

BESS WorkGroup
Internet-Draft
Intended status: Informational
Expires: March 14, 2018

S. Mohanty
A. Sreekantiah
D. Rao
Cisco Systems
K. Patel
Arrcus, Inc
September 10, 2017

BGP Multipath in Inter-AS Option-B
draft-mohanty-bess-mutipath-interas-01

Abstract

By default, The Border Gateway Protocol, BGP only installs the best-path to the IP Routing Table. BGP multi-path is a well known feature that enables installation of multiple paths to the IP Routing Table. This is done to achieve load balancing while forwarding traffic. For a path to be eligible as a multi-path, certain criteria need to be fulfilled. Inter-AS VPNs are commonly deployed to span organizations across Service Provider boundaries. In this draft, we describe an issue relating to multi-path load balancing that can arise in an Option B Inter-AS Deployment. With the help of a representative topology, we illustrate the problem and then present two simple schemes as the solution to the problem. We also note as a matter of independent interest that the same underlying issue is applicable to deployments that employ next-hop-self behavior (implicit or explicit) downstream and the multi-path feature upstream.

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 March 14, 2018.

Internet-Draft

BGP Multipath in Inter-AS Option-B

September 2017

Copyright Notice

Copyright (c) 2017 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.	Requirements Language	3
3.	Topology notation	3
4.	Problem Description	4
5.	BGP ADDpath with the non-unique RD case	4
6.	BGP Labeled unicast with Add-Path	5
7.	BGP Multi-path Inter-As Solution 1	5
8.	BGP Multi-path Inter-As Solution 2	5
9.	Protocol Considerations	6
10.	Operational Considerations	6
11.	Security Considerations	6
12.	Acknowledgements	6
13.	References	6
13.1.	Normative References	6
13.2.	Informative References	7
	Authors' Addresses	7

[1.](#) Introduction

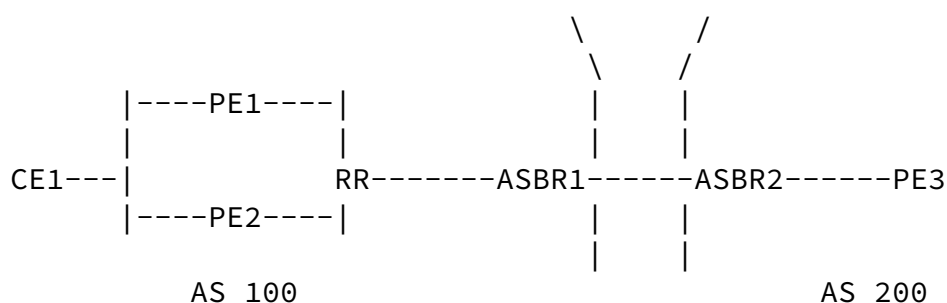
By Default BGP [[RFC4271](#)] only advertises the best-path to a peer and also installs the best-path to the IP Routing Table (RIB) and thereby to the Forwarding Information Base (FIB). BGP multi-path is a feature where more than one received BGP route, rather than only the one corresponding to the BGP best-path, are installed in the IP Routing Table and the Forwarding Information Base. This offers benefits of load balancing, efficient utilization of system resources

network-wide, and enabling high throughput for traffic flows which would be lacking otherwise. It also has the added benefit of providing redundancy in case one of the BGP paths are withdrawn due to a link going down or some other event. Often vendors have a

configurable knob which dictates how many paths to a given destination can be installed in the forwarding.

BGP Multi-path is widely deployed in practice and when augmented with the Demilitarized Link Bandwidth (DMZ LB) [[I-D.ietf-idr-link-bandwidth](#)] can be used to provide unequal cost load balancing as per user control.

The BGP best-path algorithm proceeds through a well-known and deterministic selection mechanism in determining the best-path. Typically, a path is deemed eligible as a multi-path, if it encounters a tie with the best-path, when it is determined that the IGP cost (metric) to the BGP next-hop is the same, as per the BGP best-path algorithm [[RFC4271](#)]. In addition, two paths, which match all criteria until the IGP metric but have the same next-hop IP address cannot both be considered as multi-paths. This is regardless of EBGp or IBGP rules. In this draft we point out an issue that limits the benefits of multi-path deployments arising out of above restrictions when the BGP path is propagated across Inter-AS Option B [[RFC4364](#)] Autonomous System Boundary Routers (ASBRs).



Inter-AS Option B.

Figure 1

2. 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 [[RFC2119](#)].

[3.](#) Topology notation

In the Figure 1. above, we consider a typical Inter-AS Option B topology, ASBR1 peering with ASBR2 over the inter-AS eBGPlink. A VPN, vpn has a presence in both the Autonomous Systems, on all the PE routers shown; i.e. a Virtual routing Forwarding (VRF) tables associated with the VPN vpn exists at each of the Provider Routers

shown. A dual-homed CE, CE1 is peering with PE1 and PE2 respectively in the context of vrf VRF1.

Denote the Route-Distinguisher (RD) of the vrf VRF1 configured in PE1 by RD1. Denote the Route-Distinguisher of the vrf VRF2 configured in PE2 by RD2. Assume that CE1 advertises an ipv4 prefix p, at ASBR1, the received VPN route prefix will be RD1:p and RD2:p, with next-hops PE1 and PE2 respectively, with the vpn (service) label as L1 and L2 respectively.

[4.](#) Problem Description

As per EBGp rules at the advertising ASBR, ASBR1, the next-hop will be reset to the ASBR1 itself. This causes the two routes RD1:p and RD2:p to be advertised to the receiving AS, AS2, with the mandatory attribute, the next-hop which points to ASBR1.

Let's say the swapped label for RD1:p and RD2:p at ASBR1 is L1 and L2 respectively. If ASBR2 does not reset the next-hop (usual behavior), then the two paths will be received at PE3 with the same next-hop, i.e. ASBR1. If ASBR2 does reset the next-hop, then the two paths will be received at PE3 with the next-hop set to ASBR2.

In either case above, the two paths received at PE3 have the same next-hop, even though the labels are different. As explained earlier, if two received BGP paths have the same next-hop, then both of them cannot be eligible for multi-paths at the same time. This means that at the PE3, only one of the routes will be installed in the forwarding.

In the Figure 1 above, even though the advertising AS (AS 100) has path redundancy, this is not visible to AS 200, and therefore load balancing cannot be done at ASBR1. Note that this is different from the classic same RD problem which one often encounters in the Route-Reflector context.

5. BGP ADDpath with the non-unique RD case

The above scenario is described in the context of the unique-RD case. Now consider the case when one has non-unique RDs configured for the vpn VRF at PE1 and PE2, and BGP Add-Path [[RFC7911](#)] is used to propagate the paths to AS200 via RR, ASBR1 and ASBR2 respectively. In this case, the ASBR1 resets the next-hop to itself in both of the add-paths thus ensuring that the two add-paths cannot be installed as primary and backup in the FIB at PE3 in AS200.

6. BGP Labeled unicast with Add-Path

A similar situation exists for non-VPN labeled traffic. Figure 2 shows a simple ebgp topology, in which R1 is in AS 1, R2 and R3 are in AS 2, R4 is in AS 3, and R5 is in AS 4. A labeled unicast [[RFC3107](#)] prefix, p, is being advertised from R1 to R5. Add-Path is configured at R4 and R5 and the capability is negotiated. Both R2 and R3, will set the next-hop to themselves. When R4 receives the prefix p from R2 and R3, the situation is similar to the add-path scenario for the VPN case as described in the earlier section. As a result only one of the paths will be advertised to R5.

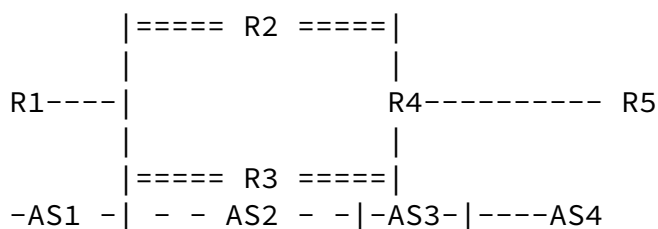


Figure 2

7. BGP Multi-path Inter-As Solution 1

The first solution is to consider the uniqueness of the label and the next-hop by considering the tuple (next-hop, label). This translates to (ASBR1, L1) and (ASBR2, L2) and therefore they can be distinguished. However many existing deployments today consider only the next-hop as the key. Therefore this solution requires upgrade to existing deployment software. An independent issue is that there should be no implications on hashing the weights assigned to the paths in the FIB due to the dependency on the label.

8. BGP Multi-path Inter-As Solution 2

The second solution is to inject two loopback ip addresses at ASBR1 into the IBGP of the receiving AS corresponding to the PE1 and PE2's configured ip address or loopbacks that are in the next-hop attribute of the vpn routes RD1:p and RD2:p. These loopback addresses need to be injected into the IGP of the receiving AS. Also ASBR2 needs to be configured with a static route pointing to ASBR1 for this purpose. Alternatively, ASBR1 can redistribute these loopbacks into EBGp. This is also equivalent to doing next-hop-self. The above solution won't require any software upgrade. However it will require the

implementation to support policy and may have security implications since routes need to be leaked from one AS to the other.

9. Protocol Considerations

No Protocol Changes are necessary

10. Operational Considerations

Any of the two methods above can be adopted. A note may be made that these solutions also are applicable to EVPN [[RFC7432](#)]

11. Security Considerations

This document raises no new security issues for L3VPN.

12. Acknowledgements

The authors would like to thank Yuri Tsier for his feedback and useful discussions

13. References

13.1. Normative References

- [I-D.ietf-idr-extcomm-iana]
Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", [draft-ietf-idr-extcomm-iana-02](#) (work in progress), December 2013.
- [I-D.ietf-idr-link-bandwidth]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", [draft-ietf-idr-link-bandwidth-06](#) (work in progress), January 2013.
- [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>>.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", [RFC 4360](#), DOI 10.17487/RFC4360, February 2006, <<https://www.rfc-editor.org/info/rfc4360>>.

- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", [RFC 7432](#), DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.

13.2. Informative References

- [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>>.

[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>>.

[RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", [RFC 6624](#), DOI 10.17487/RFC6624, May 2012, <<https://www.rfc-editor.org/info/rfc6624>>.

[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

Satya Ranjan Mohanty
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: satyamoh@cisco.com

Arjun Sreekantiah
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: asreekan@cisco.com

Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: dhrao@cisco.com

Keyur Patel
Arrcus, Inc

Email: keyur@arrcus.com