INTERNET-DRAFT   Danny McPherson
                                          Arbor Networks, Inc.
                                               Vijay Gill
                                                     AOL
Category   Informational
Expires: December 2005   June 2005

                        BGP MED Considerations
              <**draft-ietf-grow-bgp-med-considerations-04.txt**>

Status of this Memo

   By submitting this Internet-Draft, each author represents that any
   applicable patent or other IPR claims of which he or she is aware have
   been or will be disclosed, and any of which he or she becomes aware
   will be disclosed, in accordance with Section 6 of BCP 79.

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF), its areas, and its working groups. Note that other
   groups may also distribute working documents as Internet-Drafts.

   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".

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

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

Copyright Notice

                              Abstract

   The BGP MED attribute provides a mechanism for BGP speakers to convey
   to an adjacent AS the optimal entry point into the local AS. While
   BGP MEDs function correctly in many scenarios, there are a number of
   issues which may arise when utilizing MEDs in dynamic or complex
   topologies.

   This document discusses implementation and deployment considerations
   regarding BGP MEDs and provides information which implementors and
   network operators should be familiar with.

Table of Contents

1. Introduction

   The BGP MED attribute provides a mechanism for BGP speakers to convey
   to an adjacent AS the optimal entry point into the local AS. While
   BGP MEDs function correctly in many scenarios, there are a number of
   issues which may arise when utilizing MEDs in dynamic or complex
   topologies.

   While reading this document it's important to keep in mind that the
   goal is to discuss both implementation and deployment considerations
   regarding BGP MEDs and provide and guidance which both implementors
   and network operators should be familiar with. In some instances
   implementation advice varies from deployment advice.

1.1. About the MULTI_EXIT_DISC (MED) Attribute

   The BGP MULTI_EXIT_DISC (MED) attribute, formerly known as the
   INTER_AS_METRIC, is currently defined in section 5.1.4 of [BGP4], as
   follows:

     The MULTI_EXIT_DISC is an optional non-transitive attribute which
     is intended to be used on external (inter-AS) links to discriminate
     among multiple exit or entry points to the same neighboring AS.
     The value of the MULTI_EXIT_DISC attribute is a four octet unsigned
     number which is called a metric. All other factors being equal, the
     exit point with lower metric SHOULD be preferred. If received over
     EBGP, the MULTI_EXIT_DISC attribute MAY be propagated over IBGP to
     other BGP speakers within the same AS (see also 9.1.2.2). The
     MULTI_EXIT_DISC attribute received from a neighboring AS MUST NOT
     be propagated to other neighboring ASs.

     A BGP speaker MUST implement a mechanism based on local
     configuration which allows the MULTI_EXIT_DISC attribute to be
     removed from a route. If a BGP speaker is configured to remove the
     MULTI_EXIT_DISC attribute from a route, then this removal MUST be
     done prior to determining the degree of preference of the route and
     performing route selection (Decision Process phases 1 and 2).

     An implementation MAY also (based on local configuration) alter the
     value of the MULTI_EXIT_DISC attribute received over EBGP. If a
     BGP speaker is configured to alter the value of the MULTI_EXIT_DISC
     attribute received over EBGP, then altering the value MUST be done
     prior to determining the degree of preference of the route and
     performing route selection (Decision Process phases 1 and 2). See

   Section 9.1.2.2 of BGP4] for necessary restrictions on this.

  Section 9.1.2.2 (c) of [BGP4] defines the following route selection
  criteria regarding MEDs:

    c) Remove from consideration routes with less-preferred
    MULTI_EXIT_DISC attributes. MULTI_EXIT_DISC is only comparable
    between routes learned from the same neighboring AS (the neighbor-
    ing AS is determined from the AS_PATH attribute). Routes which do
    not have the MULTI_EXIT_DISC attribute are considered to have the
    lowest possible MULTI_EXIT_DISC value.

    This is also described in the following procedure:

    for m = all routes still under consideration
     for n = all routes still under consideration
      if (neighborAS(m) == neighborAS(n)) and (MED(n) < MED(m))
        remove route m from consideration

    In the pseudo-code above, MED(n) is a function which returns the
    value of route n's MULTI_EXIT_DISC attribute. If route n has no
    MULTI_EXIT_DISC attribute, the function returns the lowest possi-
    ble MULTI_EXIT_DISC value, i.e. 0.

    If a MULTI_EXIT_DISC attribute is removed before re-advertising a
    route into IBGP, then comparison based on the received EBGP
    MULTI_EXIT_DISC attribute MAY still be performed. If an
    implementation chooses to remove MULTI_EXIT_DISC, then the optional
    comparison on MULTI_EXIT_DISC if performed at all MUST be performed
    only among EBGP learned routes. The best EBGP learned route may
    then be compared with IBGP learned routes after the removal of the
    MULTI_EXIT_DISC attribute. If MULTI_EXIT_DISC is removed from a
    subset of EBGP learned routes and the selected "best" EBGP learned
    route will not have MULTI_EXIT_DISC removed, then the
    MULTI_EXIT_DISC must be used in the comparison with IBGP learned
    routes. For IBGP learned routes the MULTI_EXIT_DISC MUST be used in
    route comparisons which reach this step in the Decision Process.
    Including the MULTI_EXIT_DISC of an EBGP learned route in the
    comparison with an IBGP learned route, then removing the
    MULTI_EXIT_DISC attribute and advertising the route has been proven
    to cause route loops.

1.2. MEDs and Potatos

   In a situation where traffic flows between a pair of hosts, each

connected to different transit networks, which are themselves
interconnected at two or more locations, each transit network has the
choice of either sending traffic to the closest peering to the
adjacent transit network or passing traffic to the interconnection
location which advertises the least cost path to the destination
host.

The former method is called "hot potato routing" (or closest-exit)
because like a hot potato held in bare hands, whoever has it tries to
get rid of it quickly. Hot potato routing is accomplished by not
passing the EGBP learned MED into IBGP. This minimizes transit
traffic for the provider routing the traffic. Far less common is
"cold potato routing" (or best-exit) where the transit provider uses
their own transit capacity to get the traffic to the point that
adjacent transit provider advertised as being closest to the
destination. Cold potato routing is accomplished by passing the EBGP
learned MED into IBGP.

If one transit provider uses hot potato routing and another uses cold
potato, traffic between the two tends to be more symmetric. However,
if both providers employ cold potato routing, or both providers
employ hot potato routing between their networks, it's likely that a
larger amount of asymmetry would exist.

Depending on the business relationships, if one provider has more
capacity or a significantly less congested backbone network, then
that provider may use cold potato routing. An example of widespread
use of cold potato routing was the NSF funded NSFNET backbone and NSF
funded regional networks in the mid 1990s.

In some cases a provider may use hot potato routing for some
destinations for a given peer AS and cold potato routing for others.
An example of this is the different treatment of commercial and
research traffic in the NSFNET in the mid 1990s. Today many
commercial networks exchange MEDs with customers but not bilateral
peers. However, commercial use of MEDs varies widely, from
ubiquitous use of MEDs to no use of MEDs at all.

In addition, many deployments of MEDs today are likely behaving
differently (e.g., resulting is sub-optimal routing) than the network
operator intended, thereby resulting not in hot or cold potatos, but
mashed potatos! More information on unintended behavior resulting
from MEDs is provided throughout this document.

2. Implementation and Protocol Considerations

There are a number of implementation and protocol peculiarities
relating to MEDs that have been discovered that may affect network
behavior. The following sections provide information on these
issues.

2.1. MULTI_EXIT_DISC is a Optional Non-Transitive Attribute

MULTI_EXIT_DISC is a non-transitive optional attribute whose
advertisement to both IBGP and EBGP peers is discretionary. As a
result, some implementations enable sending of MEDs to IBGP peers by
default, while others do not. This behavior may result in sub-
optimal route selection within an AS. In addition, some
implementations send MEDs to EBGP peers by default, while others do
not. This behavior may result in sub-optimal inter-domain route
selection.

2.2. MED Values and Preferences

Some implementations consider an MED value of zero as less preferable
than no MED value. This behavior resulted in path selection
inconsistencies within an AS. The current draft version of the BGP
specification [BGP4] removes ambiguities that existed in [RFC 1771]
by stating that if route n has no MULTI_EXIT_DISC attribute, the
lowest possible MULTI_EXIT_DISC value (i.e. 0) should be assigned to
the attribute.

It is apparent that different implementations and different versions
of the BGP draft specification have been all over the map with
interpretation of missing-MED. For example, earlier versions of the
specification called for a missing MED to be assigned the highest
possible MED value (i.e., $2^{32}-1$).

In addition, some implementations have been shown to internally
employ a maximum possible MED value ($2^{32}-1$) as an "infinity" metric
(i.e., the MED value is used to tag routes as unfeasible), and would
upon on receiving an update with an MED value of $2^{32}-1$ rewrite the
value to $2^{32}-2$. Subsequently, the new MED value would be propagated
and could result in routing inconsistencies or unintended path
selections.

As a result of implementation inconsistencies and protocol revision
variances, many network operators today explicitly reset (i.e., set
to zero or some other 'fixed' value) all MED values on ingress to
conform to their internal routing policies (i.e., to include policy
that requires that MED values of 0 and 2^32-1 NOT be used in
configurations, whether the MEDs are directly computed or
configured), so as to not have to rely on all their routers having
the same missing-MED behavior.

Because implementations don't normally provide a mechanism to disable
MED comparisons in the decision algorithm, "not using MEDs" usually
entails explicitly setting all MEDs to some fixed value upon ingress
to the routing domain. By assigning a fixed MED value consistently
to all routes across the network, MEDs are a effectively a non-issue
in the decision algorithm.

2.3. Comparing MEDs Between Different Autonomous Systems

The MED was intended to be used on external (inter-AS) links to
discriminate among multiple exit or entry points to the same
neighboring AS. However, a large number of MED applications now
employ MEDs for the purpose of determining route preference between
like routes received from different autonomous systems.

A large number of implementations provide the capability to enable
comparison of MEDs between routes received from different neighboring
autonomous systems. While this capability has demonstrated some
benefit (e.g., that described in [RFC 3345]), operators should be
wary of the potential side effects with enabling such a function.
The deployment section below provides some examples as to why this
may result in undesirable behavior.

2.4. MEDs, Route Reflection and AS Confederations for BGP

In particular configurations, the BGP scaling mechanisms defined in
"BGP Route Reflection - An Alternative to Full Mesh IBGP" [RFC 2796]
and "Autonomous System Confederations for BGP" [RFC 3065] will
introduce persistent BGP route oscillation [RFC 3345]. The problem
is inherent in the way BGP works: a conflict exists between
information hiding/hierarchy and the non-hierarchical selection
process imposed by lack of total ordering caused by the MED rules.

Given current practices, we see the problem most frequently manifest
itself in the context of MED + route reflectors or confederations.

One potential way to avoid this is by configuring inter-Member-AS or
inter-cluster IGP metrics higher than intra-Member-AS IGP metrics
and/or using other tie breaking policies to avoid BGP route selection
based on incomparable MEDs. Of course, IGP metric constraints may be
unreasonably onerous for some applications.

Comparing MEDs between differing adjacent autonomous systems
discussed in section 2.3), or not utilizing MEDs at all,
significantly decreases the probability of introducing potential
route oscillation conditions into the network.

Although perhaps "legal" as far as current specifications are
concerned, modifying MED attributes received on any type of IBGP
session (e.g., standard IBGP, AS confederations EIBGP, route
reflection, etc..) is NOT recommended.

2.5. Route Flap Damping and MED Churn

MEDs are often derived dynamically from IGP metrics or additive costs
associated with an IGP metric to a given BGP NEXT_HOP. This
typically provides an efficient model for ensuring that the BGP MED
advertised to peers used to represent the best path to a given
destination within the network is aligned with that of the IGP within
a given AS.

The consequence with dynamically derived IGP-based MEDs is that
instability within an AS, or even on a single given link within the
AS, can result in wide-spread BGP instability or BGP route
advertisement churn that propagates across multiple domains. In
short, if your MED "flaps" every time your IGP metric flaps, you're
routes are likely going to be suppressed as a result of BGP Route
Flap Damping [RFC 2439].

Employment of MEDs may compound the adverse effects of BGP flap
dampening behavior because it many cause routes to be re- advertised
solely to reflect an internal topology change.

Many implementations don't have a practical problem with IGP
flapping, they either latch their IGP metric upon first advertisement
or they employ some internal suppression mechanism. Some
implementations regard BGP attribute changes as less significant than
route withdrawals and announcements to attempt to mitigate the impact

of this type of event.

[2.6](#). **Effects of MEDs on Update Packing Efficiency**

Multiple unfeasible routes can be advertised in a single BGP Update
message. The BGP4 protocol also permits advertisement of multiple
prefixes with a common set of path attributes to be advertised in a
single update message, this is commonly referred to as "update
packing". When possible, update packing is recommended as it
provides a mechanism for more efficient behavior in a number of
areas, to include:

 o Reduction in system overhead due to generation or receipt of
   fewer Update messages.

 o Reduction in network overhead as a result of fewer packets and
   lower bandwidth consumption.

 o Allows processing of path attributes and searches for matching
   sets in your AS_PATH database (if you have one) less frequently.
   Consistent ordering of the path attributes allows for ease of
   matching in the database as you don't have different
representations
   of the same data.

Update packing requires that all feasible routes within a single
update message share a common attribute set, to include a common
MULTI_EXIT_DISC value. As such, potential wide-scale variance in MED
values introduces another variable and may resulted in a marked
decrease in update packing efficiency.

[2.7](#). **Temporal Route Selection**

Some implementations have had bugs which lead to temporal behavior in
MED-based best path selection. These usually involved methods used
to store the oldest route along with ordering routes for MED in
earlier implementations that cause non-deterministic behavior on
whether the oldest route would truly be selected or not.

The reasoning for this is that older paths are presumably more
stable, and thus more preferable. However, temporal behavior in
route selection results in non-deterministic behavior, and as such,

is often undesirable.

3. Deployment Considerations

It has been discussed that accepting MEDs from other autonomous
systems have the potential to cause traffic flow churns in the
network. Some implementations only ratchet down the MED and never
move it back up to prevent excessive churn.

However, if a session is reset, the MEDs being advertised have the
potential of changing. If an network is relying on received MEDs to
route traffic properly, the traffic patterns have the potential for
changing dramatically, potentially resulting in congestion on the
network. Essentially, accepting and routing traffic based on MEDs
allows other people to traffic engineer your network. This may or may
not be acceptable to you.

As previously discussed, many network operators choose to reset MED
values on ingress. In addition, many operators explicitly do not
employ MED values of 0 or 2^32-1 in order to avoid inconsistencies
with implementations and various revisions of the BGP specification.

3.1. Comparing MEDs Between Different Autonomous Systems

Although the MED was meant to only be used when comparing paths
received from different external peers in the same AS, many
implementations provide the capability to compare MEDs between
different autonomous systems as well. AS operators often use
LOCAL_PREF to select the external preferences (primary, secondary
upstreams, peers, customers, etc.), using MED instead of LOCAL_PREF
would possibility lead to an inconsistent distribution of best routes
as MED is compared only after the AS_PATH length.

Though this may seem a fine idea for some configurations, care must
be taken when comparing MEDs between different autonomous systems.
BGP speakers often derive MED values by obtaining the IGP metric
associated with reaching a given BGP NEXT_HOP within the local AS.
This allows MEDs to reasonably reflect IGP topologies when
advertising routes to peers. While this is fine when comparing MEDs
between multiple paths learned from a single AS, it can result in
potentially "weighted" decisions when comparing MEDs between
different autonomous systems. This is most typically the case when

   the autonomous systems use different mechanisms to derive IGP
   metrics, BGP MEDs, or perhaps even use different IGP protocols with
   vastly contrasting metric spaces (e.g., OSPF v. traditional metric
   space in IS-IS).

3.2. Effects of Aggregation on MEDs`

   Another MED deployment consideration involves the impact that
   aggregation of BGP routing information has on MEDs. Aggregates are
   often generated from multiple locations in an AS in order to
   accommodate stability, redundancy and other network design goals.
   When MEDs are derived from IGP metrics associated with said
   aggregates the MED value advertised to peers can result in very
   suboptimal routing.

4. IANA Considerations

   This document introduces no new IANA considerations.

5. Security Considerations

   The MED was purposely designed to be a "weak" metric that would only
   be used late in the best-path decision process. The BGP working
   group was concerned that any metric specified by a remote operator
   would only affect routing in a local AS IF no other preference was
   specified. A paramount goal of the design of the MED was to ensure
   that peers could not "shed" or "absorb" traffic for networks that
   they advertise. As such, accepting MEDs from peers may in some sense
   increase a network's susceptibility to exploitation by peers.

5.1. Acknowledgments

   Thanks to John Scudder for applying his usual keen eye and
   constructive insight. Also, thanks to Curtis Villamizar, JR Mitchell
   and Pekka Savola for their valuable feedback.

6. References

6.1. Normative References

   [RFC 1519] Fuller, V., Li. T., Yu J., and K. Varadhan, "Classless
              Inter-Domain Routing (CIDR): an Address Assignment and
              Aggregation Strategy", RFC 1519, September 1993.

   [RFC 1771] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4
              (BGP-4)", RFC 1771, March 1995.

   [RFC 2796] Bates, T., Chandra, R., Chen, E., "BGP Route Reflection
              - An Alternative to Full Mesh IBGP", RFC 2796, April
              2000.

   [RFC 3065] Traina, P., McPherson, D., Scudder, J.. "Autonomous System
              Confederations for BGP", RFC 3065, February 2001.

   [BGP4] Rekhter, Y., T. Li., and Hares. S, Editors, "A Border
          Gateway Protocol 4 (BGP-4)", BGP Draft, Work in Progress.

6.2. **Informative References**

   [RFC 2439] Villamizar, C. and Chandra, R., "BGP Route Flap Damping",
              RFC 2439, November 1998.

   [RFC 3345] McPherson, D., Gill, V., Walton, D., and Retana, A, "BGP
              Persistent Route Oscillation Condition", RFC 3345,
              August 2002.

7. **Authors' Addresses**

   Danny McPherson
   Arbor Networks
   Email: danny@arbor.net

   Vijay Gill
   AOL
   Email: VijayGill9@aol.com

   this standard. Please address the information to the IETF at
   ietf-ipr@ietf.org.

Disclaimer of Validity

Copyright Statement

Acknowledgment