Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 4, 2017

P. Lapukhov
Facebook

October 31, 2016

Equal-Cost Multipath Considerations for BGP
draft-lapukhov-bgp-ecmp-considerations-00

Abstract

BGP routing protocol defined in ([RFC4271]) employs tie-breaking logic to elect single best path among multiple possible. At the same time, it has been common in virtually all BGP implementations to allow for "equal-cost multipath" (ECMP) election and programming of multiple next-hops in routing tables. This documents summarizes some common considerations for the ECMP logic, with the intent of providing common reference on otherwise unstandardized feature.

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 May 4, 2017.

Copyright Notice

Table of Contents

1. Introduction

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

2. AS-PATH attribute comparison

A mandatory requirement is for all paths that are candidates for ECMP selection to have the same AS_PATH length, computed using the standard 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.

Assuming that AS_PATH lengths computed in this fashion are the same, many implementations require that content of AS_SEQUENCE segment MUST be the same among all equivalent paths. Two common configuration knobs are usually provided: one allowing only the length of AS_PATH to be the same, 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 algorithm). This document refer to those two as "multipath as-path relaxed" and "multipath same peer-as" knobs.

3. Multipath among eBGP-learned paths

Step (d) in Section 9.1.2.2 of [RFC4271] instructs to remove all iBGP paths from considerations if an eBGP path is present in the candidate set. This leaves the BGP process 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 advertising neighbor. In this case, it is common for implementation 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 MED attribute is present in some paths, the set of allowed multipath routes 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 the implementation provided a configuration knob to always compare MED attributes across all paths, as recommended in [RFC4451]. In the latter case, the presence of MED attribute does not automatically narrow the candidate path set only to the same peer AS.

4. Multipath among iBGP learned paths

When all paths for a prefix are learned via iBGP, the tie-breaking commonly occurs based on IGP metric of the NEXT_HOP attribute, since in most cases iBGP is used along with an underlying IGP. It is possible, in some implementations, 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 to as "skip igp check" in this document. Notice that there is no standard way for a BGP speaker to detect presence of such tunneling techniques other than relying on 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.

Similar to the route-reflector scenario, the use of BGP confederations assumes presence of an IGP for proper loop prevention in multipath scenarios, and use the IGP metric as the final tie-breaker for multipath routing. In addition to this, and similar to eBGP case, implementation often require that equivalent paths belong to the same peer member AS as the best-path. It is useful to have two configuration knobs, one enabling "multipath same confederation member peer-as" and another enabling less restrictive "confed as-path multipath relaxed", which allows selecting multipath routes going 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, relying on IGP cost instead for loop prevention.

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

Per [RFC5065] the path 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 model would work together. Therefore, this document does not make recommendations or considers this 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 per [RFC5065] is treated the same as iBGP from perspective of best-path selection). In some case, allowing multipath routing between eBGP and iBGP learned paths might be beneficial. This is only possible if some sort of tunneling technique is used to reach both the eBGP and iBGP path. If this feature is enabled, the equivalent routes are selection by stopping the tie-breaking process prior at the MED comparison step (c) in Section 9.1.2.2 of [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. 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

Event though multiple equivalent paths may be selected for programming into the routing table, the BGP speaker always announces single best-path to its peers, unless BGP "Add-Path" feature has been enabled as described in [I-D.ietf-idr-add-paths]. 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 using the oldest path rule. 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 still be advertised, possibly affecting the tie-breaking rules on down-stream device, for example if the AS_PATH contents are different from previous.

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 the "weights" associated with the multiple paths. The document defines the applicability only in iBGP case, though there are implementations that apply it to eBGP multipath as well. The proposal does not change the equal-cost multipath selection logic, only associates additional load-sharing attributes with equivalent paths.

10. Informative References

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <http://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, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc4456>.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous

System Confederations for BGP", RFC 5065,
DOI 10.17487/RFC5065, August 2007,
<http://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,
<http://www.rfc-editor.org/info/rfc7311>.

[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-15 (work in progress), May 2016.

[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.

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

Author's Address

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

Email: petr@fb.com