Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 2 September 2026

X. Xu
China Mobile
S. Hegde
HPE
K. Talaulikar
Individual
M. Boucadair
C. Jacquenet
France Telecom
J. Dong
Huawei
1 March 2026

BGP Performance-aware Routing Mechanism
draft-ietf-idr-performance-routing-06

Abstract

The current Border Gateway Protocol (BGP) specification does not incorporate network performance metrics, such as network latency, into its route selection process. This document outlines a performance-aware BGP routing mechanism that integrates network latency as a critical criterion for route selection. This innovative approach is particularly beneficial for server providers with a global presence, enabling them to offer low-latency network connectivity service as a value-added service to their customers.

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].

This Internet-Draft will expire on 2 September 2026.

Copyright Notice

Table of Contents

1. Introduction

Cloud and/or network service providers (service providers in short) with global reach aim to deliver low-latency network connectivity service to their customers as a competitive advantage. Sometimes, the network connectivity may travel across more than one Autonomous System (AS) under their administration, which usually spans multiple continents. However, the BGP [RFC4271] protocol, which is used for path selection across ASes, doesn't use the network latency metric in the route selection process. As such, the best route selected based on the existing BGP route selection criteria may not be the best from the customer experience perspective.

This document describes a performance-aware BGP routing paradigm in which the network latency metric is disseminated via a new TLV of the AIGP attribute [RFC7311] and then is used as an input to the route selection process. This mechanism is useful for those server providers with global reach, which usually own more than one AS, to deliver low-latency network connectivity service to their customers.

Furthermore, to ensure backward compatibility with existing BGP implementations and maintain the stability of the overall routing system, it is expected that the performance-aware routing paradigm could coexist with the vanilla routing paradigm. As such, service providers could provide low-latency network connectivity service as a value-added service while still offering the vanilla routing service to meet customers’ different requirements.

For the sake of simplicity, this document considers only one network performance metric: the network latency metric. The support of multiple network performance metrics is out of scope of this document. In addition, this document focuses exclusively on BGP matters, and therefore all BGP-irrelevant matters, such as the mechanisms for measuring network latency are outside the scope of this document.

The performance-aware BGP routing paradigm has been successfully implemented in SONiC and is set to be open-sourced shortly. In addition, a variant of this performance-aware BGP routing paradigm has been implemented as well (see http://www.ist-mescal.org/roadmap/qbgp-demo.avi).

2. Terminology

This memo makes use of the terms defined in [RFC4271].

Network latency indicates the amount of time it takes for a packet to traverse a given network path [RFC2679]. Provided a packet is forwarded along a path that contains multiple links and routers, the network latency would be the sum of the transmission latency of each link (i.e., link latency), plus the sum of the internal delay occurred within each router (i.e., router latency) which includes queuing latency and processing latency. The sum of the link latency is also known as the cumulative link latency. In today’s service provider networks which usually span a wide geographical area, the cumulative link latency becomes the major part of the network latency since the total of the internal latency occured within each high-capacity router seems trivial compared to the cumulative link latency. In other words, the cumulative link latency could approximately represent the network latency in the above networks.

Furthermore, since the link latency is more stable than the router latency, the approximate network latency represented by the cumulative link latency is also more stable. Therefore, if there was a way to calculate the cumulative link latency of a given network path, it is strongly recommended to use such cumulative link latency to approximately represent the network latency. Otherwise, the network latency would have to be measured frequently by some means (e.g., PING or other measurement tools).

3. Performance-aware Route Advertisement

Performance-aware (i.e., latency-aware in the context of this document) routes SHOULD be exchanged between BGP peers by means of a specific Subsequent Address Family Identifier (SAFI) of TBD (see IANA Section) and also be carried as labeled routes as per [RFC3107]. To some extent, performance-aware routes can then be looked as specific labeled routes which are associated with the network latency metric.

A BGP speaker SHOULD NOT advertise performance-aware routes to a particular BGP peer unless that peer indicates, through BGP capability advertisement (see Section 4), that it can process update messages with that specific SAFI field.

Network latency metrics are attached to the performance-aware routes via a new TLV of the AIGP attribute, referred to as NETWORK_LATENCY TLV. The value of this TLV indicates the network latency in microseconds from the BGP speaker depicted by the NEXT_HOP path attribute to the address depicted by the NLRI prefix. The type code of this TLV is TBD (see IANA Section), and the value field is 4 octets in length. In some abnormal cases, if the cumulative link latency exceeds the maximum value of 0xFFFFFFFF, the value field SHOULD be set to 0xFFFFFFFF. Note that the NETWORK_LATENCY TLV MUST NOT co-exist with the AIGP TLV within the same AIGP attribute.

A BGP speaker SHOULD be configurable to enable or disable the origination of performance-aware routes. If enabled, a local network latency value for a given to-be-originated performance-aware route MUST be configured to the BGP speaker so that it can be filled in the NETWORK_LATENCY TLV of that performance-aware route.

A BGP speaker that is enabled to process NETWORK_LATENCY but was not provisioned with the local network latency value SHOULD set the value of the NETWORK_LATENCY attribute to zero when it advertises the corresponding route that it originated.

When distributing a performance-aware route learnt from a BGP peer, if this BGP speaker has set itself as the NEXT_HOP of such route, the value of the NETWORK_LATENCY TLV SHOULD be increased by adding the network latency from itself to the previous NEXT_HOP of such route. Otherwise, the NETWORK_LATENCY TLV of such route MUST NOT be modified.

As for how to obtain the network latency to a given BGP NEXT_HOP, this is outside the scope of this document. However, note that the path latency to the NEXT_HOP SHOULD approximately represent the network latency of the exact forwarding path towards the NEXT_HOP. For example, if a BGP speaker uses a Traffic Engineering (TE) Label Switching Path (LSP) or a SR policy route [RFC9256] from itself to the NEXT_HOP, rather than the shortest path calculated by the Interior Gateway Protocol (IGP), the latency to the NEXT_HOP SHOULD reflect the network latency of that TE LSP path or SR policy route , rather than the IGP shortest path. In cases where the latency to the NEXT_HOP could not be obtained due to some reason(s), that latency SHOULD be set to 0xFFFFFFFF by default.

To keep performance-aware routes stable enough, a BGP speaker SHOULD use a configurable threshold for network latency fluctuation to avoid sending any update which would otherwise be triggered by a minor network latency fluctuation below that threshold.

4. Capability Advertisement

A BGP speaker that uses multiprotocol extensions to advertise performance-aware routes SHOULD use the Capabilities Optional Parameter, as defined in [RFC5492], to inform its peers about this capability.

The MP_EXT Capability Code, as defined in [RFC4760], is used to advertise the (AFI, SAFI) pairs available on a particular connection.

A BGP speaker that implements the Performance-aware Routing Capability MUST support the BGP labeled route capability by default. In other words, a BGP speaker that advertises the Performance-aware Routing Capability to a peer using BGP Capabilities advertisement [RFC5492] does not have to advertise the BGP labeled route capability to that peer explicitly.

5. Performance-aware Route Selection

Performance-aware route selection only requires the following modification to the tie-breaking procedures of the BGP route selection decision (phase 2) described in [RFC4271]: the network latency metric comparison SHOULD be executed just ahead of the AS-Path Length comparison step. Prior to executing the network latency metric comparison, the value of the NETWORK_LATENCY TLV SHOULD be increased by adding the network latency from the BGP speaker to the

NEXT_HOP of that route.

The Loc-RIB of the performance-aware routing paradigm is independent of that of the vanilla routing paradigm. Accordingly, the routing table of the performance-aware routing paradigm is independent of that of the vanilla routing paradigm.

Whether the performance-aware routing paradigm or the vanilla routing paradigm would be applied to a given packet is a local policy issue which is outside the scope of this document. For example, by leveraging the color-based BGP route revolution method, those service routes marked with a certain color could be resolved over the performance-aware routes marked with the same color, which in turn could be resolved over the intra-AS routes (e.g., SR policy routes [RFC9256]) marked with the same color. Alternatively, by leveraging the Cos-Based Forwarding (CBF) capability which allows routers to have distinct routing and forwarding tables for each type of traffic, the selected performance-aware routes could be installed in the routing and forwarding tables corresponding to high-priority traffic.

5.1. Deployment Considerations

This section is not normative.

Enabling performance-aware BGP routing at large (i.e., among domains that do not belong to the same administrative entity) may be conditioned by other administrative settlement considerations that are out of the scope of this document. Nevertheless, this document does not require nor exclude activating the proposed route selection scheme between domains managed by distinct administrative entities.

The main deployment case targeted by this specification is where involved domains are managed by the same administrative entity. Concretely, this performance-aware BGP routing mechanism can advantageously be enabled in a multi-domain environment, where all the involved domains are operated by the same administrative entity so that the processing of low-latency routes can be consistent throughout the domains. Besides security considerations that may arise (which are further discussed in Section 9), there is indeed a need to consistently enforce a performance-aware BGP routing policy within a set of domains that belong to the same administrative entity. This is motivated by the processing of traffic which is of very different nature and may have different QoS requirements. For instance, a BGP color extended community could be attached to the performance-aware routes so as to associate it with a low-latency Segment Routing (SR) policy route towards the BGP NEXT_HOP that is configured with the same color. In this way, traffic matching the performance-aware BGP routes would be forwarded to the BGP NEXT_HOP via the low-latency SR policy routes towards that BGP NEXT_HOP. Alternatively, the combined use of BGP performance-aware routing with traffic engineering tools that would lead to the computation and establishment of traffic-engineered paths between "performance-aware-routing"-enabled BGP peers based upon the manipulation of the Unidirectional Link delay sub-TLV [RFC7810] [RFC7471] would contribute to guaranteeing the overall consistency of the low -atency information within each domain.

In network environments where router reflectors are deployed but next-hop-self is disabled on them, route reflectors usually reflect those received routes which are optimal (i.e., lowest latency) from their perspectives but may not be optimal from the receivers' perspectives. Some existing solutions, as described in [RFC7911], [I-D.ietf-idr-bgp-optimal-route-reflection], and [RFC6774], can be used to address this issue.

6. Contributors

Ning So
Reliance
Email: Ning.So@ril.com

Yimin Shen
Juniper
Email: yshen@juniper.net

Uma Chunduri
Huawei
Email: uma.chunduri@huawei.com

Hui Ni
Huawei
Email: nihui@huawei.com

Yongbing Fan
China Telecom
Email: fanyb@gsta.com

Luis M. Contreras
Telefonica I+D
Email: luismiguel.contrerasmurillo@telefonica.com

7. Acknowledgements

Thanks to Joel Halpern, Alvaro Retana, Jim Uttaro, Robert Raszuk, Eric Rosen, Bruno Decraene, Qing Zeng, Jie Dong, Mach Chen, Saikat Ray, Wes George, Jeff Haas, John Scudder, Stephane Litkowski and Sriganesh Kini for their valuable comments on this document. Special thanks should be given to Jim Uttaro and Eric Rosen for their proposal of using a new TLV of the AIGP attribute to convey the network latency metric. Thanks Shawn Zhang for proposing the new name of this performance-based BGP routing paradigm: Performance-aware Routing, abbreviated as PAR.

8. IANA Considerations

A new BGP Capability Code for the Performance-aware Routing Capability, a new SAFI specific for performance-aware routing paradigm and a new type code for the NETWORK_LATENCY TLV of the AIGP attribute are required to be allocated by IANA.

9. Security Considerations

In addition to the considerations discussed in [RFC4271], the following items should be considered as well:

a. Tweaking the value of the NETWORK_LATENCY by an illegitimate party may influence the route selection results. Therefore, the Performance-aware Routing Capability negotiation between BGP peers which belong to different administration domains MUST be disabled by default. Furthermore, a BGP speaker MUST discard all performance-aware routes received from the BGP peer for which the Performance-aware Routing Capability negotiation has been disabled.

b. Frequent updates of the NETWORK_LATENCY TLV may have a severe impact on the stability of the routing system. Such practice SHOULD be avoided by setting a reasonable threshold for network latency fluctuation.

10. References

10.1. 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>.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <https://www.rfc-editor.org/info/rfc3107>.

[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>.

[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>.

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

10.2. Informative References

[I-D.ietf-idr-bgp-optimal-route-reflection] Raszuk, R., Decraene, B., Cassar, C., Aman, E., and K. Wang, "BGP Optimal Route Reflection (BGP ORR)", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-optimal-route-reflection-28, 17 June 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-optimal-route-reflection-28>.

[RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, DOI 10.17487/RFC2679, September 1999, <https://www.rfc-editor.org/info/rfc2679>.

[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>.

[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>.

[RFC6774] Raszuk, R., Ed., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of Diverse BGP Paths", RFC 6774, DOI 10.17487/RFC6774, November 2012, <https://www.rfc-editor.org/info/rfc6774>.

[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>.

[RFC7810] Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 7810, DOI 10.17487/RFC7810, May 2016, <https://www.rfc-editor.org/info/rfc7810>.

[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>.

[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>.

Authors' Addresses

Xiaohu Xu
China Mobile
Email: xuxiaohu_ietf@hotmail.com

Shraddha Hegde
HPE
Email: shraddha.hegde@hpe.com

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

Mohamed Boucadair
France Telecom
Email: mohamed.boucadair@orange.com

Christian Jacquenet
France Telecom
Email: christian.jacquenet@orange.com

Jie
Huawei
Email: jie.dong@huawei.com