Network Working Group
Internet Draft
Category: Standard Track

X. Xu
H. Ni
Huawei

M. Boucadair
C. Jacquenet
France Telecom

N. So
Tata Communications

Y. Fan
China Telecom

Expires: July 2014 January 16, 2014

Performance-based BGP Routing Mechanism

draft-xu-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.

This Internet-Draft will expire on July 16, 2014.

Copyright Notice

Conventions used in this document

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].

Table of Contents

1. Introduction

Network performance, especially network latency is widely recognized as one of major obstacles in migrating business applications to the cloud, especially in the case where the network paths between cloud users and cloud data centers traverse more than one Autonomous System (AS), and would therefore stretch the forwarding path. However, the current Border Gateway Protocol (BGP) specification [RFC4271] which is used for path selection across ASes (Autonomous Systems) doesn't use network performance metrics (e.g., 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 mechanism in which network performance metrics are conveyed as additional path attributes of the Network Layer Reachability Information (NLRI) and used in the route selection decisions. So far it's only the network latency metric that would be used in the performance-based route selection decisions. 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.

For the sake of simplicity, this document considers only one performance metric that's the network latency metric. The support of multiple attributes is out of scope of this document.

To make the performance routing paradigm and the vanilla routing paradigm coexist, performance routes should be exchanged as labeled routes as per [RFC3107] while using a specified Subsequent Address Family Identifier (SAFI). As such, network providers deploying such mechanism in their networks may provide the performance routing service as a value-added service to those customers with low latency need, while continually offering the vanilla routing service to the remaining customers as before.

A variant of this performance-based BGP routing is implemented [URL: 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 routes SHOULD be exchanged between BGP peers by using a specified Subsequent Address Family Identifier (SAFI) of TBD (see IANA Section). Meanwhile, these routes SHOULD be carried as labeled routes as per [RFC3107].

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 the specified SAFI field.

Network latency metric is attached to the performance routes as one additional path attribute, referred to as NETWORK_LATENCY path attribute, which is a well-known mandatory attribute. This attribute 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 attribute 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.

A BGP speaker SHOULD be configurable to enable or disable the origination/creation 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 attribute of that performance route.

When distributing a selected performance route learnt from one BGP peer to another, unless this BGP speaker has set itself as the NEXT_HOP of such route, the NETWORK_LATENCY path attribute of such route MUST NOT be modified. Otherwise when setting itself as the NEXT_HOP of such route, this BGP speaker SHOULD increase the value of the NETWORK_LATENCY path attribute by adding the network latency value from itself to the previous NEXT_HOP of such route. It is RECOMMENDED to use the cumulative link latency from this BGP speaker to the NEXT_HOP to represent the network latency between them if possible. Otherwise, the measured network latency between them can be used instead. It is RECOMMENDED that the type of network latency SHOULD be kept consistent across all these AS's (i.e., either cumulative link latency or measured network latency, choose one).

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.

To keep performance routes stable enough, a BGP speaker SHOULD use a configurable threshold of network latency fluctuation to suppress any update which would otherwise be triggered just 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 path attribute 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 [ADD-PATH], 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 RECOMMENDED to deploy this performance-based BGP routing mechanism across multiple ASes which are within a single administrative domain. Within each AS, it is RECOMMENTED to deliver a packet from a BGP speaker to the BGP NEXT_HOP via tunnels, especially TE LSP tunnels. Furthermore, it is RECOMMENDED to use the latency metric carried in Unidirectional Link Delay Sub-TLV [OSPF-TE-EXT] [ISIS-TE-EXT] if possible, rather than the TE metric [RFC3630] [RFC5305] to perform the C-SPF calculation, unless the TE metric has already been set to the link latency metric. In this way, it could avoid the need for timely measurement of network latency between IBGP peers.

7. Security Considerations

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

Tweaking the value of the NETWORK_LATENCY by an illegitimate party may influence the route selection process. Means to check the integrity of BGP messages are RECOMMENDED.

Frequent updates of the NETWORK_LATENCY attribute may have a severe impact on the stability of the routing system. Such practice SHOULD be avoided.

8. IANA Considerations

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

9. Acknowledgements

Thanks to Joel Halpern, Alvaro Retana, Jim Uttaro, Robert Raszuk, Eric Rosen, Qing Zeng, Jie Dong and Mach Chen for their valuable comments on the initial idea of this document.

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.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

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

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

10.2. Informative References

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

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

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

[OSPF-TE-EXT] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-02 (work in progress), December 2012.

[ISIS-TE-EXT] Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., and C. Filsfils, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-previdi-isis-te-metric-extensions-02 (work in progress), October 2012.

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

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

[ADD-PATH] D. Walton, A. Retana, E. Chen, J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-09 (work in progress), October 2013.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
Beijing, China
Phone: +86-10-60610041
Email: xuxiaohu@huawei.com

Hui Ni
Huawei Technologies,
Beijing, China
Phone: +86-10-606100212
Email: nihui@huawei.com

Mohamed Boucadair
France Telecom
Rennes, France
EMail: mohamed.boucadair@orange.com

Christian Jacquenet

Orange
Rennes France
Email: christian.jacquenet@orange.com

Ning So
Tata Communications
Plano, TX 75082, USA
Email: ning.so@tatacommunications.com

Yongbing Fan
China Telecom
Guangzhou, China.
Phone: +86 20 38639121
Email: fanyb@gsta.com