Network Working Group   Pradosh Mohapatra
Internet Draft   Rex Fernando
Intended Status: Proposed Standard   Eric C. Rosen
Expires: June 12, 2012   Cisco Systems, Inc.

                                                  James Uttaro
                                                          ATT

                                            December 12, 2011

                 The Accumulated IGP Metric Attribute for BGP

                     draft-ietf-idr-aigp-07.txt

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.

material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html.

Copyright and License Notice

Table of Contents

1. 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].

2. 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 which 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 inter-AS
    metric in the Internet would have exactly these effects.

There are, however, deployments in which a single administration runs
a network which 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. We will refer to the set of
ASes in a common administrative domain as an "AIGP 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
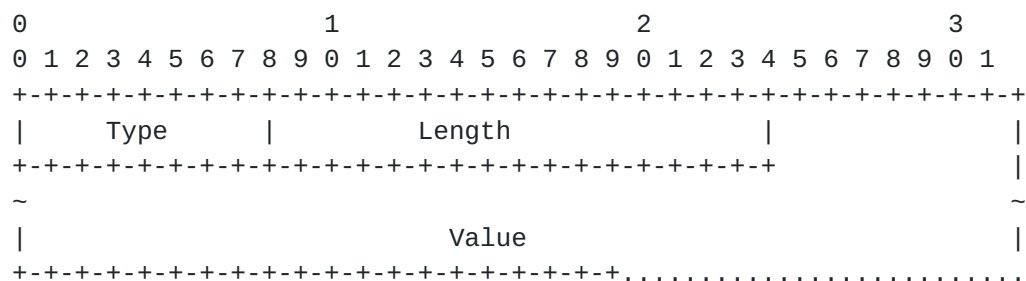   attribute", and specify the procedures for using it.

   The specified procedures prevent the AIGP attribute from "leaking
   out" past an AIGP administrative domain boundary into the Internet.

   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.

[3](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.


```
        0                   1                   2                   3
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |     Type      |             Length            |             |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+             |
       ~                                                             ~
       |                          Value                              |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+.........................
```


                              AIGP TLV
                              Figure 1

    - Type: A single octet encoding the TLV Type. Only type 1, "AIGP
      TLV", is defined in 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.)

- A value 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

- Accumulated IGP Metric.

  The value field of the AIGP TLV is always 8 bytes long. IGP
  metrics are frequently expressed as 4-octet values, and this
  ensures that the AIGP attribute can be used to hold the sum of an
  arbitrary number of 4-octet values.

[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 networks
that do not use tunneling 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]. 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.

  - The default value of AIGP_SESSION, for EBGP sessions, MUST be
    "disabled".

  - The default value of AIGP_SESSION, for IBGP and confederation-
    EBGP sessions, MUST be "enabled."

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).

3.3. Creating and Modifying the AIGP Attribute

3.3.1. 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. It may be added only to routes that satisfy one of
   the following conditions:

     - The route is a static route 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.

     - 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 MUST NOT add the AIGP attribute to any route for which
   it has 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 "ISIS"). 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 attribute is set according to policy. There are
   a number of useful policies, some of which are in the following list:

     - When a BGP speaker 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 AIGP
       attribute.

   - A BGP speaker 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 AIGP
     attribute.

   - 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. The "distance" from R to N
     MAY be assigned as the value of the AIGP attribute of the route
     to P.

       * If R has an IGP route to N, the IGP-computed distance from R
         to N MAY be used.

       * If R has a BGP route to N, and an AIGP attribute value has
         been computed for that route (see section 3.3.3), that value
         MAY be used as the AIGP attribute value of the route to P.

3.3.2. Modifications by the Originator

   If BGP speaker R is the originator of the AIGP attribute of prefix P,
   and at some point the "distance" from R to P changes, R SHOULD issue
   a new BGP update containing the new value of the AIGP attribute.
   (Here we use the term "distance" to refer to whatever value the
   originator assigns to the AIGP attribute, however it is computed; see
   section 3.3.1.) However, if the difference between the new distance
   and the distance advertised in the AIGP attribute is less than a
   configurable threshold, the update MAY be suppressed.

3.3.3. Modifications by a Non-Originator

   Suppose a BGP speaker R1 receives a route with an AIGP attribute
   whose value is A, and 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.

   If R1 changes the Next Hop of the route from R2 to R1, and if R1's
   route to R2 is an IGP-learned route, or a static route that does not
   require recursive next hop resolution, then R1 must increase the
   value of the AIGP attribute 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.

Note that if R1 and R2 above are EBGP neighbors, and 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 metric 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.
E.g., 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.

If R1 changes the Next Hop of the route from R2 to R1, and if R1's
route to R2 is a BGP-learned route, or a static route that requires
recursive next hop resolution, then the AIGP attribute 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 attribute value.

    2. Initialize Xattr to A.

    3. Set the 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 attribute value to Xattr+D. If D is below a configurable
       threshold, set the AIGP attribute value to Xattr. In either
       case, exit this procedure.

    6. If the route to XNH is a BGP-learned route, and the route does
       NOT have an AIGP attribute, then exit this procedure and do not
       pass on any AIGP attribute.

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

    8. Go to step 4.

The AIGP value of a given route depends on (a) the AIGP 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 below 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, remove from consideration all
routes that do not have an AIGP attribute.

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

   - then R must compute the value A, defined as follows: set A to the
     sum of (a) T's AIGP attribute 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 attribute values, or else

   - neither of the two routes has an AIGP attribute. The BGP
     decision process as specified in [BGP] makes use, in its tie
     breaker 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."

   Suppose route T has a next hop of N. We modify the notion of the
   "interior cost" from node R1 to node N as follows:

   - Let R2 be the 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, set A to the
     AIGP value of the route to N, computing the AIGP value of the
     route according to the procedure of section 3.3.3.

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

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

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 is it highly recommended that the procedures of
   [BESTEXT] and [ADDPATH] 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.

6. IANA Considerations

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

   IANA shall create a registry for "BGP AIGP Attribute Types". The
   type field consists of a single octet, with possible values from 0 to
   255. The allocation policy for this field is to be "Standards Action
   with Early Allocation". Type 1 should be defined as "AIGP", and
   should refer to this document.

7. Security Considerations

   The spurious introduction, though 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, Clarence
   Filsfils, Robert Raszuk, Yakov Rekhter, Samir Saad, John Scudder, and
   Shyam Sethuram for their input.

9. Authors' Addresses

    Rex Fernando
    Cisco Systems, Inc.
    170 Tasman Drive
    San Jose, CA 95134
    Email: rex@cisco.com

    Pradosh Mohapatra
    Cisco Systems, Inc.
    170 Tasman Drive
    San Jose, CA 95134
    Email: pmohapat@cisco.com

    Eric C. Rosen
    Cisco Systems, Inc.
    1414 Massachusetts Avenue
    Boxborough, MA, 01719
    Email: erosen@cisco.com

    James Uttaro
    AT&T
    200 S. Laurel Avenue
    Middletown, NJ 07748
    Email: uttaro@att.com

10. Normative References

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

11. Informative References

[ADDPATH] "Fast Connectivity Restoration Using BGP Add-Path", P.
Mohapatra, R. Fernando, C. Filsfils, R. Raszuk, draft-pmohapat-idr-
fast-conn-restore-02.txt, October 2011.

[BESTEXT], "Advertisement of the Best External Route in BGP", P.
Marques, R. Fernando, E. Chen, P. Mohapatra, H. Gredler, draft-ietf-
idr-best-external-04.txt, April 2011.

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