Network Working Group
Request for Comments: 4451
Category: Informational

D. McPherson
Arbor Networks, Inc.
V. Gill
AOL

March 2006

BGP MULTI_EXIT_DISC (MED) Considerations

Status of This Memo

Copyright Notice

Abstract

The BGP MULTI_EXIT_DISC (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, a number of issues may arise when utilizing MEDs in dynamic or complex topologies.

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

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, a number of issues may arise when utilizing MEDs in dynamic or complex topologies.

While reading this document, note that the goal is to discuss both implementation and deployment considerations regarding BGP MEDs. In addition, the intention is to provide guidance that both implementors and network operators should be familiar with. In some instances, implementation advice varies from deployment advice.

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].

2.1. 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 that 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, called a metric. All other factors being equal, the exit point with the lower metric SHOULD be preferred. If received over External BGP (EBGP), the MULTI_EXIT_DISC attribute MAY be propagated over Internal BGP (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 ASes.
>
> A BGP speaker MUST implement a mechanism (based on local configuration) that 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 prior to 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 prior to performing route selection (Decision Process phases 1 and 2). See Section 9.1.2.2 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 neighboring AS is determined from the AS_PATH attribute). Routes that 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 that 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 possible MULTI_EXIT_DISC value (i.e., 0).

Similarly, neighborAS(n) is a function that returns the neighbor AS from which the route was received. If the route is learned via IBGP, and the other IBGP speaker didn't originate the route, it is the neighbor AS from which the other IBGP speaker learned the route. If the route is learned via IBGP, and the other IBGP speaker either (a) originated the route, or (b) created the route by aggregation and the AS_PATH attribute of the aggregate route is either empty or begins with an AS_SET, it is the local AS.

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, 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 that 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.

2.2. MEDs and Potatoes

Let’s consider a situation where traffic flows between a pair of hosts, each connected to a different transit network, which is in itself 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 that 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 EBGP-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 its 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 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 with bilateral peers. However, commercial use of MEDs varies widely, from ubiquitous use to none at all.

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

3. 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.

3.1. MULTI_EXIT_DISC Is an 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.

3.2. MED Values and Preferences

Some implementations consider an MED value of zero less preferable than no MED value. This behavior resulted in path selection inconsistencies within an AS. The current version of the BGP specification [BGP4] removes ambiguities that existed in [RFC1771] 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 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); upon receiving an update with an MED value of 2^32-1, they would 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 not to 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.

3.3. 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 [RFC3345]), operators should be wary of the potential side effects of enabling such a function. The deployment section below provides some examples as to why this may result in undesirable behavior.

3.4. 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" [RFC2796] and "Autonomous System Confederations for BGP" [RFC3065] will introduce persistent BGP route oscillation [RFC3345]. 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 manifest itself most frequently 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.

Not comparing MEDs between multiple paths for a prefix learned from different adjacent autonomous systems, as 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, EBGP sessions between Member-ASes of a BGP confederation, route reflection, etc.) is not recommended.

3.5. 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 widespread BGP instability or BGP route advertisement churn that propagates across multiple domains. In short, if your MED "flaps" every time your IGP metric flaps, your routes are likely going to be suppressed as a result of BGP Route Flap Damping [RFC2439].

Employment of MEDs may compound the adverse effects of BGP flap-dampening behavior because it may 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 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.

3.6. 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, including the following:

- Reduction in system overhead due to generation or receipt of fewer Update messages.
- Reduction in network overhead as a result of fewer packets and lower bandwidth consumption.
- Less frequent processing of path attributes and searches for matching sets in your AS_PATH database (if you have one). 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 result in a marked decrease in update packing efficiency.

3.7. Temporal Route Selection

Some implementations had bugs that led to temporal behavior in MED-based best path selection. These usually involved methods to store the oldest route and to order routes for MED, which caused non-deterministic behavior as to whether or not the oldest route would truly be selected.

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

4. Deployment Considerations

It has been discussed that accepting MEDs from other autonomous systems has 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 a 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.

4.1. Comparing MEDs between Different Autonomous Systems

Although the MED was meant to be used only 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 possibly lead to an inconsistent distribution of best routes, as MED is compared only after the AS_PATH length.

Though this may seem like 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 or BGP MEDs, or when they perhaps even use different IGP protocols with vastly contrasting metric spaces (e.g., OSPF vs. traditional metric space in IS-IS).

4.2. 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.

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.

6. Acknowledgements

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.

7. References

7.1. Normative References

[RFC1771] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

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

[RFC2796] Bates, T., Chandra, R., and E. Chen, "BGP Route Reflection - An Alternative to Full Mesh IBGP", RFC 2796, April 2000.

[RFC3065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 3065, February 2001.

[BGP4] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

7.2. Informative References

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, November 1998.

[RFC3345] McPherson, D., Gill, V., Walton, D., and A. Retana, "Border Gateway Protocol (BGP) Persistent Route Oscillation Condition", RFC 3345, August 2002.

Authors' Addresses

Danny McPherson
Arbor Networks

EMail: danny@arbor.net

Vijay Gill
AOL

EMail: VijayGill9@aol.com

Full Copyright Statement

Intellectual Property

Acknowledgement

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).