

IDR
Internet-Draft
Intended status: Standards Track
Expires: 7 January 2026

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
6 July 2025

Accumulated Metric in NHC attribute
draft-ietf-idr-bgp-generic-metric-01

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 7 January 2026.

Copyright Notice

Copyright (c) 2025 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	3
2. Requirements Language	5
3. Multiple types of metrics in a network	5
4. Discontinuity in end-to-end intent	6
5. Accumulated Metric Encoding	6
6. Leverage Next Hop Dependent Characteristics (NHC) Attribute	8
7. Comparison between AIGP and AMetric	8
8. Usage of Accumulated Metric (AMetric)	9
8.1. Generation of Accumulated Metric	9
8.1.1. Originator of route into BGP	10
8.1.2. Non-Originator of route into BGP	10
8.2. Reception of Accumulated Metric	11
9. Updates to Decision Procedure	12
10. Use-case: Different Metrics across Domains	13
10.1. Scenario 1: Find delay-based end-to-end path.	14
10.2. Scenario 2: Find IGP-metric based end-to-end path leveraging domain-specific path.	15
10.3. Scenario 3: Path selection when a router does not understand the new metric-type.	15
11. Use-case: Different Metric in Data Center (DC) Network	15
12. Deployment Considerations	17
13. Manifestations of Discontinuity in end-to-end path	18
13.1. Handling discontinuous path with AIGP attribute as per RFC7311	18
13.2. Handling discontinuous path with AMetric of NHC attribute	19
13.3. Contiguity Compliance	19
14. Security Considerations	20
15. IANA Considerations	20
16. Limitations of RFC7311	20
17. Acknowledgements	20

18. References	20
18.1. Normative References	20
18.2. Informative References	21
Authors' Addresses	22

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 satisfies 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 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 Classful Transport Planes as described in [I-D.draft-ietf-idr-bgp-classful-transport-planes] 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.

The metric accumulation across domains may be applicable in Data Center Networks. The Data Center networks run IP-Fabric and adopt BGP routing paradigm without running any IGP protocol. Routers in 'n-layer' Clos network have eBGP sessions with routers in adjacent layers and each router has unique Autonomous System number. It would be desirable to determine the cost for an end-to-path across the Data Center Clos network, especially when different metrics are needed.

This document proposes "Accumulated Metric" TLV in the Next-Hop Dependent Characteristics (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.

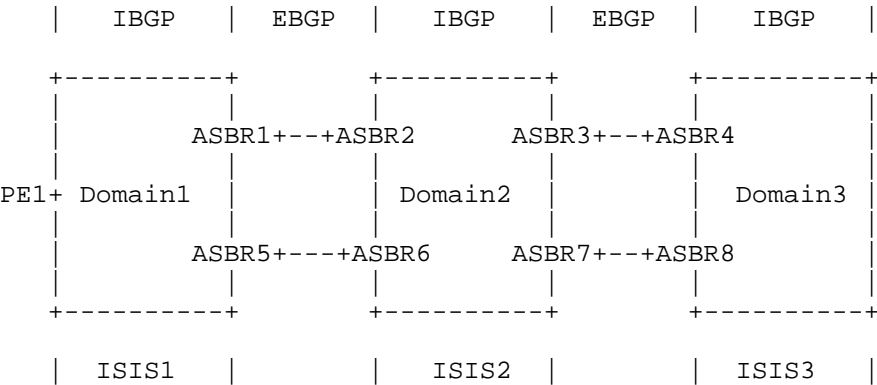


Figure 1: WAN Network

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. Accumulated Metric Encoding

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

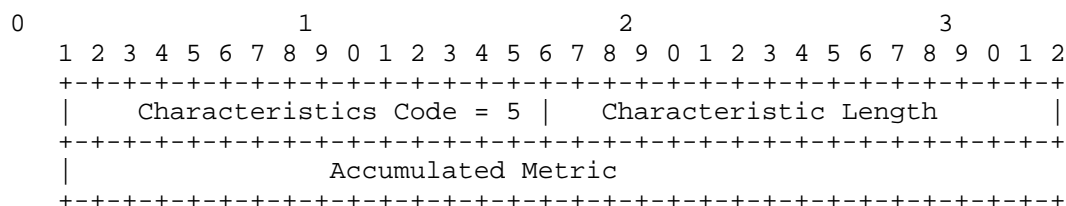


Figure 2: Ametric in NHC

Characteristics Code: 5. See [I-D.ietf-idr-entropy-label]

Characteristics Length: Size of Accumulated Metric(s) in octets.

There can be one or more Accumulated Metric encoded in the NHC. The format of Accumulated Metric is shown below.

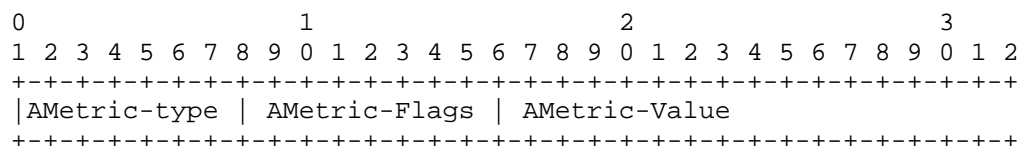


Figure 3: Accumulated Metric format

AMetric-type (1 octet): Code for metric-type from IGP-Protocol registry for metric-types.

AMetric-flags (1 octet): Bits defined below.

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

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

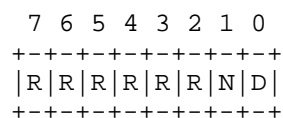


Figure 4: Accumulated Metric flags

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 Characteristics (NHC) Attribute

The Next Hop Dependent Characteristics attribute has two important aspects 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, it will also carry the AMetric for the desired metric-type(s). If any non-originator router upon learning such a route, does not modify the next hop, it will advertise it along with AMetric without modification. If any non-originator router modifies the next hop, it will perform two actions: Encode the next hop address in "Network Address of Next Hop" field of the NHC attribute; and recompute the AMetric and encode it in the NHC attribute, before advertising the route. The NHC's next hop scoping check will help determine if the advertising BGP speaker did not support NHC or AMetric and as a result the AMetric has not been updated.

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 encoded in AIGP attribute, is 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 similar to 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(s) in the NHC attribute, each AMetric to encode 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 metric-types and their metric-values as present in the AMetric and for each of the metric type 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 of metric-flags sub-field MUST be set to 1. If the BGP speaker S does not understand the type of metric, then "D" bit of metric-flags sub-field 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.
- If there is more than one AMetric in the NHC attribute, the first occurrence MUST be processed, and the other occurrences MUST be ignored. The BGP speaker MUST process all metric-types and their values if there is more than one metric type in the AMetric information.
- For each metric in the AMetric, if the BGP speaker S recognizes the metric-type sub-field, it MUST process 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 the received AMetric, 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 the received AMetric, 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 metric types in 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.

When there is more than one metric-types in AMetric, a local policy may provide guidance indicating metric-types as primary and secondary. The secondary metric-type may be used to break the tie among equal cost paths based on primary metric-type.

10. Use-case: Different Metrics across Domains

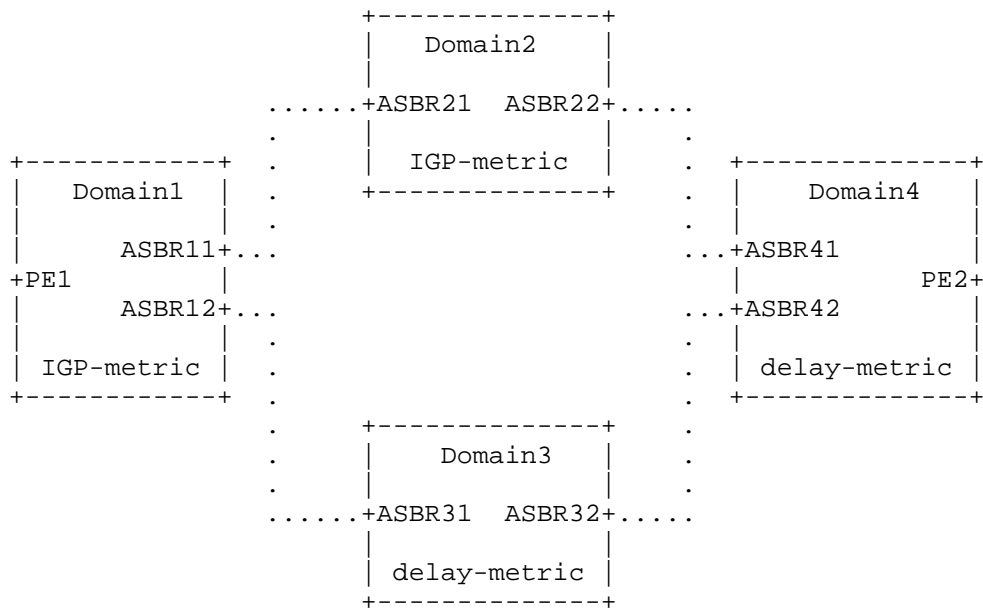


Figure 5: Different metric across network

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. Use-case: Different Metric in Data Center (DC) Network

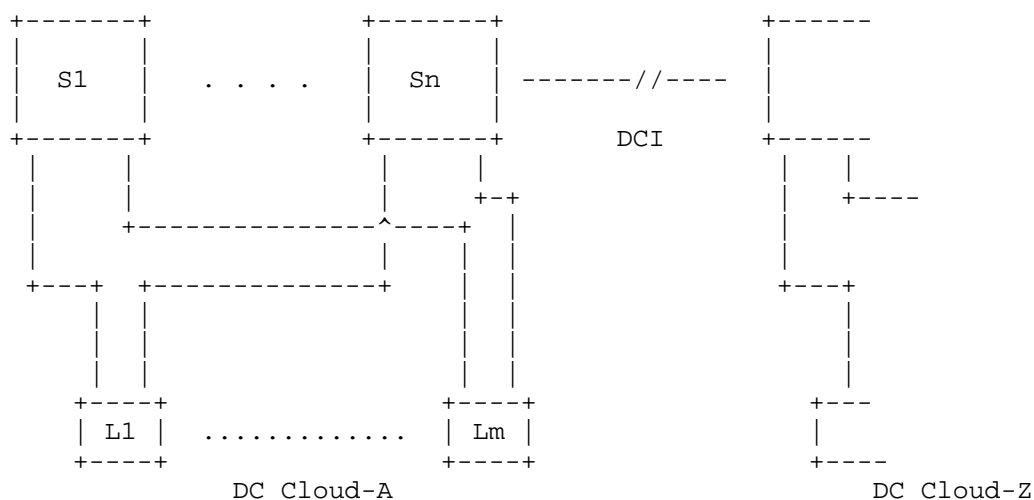


Figure 6: Different metric across DC

Typically, in a Data Center (DC) network all routers at spine layer are connected to all routers at leaf layer. As depicted above, S1..Sn form the spine layer while L1..Lm form the leaf layer in this 3-stage clos network. Routers at leaf layer do single-hop eBGP peering with routers at spine layer. Each router can be in separate Autonomous System by itself. Such DC clouds are connected over Data Center Interconnect (DCI) links. The DCI network can span larger distances as compared to DC cloud.

The cost of the end-to-end path is desirable to make better routing decisions. The metric accumulation happens via AMetric in NHC attribute. This can be used for influencing the best path selection.

Sometimes, the operator may wish to have multiple planes in DC network such that one or more spine router participates in a particular plane and, the links (and the associated BGP peering) connecting the leaf routers with those spine routers become part of that plane. Each plane can choose to adopt a separate metric-type and can independently compute the cost of end-of-end path cost by using AMetric accumulation.

12. 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 Accumulated Metric functionality. In a network without any tunneling technology, it is recommended that all routers MUST support Accumulated 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 AMetric 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.

13. 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.

13.1. 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.

13.2. 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 AMetric even though unrecognized metric is not accumulated. This is Type-C discontinuity. The procedures in this document specifies the "D" bit of the unrecognized AMetric to be set to 1 by the advertising router and hence the receiving router deterministically identifies the discontinuity.

13.3. 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.

14. 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].

15. IANA Considerations

NHC Characteristic Code 5 has been assigned in Section 5 of [I-D.ietf-idr-entropy-label] for the Accumulated Metric characteristic defined in this document. The metric-type field of AMetric refers to the IGP-Protocols registry for metric-type defined in [I-D.ietf-lsr-flex-algo-bw-con]

16. 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.

17. Acknowledgements

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

18. References

18.1. Normative References

[I-D.ietf-idr-entropy-label]

Decraene, B., Scudder, J., Kompella, K., Satya, M. R., Wen, B., Wang, K., and S. Krier, "BGP Next Hop Dependent Characteristics Attribute", Work in Progress, Internet-Draft, draft-ietf-idr-entropy-label-17, 30 March 2025, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-entropy-label-17>>.

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

[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>>.

18.2. 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-04, 31 January 2025, <<https://datatracker.ietf.org/doc/html/draft-hr-spring-intentaware-routing-using-color-04>>.

[I-D.ietf-idr-bgp-car]

Rao, D. and S. Agrawal, "BGP Color-Aware Routing (CAR)", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-car-16, 20 February 2025, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-car-16>>.

[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-39, 28 February 2025, <<https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ct-39>>.

`[I-D.ietf-lsr-flex-algo-bw-con]`

Hegde, S., Britto, W., Shetty, R., Decraene, B., Psenak, P., and T. Li, "IGP Flexible Algorithms: Bandwidth, Delay, Metrics and Constraints", Work in Progress, Internet-Draft, draft-ietf-lsr-flex-algo-bw-con-22, 13 February 2025, <<https://datatracker.ietf.org/doc/html/draft-ietf-lsr-flex-algo-bw-con-22>>.

[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