Interdomain Working Group
Internet-Draft
Updates: 4271 (if approved)
Intended status: Standards Track
Expires: 20 April 2026

O. Vroonen, Ed.
S. Litkowski
Cisco

17 October 2025

BGP best path next-hop selection enhancements
draft-vroonen-idr-bgp-bestpath-nh-selection-00

Abstract

BGP [RFC4271] has originally been designed to carry IPv4 routing information over the Internet. IP routing being “hop-by-hop” in nature, [RFC4271] defines the NEXT_HOP attribute which purpose is to carry the address of the next router to send the IP packet to. In BGP, the next-hop may not be a directly connected router, hence, when evaluating paths, BGP needs to figure out if the next-hop is resolvable and, when needed, needs to figure out what is the internal cost to reach this next-hop.

The incremental use of tunneling technologies to carry traffic between routers (e.g.: GRE, MPLS, SR-MPLS, SRv6…) may violate the assumption that the address carried in the NEXT_HOP attribute is representative of the actual forwarding next-hop. These technologies decouple the BGP control-plane’s view of the next-hop from the data-plane’s actual forwarding endpoint. This document describes the problems that arise from this decoupling. These problems include sub-optimal path selection, incorrect resolvability tracking of the forwarding path leading to traffic drop or misrouting, and others. This document proposes some modification of BGP path selection procedures to accommodate these use cases.

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 20 April 2026.

Table of Contents

1. Introduction

BGP [RFC4271] is designed to exchange network reachability information between routing domains. A BGP update typically contains a prefix and a set of path attributes, including the well-known mandatory NEXT_HOP attribute. The receiving router uses this next-hop address to determine the egress point for traffic destined for the advertised prefix. The assumption is the next hop address represents the next router from a traffic forwarding point of view.

BGP NEXT_HOP attribute as defined in [RFC4271] and BGP NEXT_HOP field in the MP_REACH_NLRI attribute as defined in [RFC4760] are referred to as NEXT_HOP attribute in this document.

[RFC4271] Section 9.1.2.1 defines the route resolvability condition: a BGP route is considered unresolvable if the BGP speaker's routing table has no route matching the BGP route next-hop address. As per [RFC4271] Section 9.1.2.2 e), when comparing paths received via internal BGP (IBGP), the Routing Table metric associated to the BGP route next-hop address is used to figure out the best path.

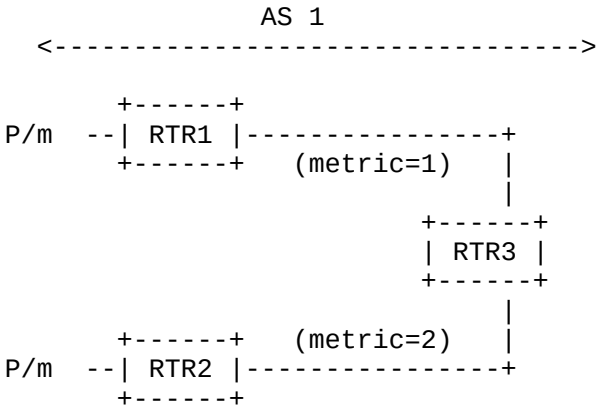

Figure 1

In Figure 1, RTR3 receives prefix P/m from RTR1 and RTR2 (IBGP peers) with NEXT_HOP N1 and N2. RTR3 resolves N1 with a cost of 1 and N2 with a cost of 2. Based on [RFC4271] procedures, RTR3 will select the path with the lowest cost to the NEXT_HOP address, so RTR3 will select the path received from RTR1. In this example, the path used to resolve N1 and N2 reflects the actual forwarding path which makes the best path decision done by RTR3 accurate.

This document describes cases where the NEXT_HOP attribute used in the BGP update is not representative of the actual forwarding path. In these cases, the resolvability condition may fail in its goal and path selection may be done on inaccurate criteria leading to suboptimal routing, network congestion, traffic drop or misrouting...

These use cases are not new and may have been partially addressed by IETF standards or drafts, some references are provided below:

* [I-D.ietf-idr-bgp-bestpath-selection-criteria] addresses the case of MPLS networks and proposes a modification of the route resolvability condition to be performed using forwarding database of a particular data plane protocol. It also proposes an optional data path verification.

* [RFC9012] Section 7 improves the route resolvability condition by verifying that there is a feasible tunnel. However, it doesn't take into account that the cost associated with the tunnel may be different from the cost associated with the BGP next-hop.

* [RFC9252] defines SRv6 overlay services signaling using BGP. The procedures involve the advertisement of an SRv6 Service TLV within the BGP Prefix SID attribute to signal the SID to be used for forwarding. RFC9252 highlights that ingress PE must perform a resolvability check for the SRv6 SID in addition to the resolvability check done on the NEXT_HOP attribute. However, it doesn't take into account that the cost associated with the tunnel may be different from the cost associated with the BGP next-hop.

This document defines generic modifications to the BGP decision process that can apply to all the use cases.

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. Use cases

2.1. MPLS path

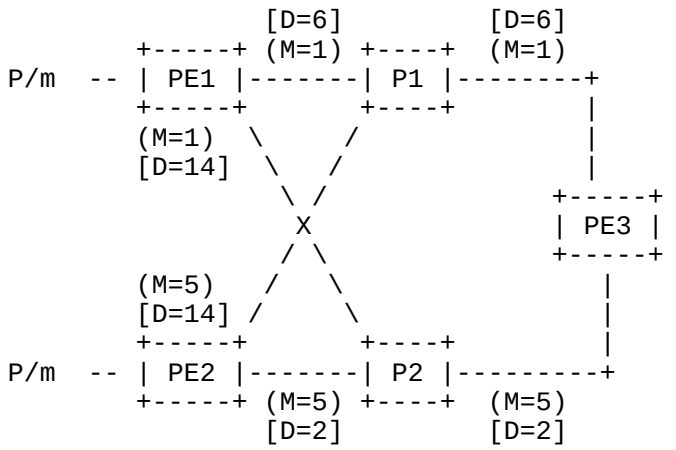

Figure 2: Next hop attribute not used for forwarding

In Figure 2, a BGP MPLS VPN (as defined in [RFC4364]) is deployed over PEs. Let's assume PE3 receives prefix P/m from PE1 and PE2 with NEXT_HOP N1 and N2. The existence of a route in PE3's IPv4 routing table to reach N1 or N2 (defined in [RFC4271]) is not sufficient to ensure that traffic can be carried from PE3 to PE1 or PE2. PE3 needs to ensure that there is a tunnel available to reach PE1 or PE2 that can carry MPLS traffic (e.g.: an MPLS Label Switched Path (LSP)). By checking only for the existence of a route in its routing table, PE3 could create a traffic drop or misrouting if there is no tunnel to carry the MPLS VPN traffic.

2.2. Segment Routing Traffic Engineering Policy

When using SR Policies [RFC9256], the BGP NEXT_HOP attribute may not accurately represent the actual forwarding path.

In Figure 2, a BGP VPN service (as defined in [RFC4364]) is deployed between the PEs. Let's assume PE3 receives prefix P/m from PE1 and PE2 with NEXT_HOP N1, N2 and color C1. On PE3, SR Policies (C1, N1) and (C1, N2) are configured to use the low latency path.

* Costs M in parentheses represents cost of IGP.

* Costs D in brackets represents cost of latency-based metric

If PE3 performs the next-hop resolution solely based on the BGP NEXT_HOP attribute, it will not verify that an SR policy (C1, Nx) exists and is up before using it. It may also pick the wrong best path by considering the cost to the NEXT_HOP address instead of the

cost of the SR policy. The best path considering SR-policy (C1, Nx) cost is via PE2, while the best path considering IGP metric is via PE1.

Prefix	Next hop	Cost
N1	P1	2
N2	P1	6

Table 1: IPv4 routing table of PE3

SR Policy	Next hop	Cost	State
(C1, N1)	P1	12	Up
(C1, N2)	P2	4	Up

Table 2: SR Policies of PE3

Considering that PE3 may take into account the cost of the SR policy and then picks up the path from PE2 as best. If the SR Policy (C1, N2) is down or not present, and considering that color C1 is using CO-flag equal to 00 ([RFC9246] Section 8.8.1), path from PE2 is authorized to use IGP path to next-hop N2 (fallback to IGP path via P1 with cost=6) as there is no policy (C1, N2) available. This situation has multiple drawbacks:

* Cost of the two BGP paths are not comparable which makes the comparison inaccurate: PE1 path uses a cost based on latency while cost of path from PE2 uses IGP metric.

* Using a path without an SR-policy may violate the intent of the service, especially if there is an alternate path (from PE1) that fulfills the intent.

It may be desirable for PE3 to prefer the path from PE1 that satisfies color C1 or even exclude any path that does not meet the color C1 requirement.

2.3. SRv6 services

When using SRv6 services as defined in [RFC9252], the BGP NEXT_HOP attribute may not be representative of the actual forwarding path.

Considering Figure 2, an SRv6-based BGP VPN is deployed between the PEs. PE1 uses locator L1 for algorithm 0 and L1_FA for flexible algorithm 128 ([RFC9350]) optimized for low-latency. Similarly, PE2 uses locator L2 for algorithm 0 and L2_FA for flexible algorithm 128. Let’s assume PE3 receives prefix P/m from PE1 and PE2 with NEXT_HOP N1, N2 and SRv6 SID S1 and S2. SID S1 and S2 are allocated respectively from L1_FA and L2_FA.

* Costs M in parentheses represents cost of algorithm 0 (as defined in [RFC8665] Section 8.5).

* Costs D in brackets represents cost of flexible algorithm [RFC9350] 129 which is using low-latency metric.

The IPv6 routing table of PE3 contains the following entries:

Prefix	Next hop	Cost
N1	P1	2
N2	P1	6
L1	P1	2
L2	P1	6
L1_FA	P1	12
L2_FA	P2	4

Table 3: IPv6 routing table of PE3

Based on [RFC4271] procedures, PE3 verifies that NEXT_HOP address of each path is resolvable. Based on [RFC9252] procedures, PE3 verifies that S1 and S2 addresses are resolvable. PE3 will then select the path with the lowest cost to the NEXT_HOP address according to [RFC4271] . Cost to N1 is lower than cost to N2, so PE3 will select the path received from PE1. However, from a latency perspective, path to PE2 is the best one.

The problem of path suboptimality may also happen with algorithm 0, if for instance SRv6 traffic for algorithm 0 needs to be offloaded from PE1, operator may increase the metric of the locator (while not changing the metric of the next-hop) on PE1. In the example above, if L1 is advertised by PE1 with an offset of 1M, then PE3 will have to cost to L1 of 1000002 but N1 will still be reachable with a cost of 2. The cost to reach the SID must also be taken into account in this scenario to ensure that the traffic offload works properly.

3. Modification of the BGP route selection process

3.1. Forwarding address

This document defines the forwarding address as the IP address of the next router to which packets are sent. The forwarding address may come from the NEXT_HOP or a different address which has been signaled in a different attribute along with the path.

The following data already defined in BGP standards SHOULD be considered as forwarding addresses:

* The Tunnel Egress Endpoint Sub-TLV contained in Tunnel Encapsulation attribute defined in [RFC9012]

* SRv6 SID Information Sub-TLV contained within L3 or L2 SRv6 SID defined in [RFC9252]

Unless a BGP update contains another type of forwarding address, the BGP NEXT_HOP attribute is considered as the forwarding address.

The forwarding address MAY be complemented by a forwarding context. The forwarding context characterizes more the forwarding path to be used. The following attributes defined in BGP standards are examples of forwarding context:

* The Tunnel Encapsulation attribute defined in [RFC9012]

* The Color extended community defined in [RFC9012]

* The SRv6 Sub-TLV and Sub-sub-TLV contained within L3 or L2 SRv6 SID TLV as defined in [RFC9252] other than the SID itself

Each new BGP extension SHOULD specify if an address carried by the extension must be considered as a forwarding address. The procedures defined in the next section SHOULD apply for any new forwarding address defined without having to redefine them.

3.2. Resolution constraints

When the forwarding address and its context are identified for a BGP route, the implementation may know the required characteristics of the route to be used to resolve the forwarding address. Depending on the type of forwarding address and context, the implementation may need to ensure that the forwarding address is resolved through a specific type of route in a specific table. This resolution constraint may come from the forwarding context and/or may be configured locally.

In some cases, operator may want to enforce that the forwarding address is resolved through a specific type of route. This can be achieved by configuring a local resolution constraint. Reusing the example defined in Section 2.1, if we consider an IPv4 unicast BGP service carried over a BGP free-core, the BGP NEXT_HOP must be reachable through a tunnel to allow the end-to-end packet delivery. Such a case cannot be derived from the BGP update context and BGP must be configured to resolve the NEXT_HOP only through tunnels (of any or specific types). Similarly, when aggregate routes are present in the routing table, user may want to prevent the forwarding address (which is a specific route) to be resolved over the aggregate routes. A resolution constraint based on prefix/mask can be done to avoid such resolution to be valid. An implementation MAY provide a set of configuration options for resolution constraints.

3.3. Route resolvability condition

This document updates [RFC4271] Section 9.1.2.1 as follows:

* The route resolvability check for the BGP NEXT_HOP attribute MUST continue to be done.

* In addition, the route resolvability criteria SHOULD be performed based on the forwarding address.

* When resolving the forwarding address, look up SHOULD be performed by applying the resolution constraints defined in Section 3.2.

The resolvability check based on the forwarding address MAY be enabled through a configuration knob.

3.4. Internal cost determination

For a prefix P/m, different BGP paths may use different forwarding addresses and contexts of various types.

```
P/m
    Path1: NH=10.0.0.1 Color: 200
    Path2: NH=2001::2 SRv6-SID: cafe:0:2:e002::
    Path3: NH=10.0.0.3 Tunnel-encap(L2TPv3, endpoint: 10.0.0.3)
    Path4: NH=2001::4 SRv6-SID: cafe:0:4:e002::
```

Figure 3

Costs retrieved from different types of forwarding addresses or contexts may not be comparable because they are based on different sets of rules. For instance, path1 may leverage an SR policy (color 200, endpoint R1) optimizing for latency, so the cost of path1 will reflect the latency to R1. Path2 may use the IGP cost to R2. Path3 may have no cost. These values are not directly comparable.

In order to compare the paths, this document introduces the concept of forwarding address preference. The preference is a local numerical value. An implementation SHOULD pick the lowest value as the most preferred.

This document updates [RFC4271] Section 9.1.2.2 e) as follows:

* Remove from consideration any routes with an highest forwarding address preference value. This preference MAY be retrieved from the resolution lookup of the forwarding address or MAY be configured locally.

* For remaining paths, the interior cost of a route is determined by the metric of the resolving route to the forwarding address applying the resolution constraints defined in Section 3.2. If the forwarding address for a route is reachable, but no cost can be determined, the cost SHOULD be set by default to the maximum allowed cost.

Using forwarding preference and forwarding address-based cost SHOULD be enabled through a configuration knob.

With the example above and the internal tables defined below, and considering lowest preference value as the most preferred one, BGP would select Path2 as best. Path2 and Path4 have the lowest preference (10), then Path2 has the lowest internal cost (12).

```
    P/m
        Path1: NH=10.0.0.1 Color: 200,
               preference 100 (from table), cost 1001
        Path2: NH=2001::2 SRv6-SID: cafe:0:2:e002::,
               preference 10 (from BGP), cost 12
        Path3: NH=10.0.0.3
               Tunnel-encap(L2TPv3, endpoint: 10.0.0.3, sessID: 1),
               preference 1000 (from BGP), cost max
        Path4: NH=2001::4 SRv6-SID: cafe:0:4:e002::,
               preference 10 (from BGP), cost 14
```

Figure 4

Prefix, Color	Preference	Metric	Forwarding data
10.0.0.1, 200	10	1001	interface IF1, label stack {L1, L2, L3}
10.0.0.2, 200	10	1002	interface IF2, label stack {L4, L5}
10.0.0.3, 200	10	1003	interface IF1, label stack {L6, L7, L8, L9}

Table 4: IPv4 Color routing table

Prefix	Preference	Metric	Forwarding data
cafe:0:1::/48	5	11	interface IF1, label stack {L1, L2, L3}
cafe:0:2::/48	5	12	interface IF2, label stack {L4, L5}
cafe:0:3::/48	5	13	interface IF1, label stack {L6, L7, L8, L9}
cafe:0:4::/48	5	14	interface IF1, label stack {L10, L11, L12}

Table 5: IPv6 routing table

Destination	Session ID	Status
10.0.0.3	1	up

Table 6: L2TP session table

Type	Preference
MPLS LSP (any signaling)	100
MPLS RSVP-TE LSP	50
SRv6 SID	10
Default	use value from table lookup, use 1000 if table provided no value

Table 7: BGP forwarding address preference configuration

3.5. Next hop and forwarding address tracking

A BGP speaker SHOULD track the resolvability of both the NEXT_HOP attribute and the forwarding address. If either the NEXT_HOP or the forwarding address becomes unresolvable or if the cost to reach either the NEXT_HOP or the forwarding address changes, the BGP speaker MUST re-evaluate the best path selection for all prefixes using the affected NEXT_HOP or forwarding address. This tracking MUST be done for all paths, including the best path and non-best paths.

4. Example

The example below illustrates the logic of forwarding address preference and cost comparison.

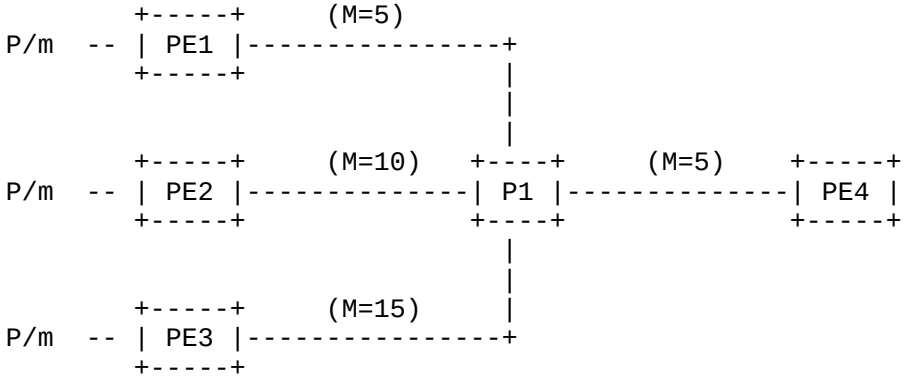

Figure 5

In Figure 5, a prefix P/m is reachable by PE4 from PE1, PE2, PE3 (with respectively NEXT_HOP N1, N2 and N3). Considering that the network is a BGP free core, traffic must be tunneled between edge devices. Traffic destined to P/m is of high bandwidth and requires traffic-engineering to spread the traffic across the available links of the core. RSVP-TE is used to provide traffic-engineering MPLS tunnels. SR-MPLS is also used to provide best-effort reachability. BGP is configured to use the route preference (or administrative distance) from table lookup as forwarding address preference. RSVP-TE is given a better route preference than SR-MPLS. RSVP-TE tunnel to PE3 cannot be established.

Prefix	Protocol	Preference	Cost
N1	IS-IS	100	10
N2	IS-IS	100	15
N3	IS-IS	100	20

Table 8: IPv4 routing table of PE4

Prefix	Protocol	Preference	Cost
N1	IS-IS SR	110	10
N1	RSVP-TE	250	1000
N2	IS-IS SR	110	15
N2	RSVP-TE	250	100
N3	IS-IS SR	110	20

Table 9: MPLS ingress tunnel table of PE4

As mentioned in Section 3.2, the case of BGP free-core requires BGP on PE4 to be configured to allow the resolution the NEXT_HOP address through tunnels (of any type). Considering that PE4 maintains a separate table for MPLS ingress tunnels, PE4 will look up for N1, N2, N3 addresses only in this table. PE4 will first check the resolvability of N1, N2 and N3. All are resolvable in the MPLS ingress tunnel table. PE4 will end-up with the following information from the MPLS ingress tunnel table to compare the path:

```
P/m
  Path1:
    from PE1, NH=N1
    cost 1000, forwarding address preference 250

  Path2:
    from PE2, NH=N2
    cost 100, forwarding address preference 250

  Path3:
    from PE3, NH=NH3
    cost 20, forwarding address preference 110
```

Figure 6

PE4 will check the forwarding address preference of the paths. PE4 will not consider the paths received from PE3 because the forwarding address preference is lower than others. Finally, PE4 will compare the internal costs between paths from PE1 and PE2 as they have the same preference and path from PE2 will be elected as best because it has the lowest cost.

As mentioned in Section 3.5, if the RSVP-TE tunnel to PE2 goes down, PE4 will re-evaluate the best path selection and will select the path from PE1 as best. This is also true if the cost of RSVP-TE LSP to reach N1 changes and becomes lower than the cost of LSP to reach N2.

5. IANA Considerations

This document does not require any IANA actions.

6. Security Considerations

This document modifies BGP route selection process by using data other than the next-hop address to perform the resolvability check as well as to compute the internal cost. This does not add any security consideration compared to using the existing NEXT_HOP attribute defined in [RFC4271].

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>.

[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>.

[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>.

[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/info/rfc4760>.

[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>.

[RFC8665] Psenak, P., Ed., Previdi, S., Ed., Filsfils, C., Gredler, H., Shakir, R., Henderickx, W., and J. Tantsura, "OSPF Extensions for Segment Routing", RFC 8665, DOI 10.17487/RFC8665, December 2019, <https://www.rfc-editor.org/info/rfc8665>.

[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/info/rfc9012>.

[RFC9246] van Brandenburg, R., Leung, K., and P. Sorber, "URI Signing for Content Delivery Network Interconnection (CDNI)", RFC 9246, DOI 10.17487/RFC9246, June 2022, <https://www.rfc-editor.org/info/rfc9246>.

[RFC9252] Dawra, G., Ed., Talaulikar, K., Ed., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "BGP Overlay Services Based on Segment Routing over IPv6 (SRv6)", RFC 9252, DOI 10.17487/RFC9252, July 2022, <https://www.rfc-editor.org/info/rfc9252>.

[RFC9256] Filsfils, C., Talaulikar, K., Ed., Voyer, D., Bogdanov, A., and P. Mattes, "Segment Routing Policy Architecture", RFC 9256, DOI 10.17487/RFC9256, July 2022, <https://www.rfc-editor.org/info/rfc9256>.

[RFC9350] Psenak, P., Ed., Hegde, S., Filsfils, C., Talaulikar, K., and A. Gulko, "IGP Flexible Algorithm", RFC 9350, DOI 10.17487/RFC9350, February 2023, <https://www.rfc-editor.org/info/rfc9350>.

7.2. Informative References

[I-D.ietf-idr-bgp-bestpath-selection-criteria] Asati, R., "BGP Bestpath Selection Criteria Enhancement", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-bestpath-selection-criteria-12, 5 June 2019, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-bestpath-selection-criteria-12>.

Acknowledgements

The authors would like to acknowledge Ketan Talaulikar, Serge Krier and Shyam Sethuram for review and comments.

Authors' Addresses

Olivier Vroonen (editor)
Cisco
Email: ovroonen@cisco.com

Stephane Litkowski
Cisco
Email: slitkows@cisco.com