IDR Working Group
Internet-Draft
Intended status: Informational
Expires: May 6, 2020

P. Lapukhov
Facebook
J. Tantsura
Apstra, Inc.

November 3, 2019

Equal-Cost Multipath Considerations for BGP
draft-lapukhov-bgp-ecmp-considerations-03

Abstract

BGP (Border Gateway Protocol) [RFC4271] employs tie-breaking logic to select a single best path among multiple paths available, known as BGP best path selection. At the same time, it has become a common practice to allow for "equal-cost multipath" (ECMP) selection and programming of multiple next-hops in routing tables. This document summarizes some common considerations for the ECMP logic when BGP is used as the routing protocol, with the intent of providing common reference for otherwise unstandardized set of features.

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 May 6, 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. AS-PATH attribute comparison 2
3. Multipath among eBGP-learned paths 3
4. Multipath among iBGP learned paths 3
5. Multipath among eBGP and iBGP paths 4
6. Multipath with AIGP 5
7. Best path advertisement 5
8. Multipath and non-deterministic tie-breaking 5
9. Weighted equal-cost multipath 5
10. Acknowledgements 6
11. Informative References 6
Authors' Addresses 7

1. Introduction

Section 9.1.2.2 of [RFC4271] defines step-by-step tie-breaking procedure for selecting a single "best-path" among multiple alternatives available for the same route. In order to improve efficiency, in densely meshed symmetric network topologies is has become a common practice to allow selection of multiple "equal" paths for the same route. Most commonly used approach is to abort the tie-breaking process after comparing IGP cost for the NEXT_HOP attribute and selecting either all eBGP or all iBGP paths that remained "equal" under the tie-breaking rules (see [BGPMP] for a vendor document explaining the logic). Basically, the steps that compare the BGP identifiers and BGP peer IP addresses (steps (f) and (g) in [RFC4271]) are ignored for the purpose of multipath routing. BGP implementations commonly have a configuration knob that specifies the maximum number of equal paths that are allowed be programmed in the routing table. Commonnly, there's also a knob to enable multipath separately for iBGP-learned or eBGP-learned paths.

2. AS-PATH attribute comparison

The mandatory requirement for all paths that are considered as the candidates for ECMP selection is to have the same AS_PATH length, computed using the logic defined in [RFC4271] and [RFC5065], i.e. ignoring the AS_SET, AS_CONFED_SEQUENCE, and AS_CONFED_SET segment lengths. The content of the latter attributes is used purely for loop detection and prevention. Assuming that AS_PATHs length computed in this fashion are the same, many implementations require

that the content of AS_SEQUENCE segment MUST be the same among all the paths considered. Two common configuration knobs to alter this behaviour are usually provided: One, to relax otherwise mandatory AS_SEQUENCE comparison rule, enforcing only the AS_PATH length rule, while ignoring the content of AS_SEQUENCE. And another requiring that the first AS numbers in first AS_SEQUENCE segment found in AS_PATH (often referred to as "peer AS" number) be the same as the one found in best path (determined by running the full tie-breaking procedure). This document refers to those two as "multipath as-path relaxed" and "multipath same peer-as".

3. Multipath among eBGP-learned paths

Step (d) in Section 9.1.2.2 of [RFC4271] mandates, in presence of an eBGP path to remove all iBGP paths from the the ECMP candidates set. This leaves the BGP tie-breaking procedure with just eBGP paths. At this point, the mandatory BGP NEXT_HOP attribute value most commonly belongs to the IP subnet that the BGP speaker shares with the advertising neighbor. In this case, it is common for implementations to treat all NEXT_HOP values as having the same "internal cost" to reach them per the guidance of step (e) of Section 9.1.2.2. In some cases, either static routing or an IGP routing protocol could be running between the BGP speakers peering over eBGP session. An implementation may use the metric discovered from the above sources to perform tie-breaking even for eBGP paths.

Notice that in case when, in some paths MED attribute is present, the set of multipath routes allowed will most likely be reduced to the ones coming from the same peer AS, per step (c) of Section 9.1.2.2. This is unless an implementation provides a configuration knob to always compare MED attributes across all paths, as recommended by [RFC4451]. In the latter case, the presence of MED attribute does not automatically reduce the candidate path set to the same peer AS only.

4. Multipath among iBGP learned paths

When all paths for a prefix are learned via iBGP, since in most cases iBGP is used along with an underlying IGP, the tie-breaking commonly occurs based on IGP metric of the NEXT_HOP attribute. In some implementations, it is however possible to ignore the IGP cost as well, if all of the paths are reachable via some kind of tunneling mechanism, such as MPLS [RFC3031]. This is enabled via a knob referred in this document as "skip igp check" . Notice that there is no standard way for a BGP speaker to detect presence of such tunneling techniques other than relying on the configuration settings.

When iBGP is deployed with BGP route-reflectors per [RFC4456] the path attribute list may include the CLUSTER_LIST attribute. Most implementations commonly ignore it for the purpose of ECMP route selection, assuming that IGP cost along should be sufficient for loop prevention. This assumption may not hold when IGP is not deployed, and instead iBGP session are configured to reset the NEXT_HOP attribute to self on every node (this also assumes the use of directly connected link addresses for session formation). In this case, ignoring CLUSTER_LIST length might lead to routing loops. It is therefore recommended for implementations to have a knob that enables accounting for CLUSTER_LIST length when performing multipath route selection. In this case, CLUSTER_LIST attribute length should be effectively used to replace the IGP metric.

Similarly to the route-reflector scenario, the use of BGP confederations in multipath scenarios assumes presence of an IGP for proper loop prevention and use the IGP metric as the final tie-breaker for multipath routing. In addition to that, and similar to eBGP case, implementations often require that in order to be considered equal, paths under consideration must belong to the same peer member AS as the best-path. It is useful to have the following two configuration knobs, one enabling "multipath same confederation member peer-as" and another enabling less restrictive "confed as-path multipath relaxed" rule, that allow selecting multipath routes reachable via any confederation member peer AS. As mentioned above, the AS_CONFED_SEQUENCE value length is usually ignored for the purpose of AS_PATH length comparison, for the loop prevention relying instead on the IGP cost .

In cases, when IGP is not present with BGP confederation deployment, and similar to route-reflection case, it may be nessesary to consider AS_CONFED_SEQUENCE length when selecting the equivalent routes, effectively using it as a substitution for an IGP metric. A separate configuration knob is needed to allow this behavior.

Per [RFC5065] paths learned over BGP intra-confederation peering sessions are treated as iBGP. There is no specification or operational document that defines how a mixed iBGP route-reflector and confederation based deplyments would work together. Therefore, this document does not make recommendations for the above case.

5. Multipath among eBGP and iBGP paths

The best-path selection algorithm explicitly prefers eBGP paths over iBGP (or learned from BGP confederation member AS, which is, as per [RFC5065] treated the same as iBGP from perspective of best-path selection). In some cases however, it might be beneficial to allow multipath routing between eBGP and iBGP learned paths. This is only

possible if some sort of tunneling technique is used to reach both the eBGP and iBGP paths. If this feature is enabled, the equal routes are selected prior to the MED comparison step (c) in Section 9.1.2.2 [RFC4271].

6. Multipath with AIGP

AIGP attribute defined in [RFC7311] must be used for best-path selection prior to running any logic of Section 9.1.2.2 [RFC4271]. Only the paths with minimal value of AIGP metric are eligible for further consideration of tie-breaking rules. The rest of multipath selection logic remains the same.

7. Best path advertisement

Unless BGP "Add-Path" feature as described in [RFC7911] is enabled and even though multiple equal paths may be selected for programming into the routing table, a BGP speaker announces to its peers single best-path only. The unique best-path is elected among the multi-path set using the standard tie-breaking rules.

8. Multipath and non-deterministic tie-breaking

Some implementations may implement non-standard tie-breaking logic, for example using the oldest path rule, IETF reference - [RFC5004], a vendor implementaion example [BGPMP]. This is generally not recommended, and may interact with multi-path route selection on downstream BGP speakers. That is, after a route flap that affects the best-path upstream, the original best path would not be recovered, and the older path would still be advertised, possibly affecting the tie-breaking rules on down-stream device if for example, the AS_PATH contents are different from previous. Another side effect of using non-standard tie-breaking could be increased number of BGP Next-Hop sets for Prefixes learned from eBGP neighbors and advertised downstream towards iBGP Neighbors. This could potentially cause ECMP group/entry tables to overrun (depending on a platform) as the prefixes will be less coalesced.

9. Weighted equal-cost multipath

The proposal in [I-D.ietf-idr-link-bandwidth] defines conditions where iBGP multipath feature might inform the routing table of "weights" associated with the multiple external paths. [I-D.ietf-idr-link-bandwidth] defines the weight extended community attribute as non-transitive, considers the applicability for iBGP only, though there are implementations that apply it to eBGP as well. The proposal does not change the equal-cost multipath selection

logic, only associates additional load-sharing attributes with equivalent paths.

10. Acknowledgements

We like to thank Diptanshu Singh for their reviews and valuable comments.

11. Informative References

[BGPMP] "BGP Best Path Selection Algorithm", <http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.html>.

[I-D.ietf-idr-link-bandwidth] Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-07 (work in progress), March 2018.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[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>.

[RFC4451] McPherson, D. and V. Gill, "BGP MULTI_EXIT_DISC (MED) Considerations", RFC 4451, DOI 10.17487/RFC4451, March 2006, <https://www.rfc-editor.org/info/rfc4451>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC5004] Chen, E. and S. Sangli, "Avoid BGP Best Path Transitions from One External to Another", RFC 5004, DOI 10.17487/RFC5004, September 2007, <https://www.rfc-editor.org/info/rfc5004>.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, DOI 10.17487/RFC5065, August 2007, <https://www.rfc-editor.org/info/rfc5065>.

[RFC7311] Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro, "The Accumulated IGP Metric Attribute for BGP", RFC 7311, DOI 10.17487/RFC7311, August 2014, <https://www.rfc-editor.org/info/rfc7311>.

[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

Petr Lapukhov
Facebook
1 Hacker Way
Menlo Park, CA 94025
US

Email: petr@fb.com

Jeff Tantsura
Apstra, Inc.
333 Middlefield Rd #200
Menlo Park, CA 94025
US

Email: jefftant.ietf@gmail.com