Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 11, 2015

X. Xu
Huawei
M. Boucadair
C. Jacquenet
France Telecom
N. So
Vinci Systems
Y. Shen
Juniper
U. Chunduri
Ericsson
H. Ni
Huawei
Y. Fan
China Telecom

January 7, 2015

Performance-based BGP Routing Mechanism
draft-ietf-idr-performance-routing-00

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.

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 http://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 July 11, 2015.

Table of Contents

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).

1.1. 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].

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-exisit 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.

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. In the case where a router reflector is deployed without next-hop-self enabled when reflecting received routes from one IBGP peer to other IBGP peer, it is RECOMMENDED to enable such route reflector to

reflect all received performance routes by using some mechanisms such as [I-D.ietf-idr-add-paths], rather than reflecting only the performance route which is the best from its own perspective. Otherwise, it may result in a non-optimal choice by its clients and/or its IBGP peers.

The Loc-RIB of performance routing paradigm is independent from that of vanilla routing paradigm. Accordingly, the routing table of performance routing paradigm is independent from that of the vanilla routing paradigm. Whether performance routing paradigm or vanilla routing paradigm would be used for a given packet is a local policy issue which is outside the scope of this document.

6. Deployment Considerations

It is strongly RECOMMENDED to deploy this performance-based BGP routing mechanism across multiple ASes which belong to a single administrative domain. Within each AS, it is RECOMMENDED to deliver a packet from a BGP speaker to the BGP NEXT_HOP via tunnels, typically TE LSP tunnels. Furthermore, if a TE LSP is used between iBGP peers, it is RECOMMENDED to use the latency metric carried in Unidirectional Link Delay Sub-TLV [I-D.ietf-isis-te-metric-extensions] [I-D.ietf-ospf-te-metric-extensions] if possible, rather than the TE metric [RFC3630][RFC5305] to calculate the cumulative link latency associated with the TE LSP and use that cumulative link latency to approximately represent the network latency. Thus, there is no need for frequent measurement of network latency between IBGP peers.

7. Acknowledgements

Thanks to Joel Halpern, Alvaro Retana, Jim Uttaro, Robert Raszuk, Eric Rosen, Qing Zeng, Jie Dong, Mach Chen, Saikat Ray, Wes George, Jeff Haas, John Scudder, Stephane Litkowski and Sriganesh Kini for their valuable comments on the initial idea of 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.

8. 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.

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 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.

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, August 2014.

10.2. Informative References

[I-D.ietf-idr-add-paths]
Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-10 (work in progress), October 2014.

[I-D.ietf-isis-te-metric-extensions]
Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., Filsfils, C., and W. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-ietf-isis-te-metric-extensions-04 (work in progress), October 2014.

[I-D.ietf-ospf-te-metric-extensions]

Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-10 (work in progress), January 2015.

[RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

Authors' Addresses

Xiaohu Xu
Huawei

Email: xuxiaohu@huawei.com

Mohamed Boucadair
France Telecom

Email: mohamed.boucadair@orange.com

Christian Jacquenet
France Telecom

Email: christian.jacquenet@orange.com

Ning So
Vinci Systems

Email: ning.so@vinci-systems.com

Yimin Shen
Juniper

Email: yshen@juniper.net

Uma Chunduri
Ericsson

Email: uma.chunduri@ericsson.com

Hui Ni
Huawei

Email: nihui@huawei.com

Yongbing Fan
China Telecom

Email: fanyb@gsta.com