

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 12, 2020

X. Xu
Alibaba, Inc
S. Hegde
Juniper
K. Talaulikar
Cisco
M. Boucadair
C. Jacquenet
France Telecom
October 10, 2019

Performance-based BGP Routing Mechanism
draft-ietf-idr-performance-routing-02

Abstract

The current BGP specification doesn't use network performance metrics (e.g., network latency) in the route selection decision process. This document describes a performance-based BGP routing mechanism in which network latency metric is taken as one of the route selection criteria. This routing mechanism is useful for those server providers with global reach to deliver low-latency network connectivity services 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](#)].

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 April 12, 2020.

Copyright Notice

Copyright (c) 2019 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 Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1.	Introduction	2
2.	Terminology	3
3.	Performance Route Advertisement	4
4.	Capability Advertisement	5
5.	Performance Route Selection	5
6.	Deployment Considerations	6
7.	Contributors	7
8.	Acknowledgements	8
9.	IANA Considerations	8
10.	Security Considerations	8
11.	References	8
11.1.	Normative References	8
11.2.	Informative References	9
	Authors' Addresses	10

[1.](#) Introduction

Network latency is widely recognized as one of major obstacles in migrating business applications to the cloud since cloud-based applications usually have very clearly defined and stringent network latency requirements. Service providers with global reach aim at delivering low-latency network connectivity services to their cloud service customers as a competitive advantage. Sometimes, the network connectivity may travel across more than one Autonomous System (AS) under their administration. However, the BGP [[RFC4271](#)] which is used for path selection across ASes doesn't use network latency in the route selection process. As such, the best route selected based upon the existing BGP route selection criteria may not be the best from the customer experience perspective.

This document describes a performance-based BGP routing paradigm in which network latency metric is disseminated via a new TLV of the AIGP attribute [RFC7311] and that metric 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 services to their customers.

Furthermore, in order to be backward compatible with existing BGP implementations and have no impact on the stability of the overall routing system, it's expected that the performance routing paradigm could coexist with the vanilla routing paradigm. As such, service providers could thus provide low-latency routing services while still offering the vanilla routing services depending on customers' requirements.

For the sake of simplicity, this document considers only one network performance metric that's 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 those BGP-irrelevant matters such as the mechanisms for measuring network latency are outside the scope of this document.

A variant of this performance-based BGP routing is implemented (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 was forwarded along a path which 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 across a wide geographical area, the cumulative link latency becomes the major part of the network latency since the total of the internal latency happened 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, such approximate network latency represented by the cumulative link latency is 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 Route Advertisement

Performance (i.e., low latency) 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](#)]. In other word, performance routes can then be looked as specific labeled routes which are associated with network latency metric.

A BGP speaker SHOULD NOT advertise performance 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 metric is attached to the performance 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 routes. If enabled, a local latency value for a given to-be-originated performance route MUST be configured to the BGP speaker so that it can be filled to the NETWORK_LATENCY TLV of that performance route.

A BGP speaker that is enabled to process NETWORK_LATENCY, but it was not provisioned with the local latency value SHOULD remove the NETWORK_LATENCY attribute when it advertises the corresponding route downstream.

When distributing a performance 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 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) from itself to the NEXT_HOP, rather than the shortest path calculated by Interior Gateway Protocol (IGP), the latency to the NEXT_HOP SHOULD reflect the network latency of that TE LSP path, rather than the IGP shortest path. In the case 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 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 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 Routing Capability MUST support the BGP Labeled Route Capability, as defined in [RFC3107]. A BGP speaker that advertises the Performance Routing Capability to a peer using BGP Capabilities advertisement [RFC5492] does not have to advertise the BGP Labeled Route Capability to that peer.

5. Performance Route Selection

Performance route selection only requires the following modification to the tie-breaking procedures of the BGP route selection decision (phase 2) described in [RFC4271]: 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 routing paradigm is independent from that of the vanilla routing paradigm. Accordingly, the routing table of the performance routing paradigm is independent from that of the vanilla routing paradigm. Whether the performance 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 Cos-Based Forwarding (CBF) capability which allows routers to have distinct routing and forwarding tables for each type of traffic, the selected performance routes could be installed in the routing and forwarding tables corresponding to high-priority traffic.

6. Deployment Considerations

This section is not normative.

Enabling the performance-based 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 scope of this document. Nevertheless, this document does not require nor exclude activating the proposed route selection scheme between domains that are 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-based 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 the low latency routes can be consistent throughout the domains. Besides security considerations that may arise (and which are further discussed in [Section 9](#)), there is indeed a need to consistently enforce a low-latency-based 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 which may have different QoS requirements. Moreover, the combined use of BGP-inferred low latency information with traffic engineering tools that would lead to the computation and the establishment of traffic-engineered LSP paths between "low latency"-enabled BGP peers based upon the manipulation of the Unidirectional Link delay sub-TLV [[RFC7810](#)] [[RFC7471](#)] would contribute to guarantee the overall consistency of the low latency information within each domain. Furthermore, a BGP color extended community could be attached to the performance routes so as to associates a low-latency Segment Routing (SR) LSP towards the BGP NEXT_HOP with these low-latency BGP routes, in this way, those traffic matching the low-latency BGP routes would be forwarded to the BGP NEXT_HOP via the low-latency SR LSP towards that BGP NEXT_HOP.

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.

From a network provider perspective, the ability to manipulate low latency routes may lead to different, presumably service-specific designs. In particular, there is a need to assess the impact of using such capability on the overall performance of the BGP peers from a route computation and selection procedure as a function of the tie-breaking operation. A typical use case would consist in selecting low latency routes for traffic that for example pertains to the VoIP, or whose nature demands the selection of the lowest latency route in the Adj-RIB-Out database of the corresponding BGP peers. Typically, live broadcasting services or some e-health services could certainly take advantage of such capability. It is out of scope of this document to exhaustively elaborate on such service-specific designs that are obviously deployment-specific.

[7.](#) 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

8. 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.

9. IANA Considerations

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

10. 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 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 routes received from the BGP peer for which the Performance 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.

11. References

11.1. 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>>.

11.2. Informative References

- [I-D.ietf-idr-bgp-optimal-route-reflection]
Raszuk, R., Cassar, C., Aman, E., Decraene, B., and K. Wang, "BGP Optimal Route Reflection (BGP-ORR)", [draft-ietf-idr-bgp-optimal-route-reflection-19](#) (work in progress), July 2019.
- [I-D.ietf-spring-segment-routing-policy]
Filsfils, C., Sivabalan, S., daniel.voyer@bell.ca, d., bogdanov@google.com, b., and P. Mattes, "Segment Routing Policy Architecture", [draft-ietf-spring-segment-routing-policy-03](#) (work in progress), May 2019.
- [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>>.

Authors' Addresses

Xiaohu Xu
Alibaba, Inc

Email: xiaohu.xxh@alibaba-inc.com

Shraddha Hegde
Juniper

Email: shraddha@juniper.net

Ketan Talaulikar
Cisco

Email: ketant@cisco.com

Mohamed Boucadair
France Telecom

Email: mohamed.boucadair@orange.com

Christian Jacquenet
France Telecom

Email: christian.jacquenet@orange.com

