INTERNET-DRAFT Danny McPherson
Arbor Networks, Inc.
Vijay Gill
AOL
Category Informational
Expires: November 2004 May 2004

BGP MED Considerations
<draft-ietf-grow-bgp-med-considerations-01.txt>

Status of this Document

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

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/ietf/1id-abstracts.txt

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

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 RFC 2119 [RFC 2119].

This document is a product of an individual. Comments are solicited and should be addressed to the author(s).

Copyright Notice

Copyright (C) The Internet Society (2004). All Rights Reserved.

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 4
1.1. About the MULTI_EXIT_DISC (MED) Attribute 4
1.2. MEDs and Potatoes 5
2. Implementation and Protocol Considerations 7
2.1. MULTI_EXIT_DISC is a Optional Non-Transitive Attribute. 7
2.2. MED Values and Preferences. 7
2.3. Comparing MEDs Between Different Autonomous Systems. 8
2.4. MEDs, Route Reflection and AS Confederations for BGP. 8
2.5. Route Flap Damping and MED Churn. 9
2.6. Effects of MEDs on Update Packing Efficiency. 10
2.7. Temporal Route Selection. 10
3. Deployment Considerations. 11
3.1. Comparing MEDs Between Different Autonomous Systems. 11
3.2. Effects of Aggregation on MEDs` 12
4. Security Considerations. 12
4.1. Acknowledgments 12
5. References 13
6. Authors' Addresses 13
7. Full Copyright Statement 13

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.

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

1.1. About the MULTI_EXIT_DISC (MED) Attribute

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

> 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 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. An MED attribute received from a neighboring AS MUST NOT be propagated to other neighboring autonomous systems.
>
> A BGP speaker MUST IMPLEMENT a mechanism based on local configuration which allows the MULTI_EXIT_DISC attribute to be removed from a route. This MAY 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. This MAY be done prior to determining the degree of preference of the route and performing route selection (decision process phases 1 and 2).

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

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 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 =3D all routes still under consideration
  for n =3D all routes still under consideration
    if (neighborAS(m) =3D=3D 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 possible 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.

Routes that have different MULTI_EXIT_DISC attribute SHALL NOT be aggregated.

1.2. MEDs and Potatoes

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.

DIAGRAM***

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.

DIAGRAMS***

If one transit provider uses hot potato routing and another uses cold potato, traffic between the two tends to be more symmetric. 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 potatoes, but mashed potatoes! 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.

DIAGRAM and MORE TEXT***

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 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.

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 (which will be discussed in later sections), 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. In addition, one or more feasible routes can be advertised in a single Update message so long as all prefixes share a common attribute set.

The BGP4 protocol 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:

- 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.

- 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

Empirical data [MFN/Ixia Monitoring Project] has shown 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 that 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.

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.

DIAGRAM AND EXPAND***

4. 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.

4.1. Acknowledgments

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

Others to be supplied.

5. 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 2439] Villamizar, C. and Chandra, R., "BGP Route Flap Damping", RFC 2439, November 1998.

[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.

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

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

[MFN/Ixia Monitoring Project] Vijay to Provide.

6. Authors' Addresses

Danny McPherson
Arbor Networks
Email: danny@arbor.net

Vijay Gill
AOL
Email: VijayGill9@aol.com

7. Full Copyright Statement

Copyright (C) The Internet Society (2004). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.