimum value between the bandwidth of the link towards the advertising node (e.g., L1 in PoD-1) and the value of the path bandwidth extended community carried in the route, and then update the path bandwidth extended community with the above minimum value before readvertising that route to remote eBGP peers. Once S1 in PoD-1 receives multiple equal-cost routes for a given prefix from multiple leaf nodes (e.g., L1 and L2 in PoD-1 in the server multi-homing scenario), for each route, it will determine the minimum value between the bandwidth of the link towards the advertising node and the bandwidth value of the path bandwidth extended community carried in the route, and then use that minimum bandwidth value as a weight value for that route when performing weighted ECMP. When readvertising the route for that prefix to remote eBGP peers, the path bandwidth extended community would be updated with the sum of the minimum bandwidth value of each route.

When a given super-spine node, such as SS1 in Plane-1, receives the route for that prefix from S1 in PoD-1, it will not update the transtive path bandwidth extended community when readvertising that route. It COULD optionally attach another path bandwidth extended community which is non-transitive to indicate the bandwith of the link towards the advertising router.

When a given spine node in another PoD, such as S1 in PoD-8, receives multiple equal-cost routes for a given prefix from super-spine nodes in Plane-1 (e.g., SS1, SS2, SS3 and SS4 in Plane-1), it will not update the value of the transitive path bandwidth extended community when readvertising that route towards remote peers (Note that the transitive path bandwidth extended community of those multiple equal-cost routes carry the same value that was set by S1 in PoD-1). Meanwhile, once each route contains a non-transitive path bandwidth extended community, for each route, it will determine the minimum value between the bandwidth of the link towards the advertising node and the bandwidth value of the non-transitive path bandwidth extended community carried in the route, and then use that minimum bandwidth value as a weight value for that route when performing weighted ECMP.

When a leaf node, such as L8 in PoD-8, receives multiple equal-cost routes for that prefix from multiple spine nodes (e.g., S1, S2, S3 and S4 in PoD-8), for each route, it will determine the minimum value between the bandwidth of the link towards the advertising node and the value of the path bandwidth extended community carried in the route, and then use that minimum bandwidth value as a weight value for that route when performing weighted ECMP.

Note that the weighted ECMP according to path bandwidth SHOULD NOT be performed unless all equal-cost routes for a given prefix carry the path bandwidth extended community.

5. Acknowledgements

TBD.

6. IANA Considerations

TBD.

7. Security Considerations

TBD.

8. References

8.1. 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>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

8.2. Informative References

[I-D.ietf-idr-link-bandwidth] Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", Work in Progress, Internet-Draft, draft-ietf-idr-link-bandwidth-07, 5 March 2018, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-link-bandwidth-07>.

[I-D.xu-lsr-fare]
Xu, X., He, Z., Wang, J., Huang, H., Zhang, Q., Wu, H., Liu, Y., Xia, Y., Wang, P., and S. Hegde, "Fully Adaptive Routing Ethernet", Work in Progress, Internet-Draft, draft-xu-lsr-fare-02, 25 February 2024, <https://datatracker.ietf.org/doc/html/draft-xu-lsr-fare-02>.

[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
China Mobile
Email: xuxiaohu_ietf@hotmail.com

Shraddha Hegde
Juniper
Email: shraddha@juniper.net

Zongying He
Broadcom
Email: zongying.he@broadcom.com

Junjie Wang
Centec
Email: wangjj@centec.com

Hongyi Huang
Huawei
Email: hongyi.huang@huawei.com

Qingliang Zhang
H3C
Email: zhangqingliang@h3c.com

Hang Wu
Ruijie Networks
Email: wuhang@ruijie.com.cn

Yadong Liu
Tencent
Email: zeepliu@tencent.com

Yinben Xia
Tencent
Email: forestxia@tencent.com

Peilong Wang
Baidu
Email: wangpeilong01@baidu.com

Tiezheng
IEIT SYSTEMS
Email: litiezheng@ieisystem.com