Internet Engineering Task Force (IETF)
Request for Comments: 7311
Category: Standards Track
ISSN: 2070-1721

P. Mohapatra
Sproute Networks
R. Fernando
E. Rosen
Cisco Systems, Inc.
J. Uttaro
AT&T

August 2014

The Accumulated IGP Metric Attribute for BGP

Abstract

Routing protocols that have been designed to run within a single administrative domain (IGPs) generally do so by assigning a metric to each link and then choosing, as the installed path between two nodes, the path for which the total distance (sum of the metric of each link along the path) is minimized. BGP, designed to provide routing over a large number of independent administrative domains (autonomous systems), does not make its path-selection decisions through the use of a metric. It is generally recognized that any attempt to do so would incur significant scalability problems as well as inter-administration coordination problems. However, there are deployments in which a single administration runs several contiguous BGP networks. In such cases, it can be desirable, within that single administrative domain, for BGP to select paths based on a metric, just as an IGP would do. The purpose of this document is to provide a specification for doing so.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7311.

Table of Contents

1. Introduction

There are many routing protocols that have been designed to run within a single administrative domain. These are known collectively as "Interior Gateway Protocols" (IGPs). Typically, each link is assigned a particular "metric" value. The path between two nodes can then be assigned a "distance", which is the sum of the metrics of all the links that belong to that path. An IGP selects the "shortest" (minimal distance) path between any two nodes, perhaps subject to the constraint that if the IGP provides multiple "areas", it may prefer the shortest path within an area to a path that traverses more than one area. Typically, the administration of the network has some routing policy that can be approximated by selecting shortest paths in this way.

BGP, as distinguished from the IGPs, was designed to run over an arbitrarily large number of administrative domains ("autonomous systems" or "ASes") with limited coordination among the various administrations. BGP does not make its path-selection decisions based on a metric; there is no such thing as an "inter-AS metric". There are two fundamental reasons for this:

- The distance between two nodes in a common administrative domain may change at any time due to events occurring in that domain. These changes are not propagated around the Internet unless they actually cause the border routers of the domain to select routes with different BGP attributes for some set of address prefixes. This accords with a fundamental principle of scaling, viz., that changes with only local significance must not have global effects. If local changes in distance were always propagated around the Internet, this principle would be violated.

- A basic principle of inter-domain routing is that the different administrative domains may have their own policies, which do not have to be revealed to other domains and which certainly do not have to be agreed to by other domains. Yet, the use of an inter-AS metric in the Internet would have exactly these effects.

There are, however, deployments in which a single administration runs a network that has been sub-divided into multiple, contiguous ASes, each running BGP. There are several reasons why a single administrative domain may be broken into several ASes (which, in this case, are not really autonomous.) It may be that the existing IGPs do not scale well in the particular environment; it may be that a more generalized topology is desired than could be obtained by use of a single IGP domain; it may be that a more finely grained routing policy is desired than can be supported by an IGP. In such deployments, it can be useful to allow BGP to make its routing

decisions based on the IGP metric, so that BGP chooses the shortest path between two nodes, even if the nodes are in two different ASes within that same administrative domain.

There are, in fact, some implementations that already do something like this, using BGP's MULTI_EXIT_DISC (MED) attribute to carry a value based on IGP metrics. However, that doesn't really provide IGP-like shortest path routing, as the BGP decision process gives priority to other factors, such as the AS_PATH length. Also, the standard procedures for use of the MED do not ensure that the IGP metric is properly accumulated so that it covers all the links along the path.

In this document, we define a new optional, non-transitive BGP attribute, called the "Accumulated IGP Metric Attribute", or "AIGP attribute", and specify the procedures for using it.

The specified procedures prevent the AIGP attribute from "leaking out" past an administrative domain boundary into the Internet. We will refer to the set of ASes in a common administrative domain as an "AIGP administrative domain".

The specified procedures also ensure that the value in the AIGP attribute has been accumulated all along the path from the destination, i.e., that the AIGP attribute does not appear when there are "gaps" along the path where the IGP metric is unknown.

2. Specification of Requirements

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. AIGP Attribute

The AIGP attribute is an optional, non-transitive BGP path attribute. The attribute type code for the AIGP attribute is 26.

The value field of the AIGP attribute is defined here to be a set of elements encoded as "Type/Length/Value" (i.e., a set of TLVs). Each such TLV is encoded as shown in Figure 1.

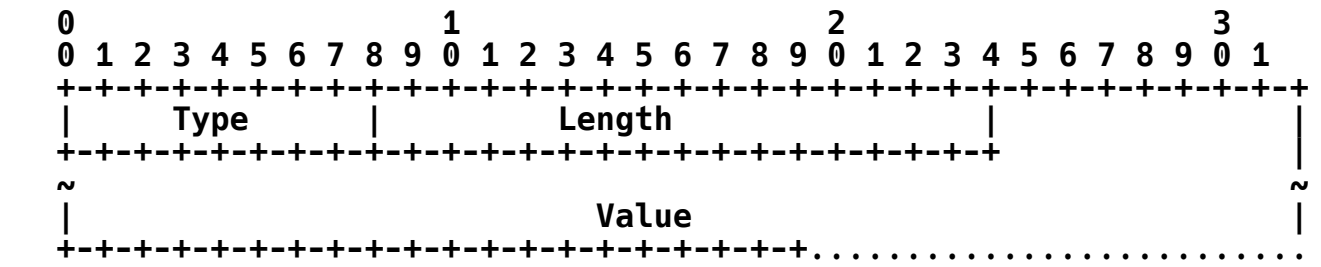

Figure 1: AIGP TLV

- Type: A single octet encoding the TLV Type. Only type 1, "AIGP TLV", is defined in this document. Use of other TLV types is outside the scope of this document.

- Length: Two octets encoding the length in octets of the TLV, including the Type and Length fields. The length is encoded as an unsigned binary integer. (Note that the minimum length is 3, indicating that no Value field is present.)

- Value: A field containing zero or more octets.

This document defines only a single such TLV, the "AIGP TLV". The AIGP TLV is encoded as follows:

- Type: 1

- Length: 11

- Value: Accumulated IGP Metric.

  The value field of the AIGP TLV is always 8 octets long, and its value is interpreted as an unsigned 64-bit integer. IGP metrics are frequently expressed as 4-octet values. By using an 8-octet field, we ensure that the AIGP attribute can be used to hold the sum of an arbitrary number of 4-octet values.

When an AIGP attribute is created, it SHOULD contain no more than one AIGP TLV. However, if it contains more than one AIGP TLV, only the first one is used as described in Sections 3.4 and 4. In the remainder of this document, we will use the term "value of the AIGP TLV" to mean the value of the first AIGP TLV in the AIGP attribute. Any other AIGP TLVs in the AIGP attribute MUST be passed along unchanged if the AIGP attribute is passed along.

3.1. Applicability Restrictions and Cautions

This document only considers the use of the AIGP attribute in networks where each router uses tunneling of some sort to deliver a packet to its BGP next hop. Use of the AIGP attribute in other scenarios is outside the scope of this document.

If a Route Reflector supports the AIGP attribute but some of its clients do not, then the routing choices that result may not all reflect the intended routing policy.

3.2. Handling a Malformed AIGP Attribute

When receiving a BGP Update message containing a malformed AIGP attribute, the attribute MUST be treated exactly as if it were an unrecognized non-transitive attribute. That is, it "MUST be quietly ignored and not passed along to other BGP peers" (see [BGP], Section 5). This is equivalent to the "attribute discard" action specified in [BGP-ERROR].

Note that an AIGP attribute MUST NOT be considered to be malformed because it contains more than one TLV of a given type or because it contains TLVs of unknown types.

If a BGP path attribute is received that has the AIGP attribute codepoint but also has the transitive bit set, the attribute MUST be considered to be a malformed AIGP attribute and MUST be discarded as specified in this section.

If an AIGP attribute is received and its first AIGP TLV contains the maximum value 0xffffffffffffffff, the attribute SHOULD be considered to be malformed and SHOULD be discarded as specified in this section. (Since the TLV value cannot be increased any further, it is not useful for metric-based path selection.)

3.3. Restrictions on Sending/Receiving

An implementation that supports the AIGP attribute MUST support a per-session configuration item, AIGP_SESSION, that indicates whether the attribute is enabled or disabled for use on that session.

- For Internal BGP (IBGP) sessions, and for External BGP (EBGP) sessions between members of the same BGP Confederation [BGP-CONFED], the default value of AIGP_SESSION SHOULD be "enabled".

- For all other External BGP (EBGP) sessions, the default value of AIGP_SESSION MUST be "disabled".

The AIGP attribute MUST NOT be sent on any BGP session for which AIGP_SESSION is disabled.

If an AIGP attribute is received on a BGP session for which AIGP_SESSION is disabled, the attribute MUST be treated exactly as if it were an unrecognized non-transitive attribute. That is, it "MUST be quietly ignored and not passed along to other BGP peers" (see [BGP], Section 5). However, the fact that the attribute was received SHOULD be logged (in a rate-limited manner).

3.4. Creating and Modifying the AIGP Attribute

3.4.1. Originating the AIGP Attribute

An implementation that supports the AIGP attribute MUST support a configuration item, AIGP_ORIGINATE, that enables or disables its creation and attachment to routes. The default value of AIGP_ORIGINATE MUST be "disabled".

A BGP speaker MUST NOT add the AIGP attribute to any route whose path leads outside the AIGP administrative domain to which the BGP speaker belongs. When the AIGP attribute is used, changes in IGP routing will directly impact BGP routing. Attaching the AIGP attribute to customer routes, Internet routes, or other routes whose paths lead outside the infrastructure of a particular AIGP administrative domain could result in BGP scaling and/or thrashing problems.

The AIGP attribute may be added only to routes that satisfy one of the following conditions:

- The route is a static route, not leading outside the AIGP administrative domain, that is being redistributed into BGP;
- The route is an IGP route that is being redistributed into BGP;
- The route is an IBGP-learned route whose AS_PATH attribute is empty; or
- The route is an EBGP-learned route whose AS_PATH contains only ASes that are in the same AIGP administrative domain as the BGP speaker.

A BGP speaker R MUST NOT add the AIGP attribute to any route for which R does not set itself as the next hop.

It SHOULD be possible to set AIGP_ORIGINATE to "enabled for the routes of a particular IGP that are redistributed into BGP" (where "a particular IGP" might be OSPF or IS-IS). Other policies determining when and whether to originate an AIGP attribute are also possible, depending on the needs of a particular deployment scenario.

When originating an AIGP attribute for a BGP route to address prefix P, the value of the AIGP TLV is set according to policy. There are a number of useful policies, some of which are in the following list:

- When a BGP speaker R is redistributing into BGP an IGP route to address prefix P, the IGP will have computed a distance from R to P. This distance MAY be assigned as the value of the AIGP TLV.

- A BGP speaker R may be redistributing into BGP a static route to address prefix P, for which a distance from R to P has been configured. This distance MAY be assigned as the value of the AIGP TLV.

- A BGP speaker R may have received and installed a BGP-learned route to prefix P, with next hop N. Or it may be redistributing a static route to P, with next hop N. Then:

  * If R has an IGP route to N, the IGP-computed distance from R to N MAY be used as the value of the AIGP TLV of the route to P.

  * If R has a BGP route to N, and an AIGP TLV attribute value has been computed for that route (see Section 3.4.3), that value MAY be used as the AIGP TLV value of the route to P.

3.4.2. Modifications by the Originator

If BGP speaker R is the originator of the AIGP attribute of prefix P, and the distance from R to P changes at some point, R SHOULD issue a new BGP update containing the new value of the AIGP TLV of the AIGP attribute. (Here we use the term "distance" to refer to whatever value the originator assigns to the AIGP TLV, however it is computed; see Section 3.4.1.) However, if the difference between the new distance and the distance advertised in the AIGP TLV is less than a configurable threshold, the update MAY be suppressed.

3.4.3. Modifications by a Non-Originator

Suppose a BGP speaker R1 receives a route with an AIGP attribute whose value is A and with a next hop whose value is R2. Suppose also that R1 is about to redistribute that route on a BGP session that is enabled for sending/receiving the attribute.

If R1 does not change the next hop of the route, then R1 MUST NOT change the AIGP attribute value of the route.

In all the computations discussed in this section, the AIGP value MUST be capped at its maximum unsigned value 0xffffffffffffffff. Increasing the AIGP value MUST NOT cause the value to wrap around.

Suppose R1 changes the next hop of the route from R2 to R1. If R1's route to R2 is either (a) an IGP-learned route or (b) a static route that does not require recursive next hop resolution, then R1 MUST increase the value of the AIGP TLV by adding to A the distance from R1 to R2. This distance is either the IGP-computed distance from R1 to R2 or some value determined by policy. However, A MUST be increased by a non-zero amount.

It is possible that R1 and R2 above are EBGP neighbors and that there is a direct link between them on which no IGP is running. Then, when R1 changes the next hop of a route from R2 to R1, the AIGP TLV value MUST be increased by a non-zero amount. The amount of the increase SHOULD be such that it is properly comparable to the IGP metrics. For example, if the IGP metric is a function of latency, then the amount of the increase should be a function of the latency from R1 to R2.

Suppose R1 changes the next hop of the route from R2 to R1 and R1's route to R2 is either (a) a BGP-learned route or (b) a static route that requires recursive next-hop resolution. Then, the AIGP TLV value needs to be increased in several steps, according to the following procedure. (Note that this procedure is ONLY used when recursive next-hop resolution is needed.)

1. Let Xattr be the new AIGP TLV value.

2. Initialize Xattr to A.

3. Set XNH to R2.

4. Find the route to XNH.

5. If the route to XNH does not require recursive next-hop resolution, get the distance D from R1 to XNH. (Note that this condition cannot be satisfied the first time through this procedure.) If D is above a configurable threshold, set the AIGP TLV value to Xattr+D. If D is below a configurable threshold, set the AIGP TLV value to Xattr. In either case, exit this procedure.

6. If the route to XNH is a BGP-learned route that does NOT have an AIGP attribute, then exit this procedure and do not pass on any AIGP attribute. If the route has an AIGP attribute without an AIGP TLV, then the AIGP attribute MAY be passed along unchanged.

7. If the route to XNH is a BGP-learned route that has an AIGP TLV value of Y, then set Xattr to Xattr+Y and set XNH to the next hop of this route. (The intention here is that Y is the AIGP TLV value of the route as it was received by R1, without having been modified by R1.)

8. Go to step 4.

The AIGP TLV value of a given route depends on (a) the AIGP TLV values of all the next hops that are recursively resolved during this procedure, and (b) the IGP distance to any next hop that is not recursively resolved. Any change due to (a) in any of these values MUST trigger a new AIGP computation for that route. Whether a change due to (b) triggers a new AIGP computation depends upon whether the change in IGP distance exceeds a configurable threshold.

If the AIGP attribute is carried across several ASes, each with its own IGP domain, it is clear that these procedures are unlikely to give a sensible result if the IGPs are different (e.g., some OSPF and some IS-IS) or if the meaning of the metrics is different in the different IGPs (e.g., if the metric represents bandwidth in some IGP domains but represents latency in others). These procedures also are unlikely to give a sensible result if the metric assigned to inter-AS BGP links (on which no IGP is running) or to static routes is not comparable to the IGP metrics. All such cases are outside the scope of the current document.

4. Decision Process

Support for the AIGP attribute involves several modifications to the tie-breaking procedures of the BGP "phase 2" decision described in [BGP], Section 9.1.2.2. These modifications are described in Sections 4.1 and 4.2.

In some cases, the BGP decision process may install a route without executing any tie-breaking procedures. This may happen, e.g., if only one route to a given prefix has the highest degree of preference (as defined in [BGP], Section 9.1.1). In this case, the AIGP attribute is not considered.

In other cases, some routes may be eliminated before the tie-breaking procedures are invoked, e.g., routes with AS-PATH attributes indicating a loop or routes with unresolvable next hops. In these cases, the AIGP attributes of the eliminated routes are not considered.

4.1. When a Route Has an AIGP Attribute

Assuming that the BGP decision process invokes the tie-breaking procedures, the procedures in this section MUST be executed BEFORE any of the tie-breaking procedures described in [BGP], Section 9.1.2.2 are executed.

If any routes have an AIGP attribute containing an AIGP TLV, remove from consideration all routes that do not have an AIGP attribute containing an AIGP TLV.

If router R is considering route T, where T has an AIGP attribute with an AIGP TLV,

- then R must compute the value A, defined as follows: set A to the sum of (a) T's AIGP TLV value and (b) the IGP distance from R to T's next hop.

- remove from consideration all routes that are not tied for the lowest value of A.

4.2. When the Route to the Next Hop Has an AIGP Attribute

Suppose that a given router R1 is comparing two BGP-learned routes, such that either:

- the two routes have equal AIGP TLV values, or else

- neither of the two routes has an AIGP attribute containing an AIGP TLV.

The BGP decision process as specified in [BGP] makes use, in its tie-breaking procedures, of "interior cost", defined as follows:

> interior cost of a route is determined by calculating the metric to the NEXT_HOP for the route using the Routing Table.

This document replaces the "interior cost" tie breaker of [BGP] with a tie breaker based on the "AIGP-enhanced interior cost". Suppose route T has a next hop of N. The "AIGP-enhanced interior cost" from node R1 to node N is defined as follows:

- Let R2 be the BGP next hop of the route to N after all recursive resolution of the next hop is done. Let m be the IGP distance (or in the case of a static route, the configured distance) from R1 to R2.

- If the installed route to N has an AIGP attribute with an AIGP TLV, set A to its AIGP TLV value, computed according to the procedure in Section 3.4.3.

- If the installed route to N does not have an AIGP attribute with an AIGP TLV, set A to 0.

- The "AIGP-enhanced interior cost" of route T is the quantity A+m.

The "interior cost" tie breaker of [BGP] is then applied, using the "AIGP-enhanced interior cost" instead of the "interior cost" as defined in [BGP].

5. Deployment Considerations

- Using the AIGP attribute to achieve a desired routing policy will be more effective if each BGP speaker can use it to choose from among multiple routes. Thus, it is highly recommended that the procedures of [BESTEXT] and [ADD-PATH] be used in conjunction with the AIGP attribute.

- If a Route Reflector does not pass all paths to its clients, then it will tend to pass the paths for which the IGP distance from the Route Reflector itself to the next hop is smallest. This may result in a non-optimal choice by the clients.

- When the procedures of this document are deployed, it must be understood that frequent changes of the IGP distance towards a certain prefix may result in equally frequent transmission of BGP updates about that prefix.

- In an IGP deployment, there are certain situations in which a network link may be temporarily assigned a metric whose value is the maximum metric value (or close to the maximum) for that IGP. This is known as "costing out" the link. A link may be "costed out" to deflect traffic from the link before the link is actually brought down or to discourage traffic from using a link until all the necessary state for that link has been set up (e.g., [LDP-IGP-SYNC]). This assumes, of course, that a path containing a "costed out" link will have a total distance that is larger than any alternate path within the same IGP area; in that case, the normal IGP decision process will choose the path that does not contain the "costed out" link.

Costing out a link will have the same effect on BGP routes that carry the AIGP attribute. The value of the AIGP TLV will be larger for a route (to a given prefix) that contains a "costed out" link than for a route (to the same prefix) that does not. It must be understood, though, that a route that carries an AIGP attribute will be preferred to a route that does not, no matter what the value of the AIGP TLV is. This is similar to the behavior in, e.g., an OSPF area, where an intra-area route is preferred to an inter-area or external route, even if the intra-area route's distance is large.

6. IANA Considerations

IANA has assigned the codepoint 26 in the "BGP Path Attributes" registry to the AIGP attribute.

IANA has created a registry for "BGP AIGP Attribute Types". The Type field consists of a single octet, with possible values from 1 to 255. (The value 0 is "Reserved".) The registration procedure for this registry is "Standards Action". Type 1 is defined as "AIGP" and refers to this document.

7. Security Considerations

The spurious introduction, through error or malfeasance, of an AIGP attribute could result in the selection of paths other than those desired.

Improper configuration on both ends of an EBGP connection could result in an AIGP attribute being passed from one service provider to another. This would likely result in an unsound selection of paths.

8. Acknowledgments

The authors would like to thank Waqas Alam, Rajiv Asati, Alia Atlas, Ron Bonica, Bruno Decraene, Brian Dickson, Clarence Filsfils, Sue Hares, Anoop Kapoor, Pratima Kini, Thomas Mangin, Robert Raszuk, Yakov Rekhter, Eric Rosenberg, Samir Saad, John Scudder, Shyam Sethuram, and Ilya Varlashkin for their input.

9. References

9.1. Normative Reference

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

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

9.2. Informative References

[ADD-PATH] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", Work in Progress, October 2013.

[BESTEXT] Marques, P., Fernando, R., Mohapatra, P., and H. Gredler, "Advertisement of the best external route in BGP", Work in Progress, January 2012.

[BGP-CONFED] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007.

[BGP-ERROR] Chen, E., Scudder, J., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", Work in Progress, June 2014.

[LDP-IGP-SYNC] Jork, M., Atlas, A., and L. Fang, "LDP IGP Synchronization", RFC 5443, March 2009.

Authors' Addresses

Pradosh Mohapatra
Sproute Networks

EMail: mpradosh@yahoo.com

Rex Fernando
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA 95134
US

EMail: rex@cisco.com

Eric C. Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA, 01719
US

EMail: erosen@cisco.com

James Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748
US

EMail: uttaro@att.com