Network Working Group
Request for Comments: 3345
Category: Informational

D. McPherson
TCB
V. Gill
AOL Time Warner, Inc.
D. Walton
A. Retana
Cisco Systems, Inc.

August 2002

Border Gateway Protocol (BGP) Persistent Route Oscillation Condition

Status of this Memo

Copyright Notice

Abstract

In particular configurations, the BGP scaling mechanisms defined in "BGP Route Reflection - An Alternative to Full Mesh IBGP" and "Autonomous System Confederations for BGP" will introduce persistent BGP route oscillation. This document discusses the two types of persistent route oscillation that have been identified, describes when these conditions will occur, and provides some network design guidelines to avoid introducing such occurrences.

1. Introduction

The Border Gateway Protocol (BGP) is an inter-Autonomous System routing protocol. The primary function of a BGP speaking system is to exchange network reachability information with other BGP systems.

In particular configurations, the BGP [1] scaling mechanisms defined in "BGP Route Reflection - An Alternative to Full Mesh IBGP" [2] and "Autonomous System Confederations for BGP" [3] will introduce persistent BGP route oscillation.

The problem is inherent in the way BGP works: locally defined routing policies may conflict globally, and certain types of conflicts can cause persistent oscillation of the protocol. Given current practices, we happen to see the problem manifest itself in the context of MED + route reflectors or confederations.

The current specification of BGP-4 [4] states that the MULTI_EXIT_DISC is only comparable between routes learned from the same neighboring AS. This limitation is consistent with the description of the attribute: "The MULTI_EXIT_DISC attribute may be used on external (inter-AS) links to discriminate among multiple exit or entry points to the same neighboring AS." [1,4]

In a full mesh iBGP network, all the internal routers have complete visibility of the available exit points into a neighboring AS. The comparison of the MULTI_EXIT_DISC for only some paths is not a problem.

Because of the scalability implications of a full mesh iBGP network, two alternatives have been standardized: route reflectors [2] and AS confederations [3]. Both alternatives describe methods by which route distribution may be achieved without a full iBGP mesh in an AS.

The route reflector alternative defines the ability to re-advertise (reflect) iBGP-learned routes to other iBGP peers once the best path is selected [2]. AS Confederations specify the operation of a collection of autonomous systems under a common administration as a single entity (i.e. from the outside, the internal topology and the existence of separate autonomous systems are not visible). In both cases, the reduction of the iBGP full mesh results in the fact that not all the BGP speakers in the AS have complete visibility of the available exit points into a neighboring AS. In fact, the visibility may be partial and inconsistent depending on the location (and function) of the router in the AS.

In certain topologies involving either route reflectors or confederations (detailed description later in this document), the partial visibility of the available exit points into a neighboring AS may result in an inconsistent best path selection decision as the routers don't have all the relevant information. If the inconsistencies span more than one peering router, they may result in a persistent route oscillation. The best path selection rules applied in this document are consistent with the current specification [4].

The persistent route oscillation behavior is deterministic and can be avoided by employing some rudimentary BGP network design principles until protocol enhancements resolve the problem.

In the following sections a taxonomy of the types of oscillations is presented and a description of the set of conditions that will trigger route oscillations is given. We continue by providing several network design alternatives that remove the potential of this occurrence.

It is the intent of the authors that this document serve to increase operator awareness of the problem, as well as to trigger discussion and subsequent proposals for potential protocol enhancements that remove the possibility of this to occur.

The oscillations are classified into Type I and Type II depending upon the criteria documented below.

2. Discussion of Type I Churn

In the following two subsections we provide configurations under which Type I Churn will occur. We begin with a discussion of the problem when using Route Reflection, and then discuss the problem as it relates to AS Confederations.

In general, Type I Churn occurs only when BOTH of the following conditions are met:

1) a single-level Route Reflection or AS Confederations design is used in the network AND

2) the network accepts the BGP MULTI_EXIT_DISC (MED) attribute from two or more ASs for a single prefix and the MED values are unique.

It is also possible for the non-deterministic ordering of paths to cause the route oscillation problem. [1] does not specify that paths should be ordered based on MEDs but it has been proven that non-deterministic ordering can lead to loops and inconsistent routing decisions. Most vendors have either implemented deterministic ordering as default behavior, or provide a knob that permits the operator to configure the router to order paths in a deterministic manner based on MEDs.

2.1. Route Reflection and Type I Churn

We now discuss Type I oscillation as it relates to Route Reflection. To begin, consider the topology depicted in Figure 1:

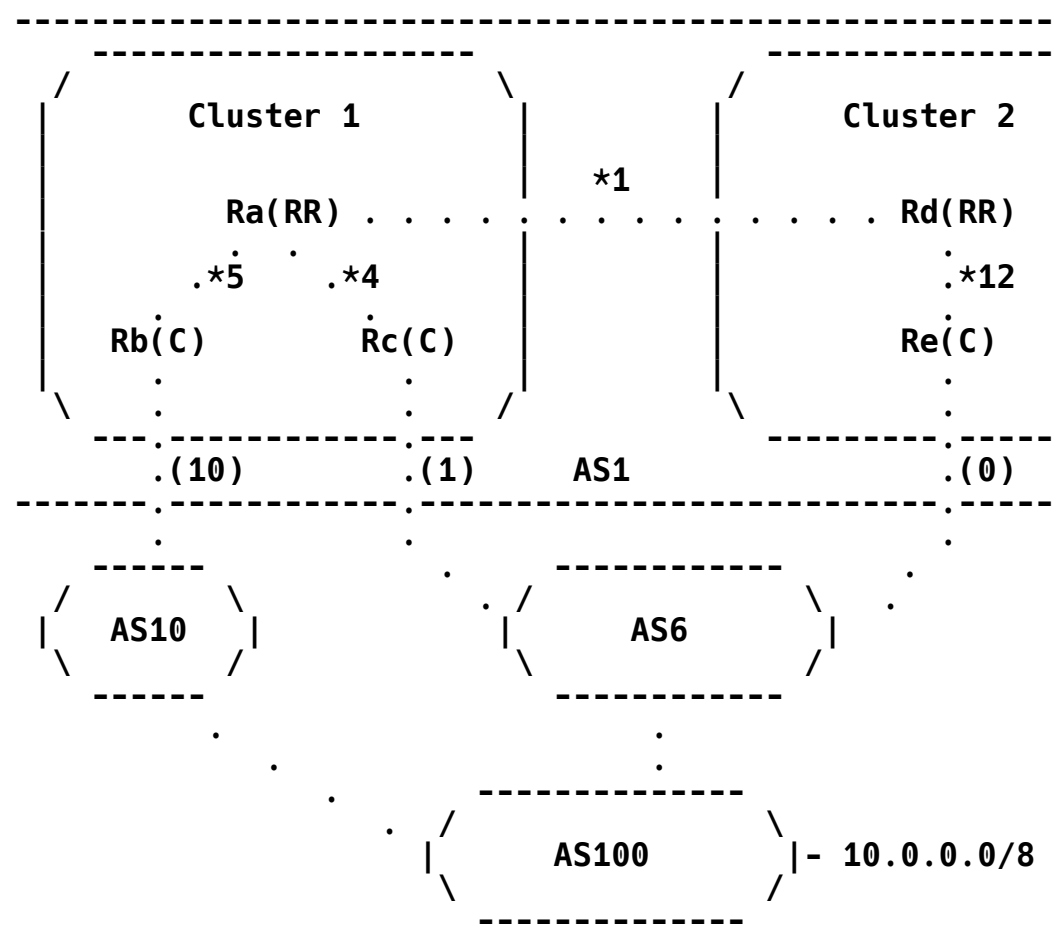

Figure 1: Example Route Reflection Topology

In Figure 1 AS1 contains two Route Reflector Clusters, Clusters 1 and 2. Each Cluster contains one Route Reflector (RR) (i.e., Ra and Rd, respectively). An associated 'RR' in parentheses represents each RR. Cluster 1 contains two RR Clients (Rb and Rc), and Cluster 2 contains one RR Client (Re). An associated 'C' in parentheses indicates RR Client status. The dotted lines are used to represent BGP peering sessions.

The number contained in parentheses on the AS1 EBGP peering sessions represents the MED value advertised by the peer to be associated with the 10.0.0.0/8 network reachability advertisement.

The number following each '*' on the IBGP peering sessions represents the additive IGP metrics that are to be associated with the BGP NEXT_HOP attribute for the concerned route. For example, the Ra IGP metric value associated with a NEXT_HOP learned via Rb would be 5; while the metric value associated with the NEXT_HOP learned via Re would be 13.

Table 1 depicts the 10.0.0.0/8 route attributes as seen by routers Rb, Rc and Re, respectively. Note that the IGP metrics in Figure 1 are only of concern when advertising the route to an IBGP peer.

Router	MED	AS_PATH
Rb	10	10 100
Rc	1	6 100
Re	0	6 100

Table 1: Route Attribute Table

For the following steps 1 through 5, the best path will be marked with a '*'.

1) Ra has the following installed in its BGP table, with the path learned via AS2 marked best:

	AS_PATH	MED	NEXT_HOP IGP Cost
	6 100	1	4
*	10 100	10	5

The '10 100' route should not be marked as best, though this is not the cause of the persistent route oscillation. Ra realizes it has the wrong route marked as best since the '6 100' path has a lower IGP metric. As such, Ra makes this change and advertises an UPDATE message to its neighbors to let them know that it now considers the '6 100, 1, 4' route as best.

2) Rd receives the UPDATE from Ra, which leaves Rd with the following installed in its BGP table:

	AS_PATH	MED	NEXT_HOP IGP Cost
*	6 100	0	12
	6 100	1	5

Rd then marks the '6 100, 0, 12' route as best because it has a lower MED. Rd sends an UPDATE message to its neighbors to let them know that this is the best route.

3) Ra receives the UPDATE message from Rd and now has the following in its BGP table:

	AS_PATH	MED	NEXT_HOP IGP Cost
	6 100	0	13
	6 100	1	4
*	10 100	10	5

The first route (6 100, 0, 13) beats the second route (6 100, 1, 4) because of a lower MED. Then the third route (10 100, 10, 5) beats the first route because of lower IGP metric to NEXT_HOP. Ra sends an UPDATE message to its peers informing them of the new best route.

4) Rd receives the UPDATE message from Ra, which leaves Rd with the following BGP table:

	AS_PATH	MED	NEXT_HOP IGP Cost
	6 100	0	12
*	10 100	10	6

Rd selects the '10 100, 10, 6' path as best because of the IGP metric. Rd sends an UPDATE/withdraw to its peers letting them know this is the best route.

5) Ra receives the UPDATE message from Rd, which leaves Ra with the following BGP table:

	AS_PATH	MED	NEXT_HOP IGP Cost
	6 100	1	4
*	10 100	10	5

Ra received an UPDATE/withdraw for '6 100, 0, 13', which changes what is considered the best route for Ra. This is why Ra has the '10 100, 10, 5' route selected as best in Step 1, even though '6 100, 1, 4' is actually better.

At this point, we've made a full loop and are back at Step 1. The router realizes it is using the incorrect best path, and repeats the cycle. This is an example of Type I Churn when using Route Reflection.

2.2. AS Confederations and Type I Churn

Now we provide an example of Type I Churn occurring with AS Confederations. To begin, consider the topology depicted in Figure 2:

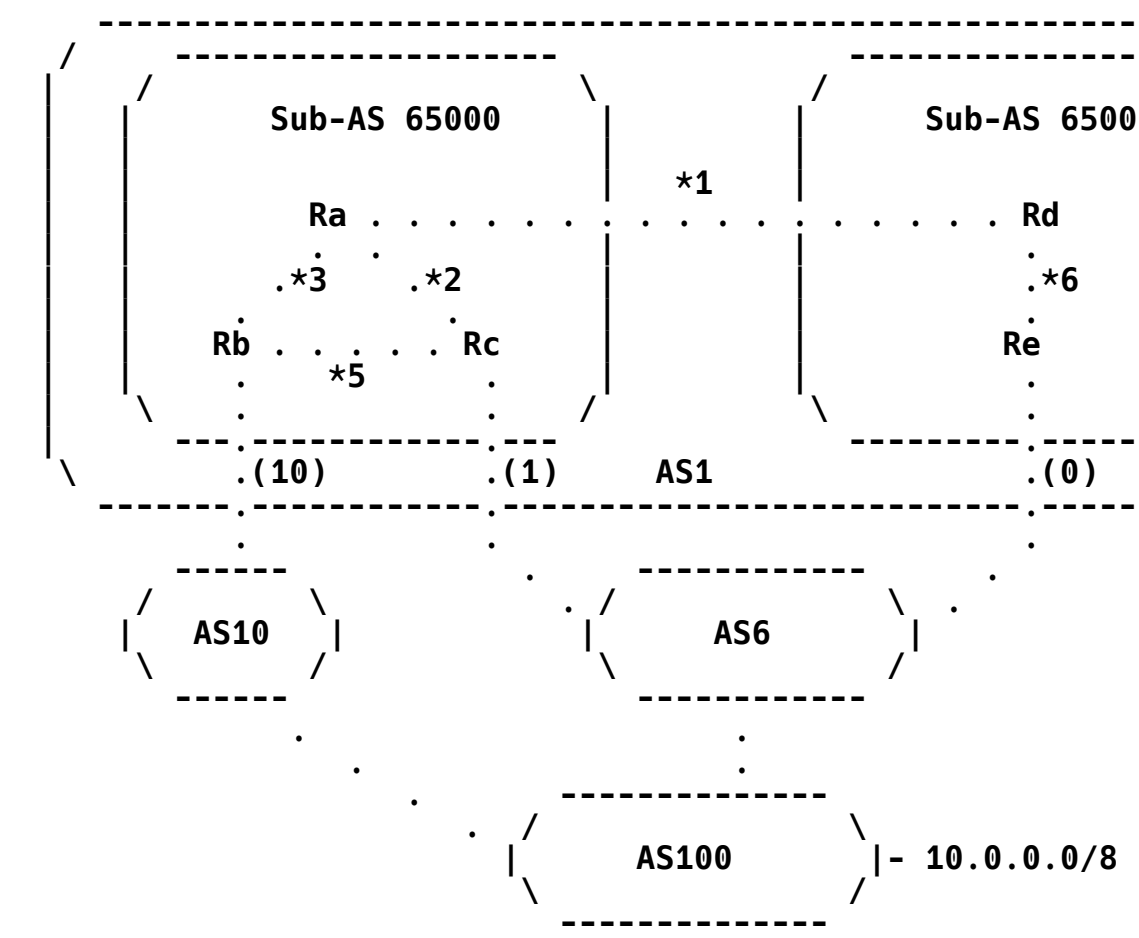

Figure 2: Example AS Confederations Topology

The number contained in parentheses on each AS1 EBGP peering session represents the MED value advertised by the peer to be associated with the 10.0.0.0/8 network reachability advertisement.

The number following each '*' on the IBGP peering sessions represents the additive IGP metrics that are to be associated with the BGP NEXT_HOP attribute for the concerned route.

For example, the Ra IGP metric value associated with a NEXT_HOP learned via Rb would be 3; while the metric value associated with the NEXT_HOP learned via Re would be 6.

Table 2 depicts the 10.0.0.0/8 route attributes as seen by routers Rb, Rc and Re, respectively. Note that the IGP metrics in Figure 2 are only of concern when advertising the route to an IBGP peer.

Router	MED	AS_PATH
Rb	10	10 100
Rc	1	6 100
Re	0	6 100

Table 2: Route Attribute Table

For the following steps 1 through 6 the best route will be marked with an '*'.

1) Ra has the following BGP table:

	AS_PATH	MED	NEXT_HOP IGP Cost
*	10 100	10	3
	(65001) 6 100	0	7
	6 100	1	2

The '10 100' route is selected as best and is advertised to Rd, though this is not the cause of the persistent route oscillation.

2) Rd has the following in its BGP table:

	AS_PATH	MED	NEXT_HOP IGP Cost
	6 100	0	6
*	(65000) 10 100	10	4

The '(65000) 10 100' route is selected as best because it has the lowest IGP metric. As a result, Rd sends an UPDATE/withdraw to Ra for the '6 100' route that it had previously advertised.

3) Ra receives the withdraw from Rd. Ra now has the following in its BGP table:

	AS_PATH	MED	NEXT_HOP IGP Cost
*	10 100	10	3
	6 100	1	2

Ra received a withdraw for '(65001) 6 100', which changes what is considered the best route for Ra. Ra does not compute the best path for a prefix unless its best route was withdrawn. This is why Ra has the '10 100, 10, 3' route selected as best, even though the '6 100, 1, 2' route is better.

4) Ra's periodic BGP scanner runs and realizes that the '6 100' route is better because of the lower IGP metric. Ra sends an UPDATE/withdraw to Rd for the '10 100' route since Ra is now using the '6 100' path as its best route.

Ra's BGP table looks like this:

	AS_PATH	MED	NEXT_HOP IGP Cost
	10 100	10	3
*	6 100	1	2

5) Rd receives the UPDATE from Ra and now has the following in its BGP table:

	AS_PATH	MED	NEXT_HOP IGP Cost
	(65000) 6 100	1	3
*	6 100	0	6

Rd selects the '6 100, 0, 6' route as best because of the lower MED value. Rd sends an UPDATE message to Ra, reporting that '6 100, 0, 6' is now the best route.

6) Ra receives the UPDATE from Rd. Ra now has the following in its BGP table:

	AS_PATH	MED	NEXT_HOP IGP Cost
*	10 100	10	3
	(65001) 6 100	0	7
	6 100	1	2

At this point we have made a full cycle and are back to step 1. This is an example of Type I Churn with AS Confederations.

2.3. Potential Workarounds for Type I Churn

There are a number of alternatives that can be employed to avoid this problem:

1) When using Route Reflection make sure that the inter-Cluster links have a higher IGP metric than the intra-Cluster links. This is the preferred choice when using Route Reflection. Had the inter-Cluster IGP metrics been much larger than the intra-Cluster IGP metrics, the above would not have occurred.

2) When using AS Confederations ensure that the inter-Sub-AS links have a higher IGP metric than the intra-Sub-AS links. This is the preferred option when using AS Confederations. Had the inter-Sub-AS IGP metrics been much larger than the intra-Sub-AS IGP metrics, the above would not have occurred.

3) Do not accept MEDs from peers (this may not be a feasible alternative).

4) Utilize other BGP attributes higher in the decision process so that the BGP decision algorithm never reaches the MED step. As using this completely overrides MEDs, Option 3 may make more sense.

5) Always compare BGP MEDs, regardless of whether or not they were obtained from a single AS. This is probably a bad idea since MEDs may be derived in a number of ways, and are typically done so as a matter of operator-specific policy. As such, comparing MED values for a single prefix learned from multiple ASs is ill-advised. Of course, this mostly defeats the purpose of MEDs, and as such, Option 3 may be a more viable alternative.

6) Use a full IBGP mesh. This is not a feasible solution for ASs with a large number of BGP speakers.

3. Discussion of Type II Churn

In the following subsection we provide configurations under which Type II Churn will occur when using AS Confederations. For the sake of brevity, we avoid similar discussion of the occurrence when using Route Reflection.

In general, Type II churn occurs only when BOTH of the following conditions are met:

1) More than one tier of Route Reflection or Sub-ASs is used in the network AND

2) the network accepts the BGP MULTI_EXIT_DISC (MED) attribute from two or more ASs for a single prefix and the MED values are unique.

3.1. AS Confederations and Type II Churn

Let's now examine the occurrence of Type II Churn as it relates to AS Confederations. Figure 3 provides our sample topology:

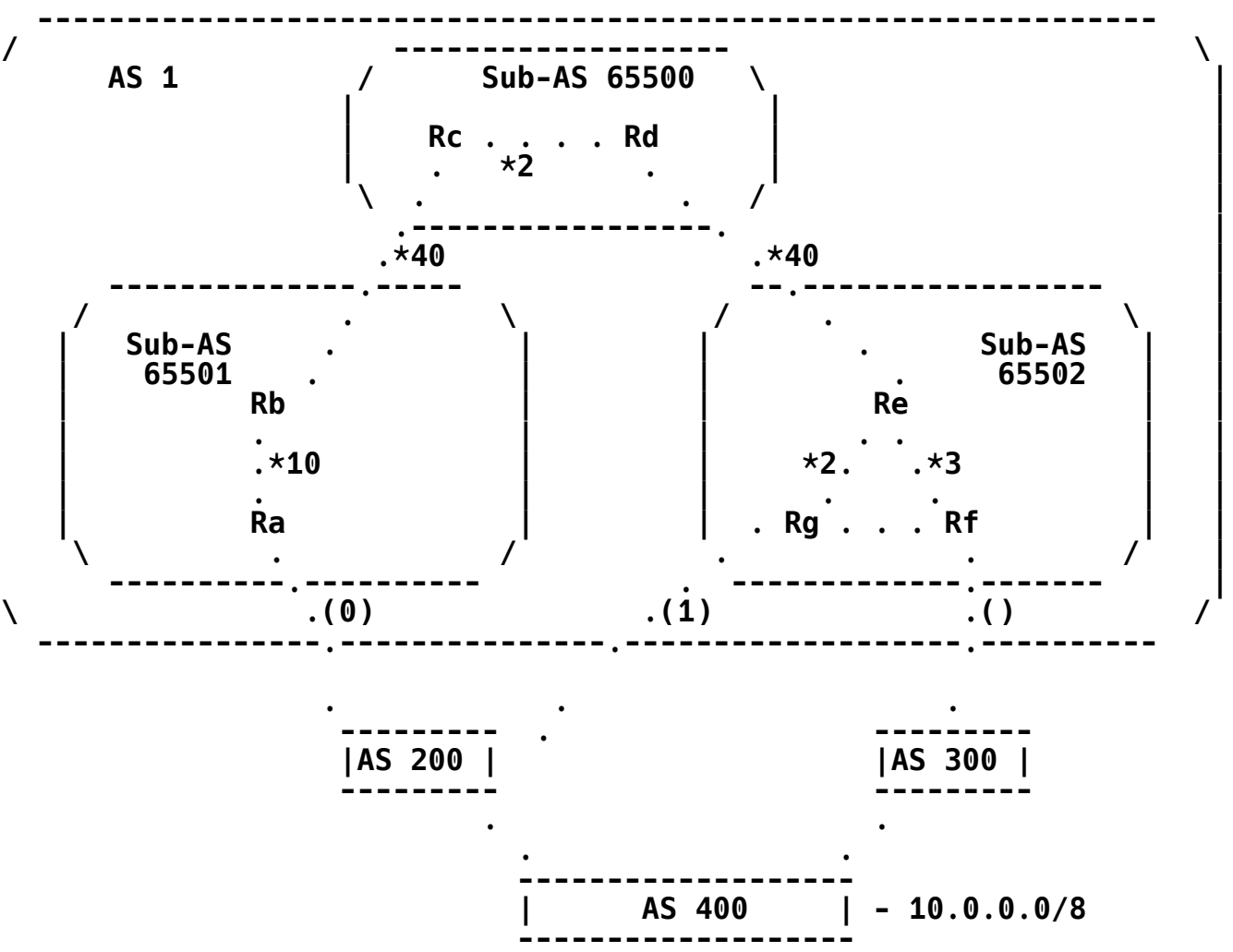

Figure 3: Example AS Confederations Topology

In Figure 3 AS 1 contains three Sub-ASs, 65500, 65501 and 65502. No RR is used within the Sub-AS, and as such, all routers within each Sub-AS are fully meshed. Ra and Rb are members of Sub-AS 65501. Rc and Rd are members of Sub-AS 65500. Ra and Rg are EBGP peering with AS 200, router Rf has an EBGP peering with AS 300. AS 200 and AS 300 provide transit for AS 400, and in particular, the 10/8 network. The dotted lines are used to represent BGP peering sessions.

The number following each '*' on the BGP peering sessions represents the additive IGP metrics that are to be associated with the BGP NEXT_HOP. The number contained in parentheses on each AS 1 EBGP peering session represents the MED value advertised by the peer to be associated with the network reachability advertisement (10.0.0.0/8).

Rc, Rd and Re are the primary routers involved in the churn, and as such, will be the only BGP tables that we will monitor step by step.

For the following steps 1 through 8 each router's best route will be marked with a '*'.

1) Re receives the AS 400 10.0.0.0/8 route advertisement via AS 200 from Rg and AS 300 from Rf. Re selects the path via Rg and AS 200 because of IGP metric (Re didn't consider MED because the advertisements were received from different ASs).

Router		AS_PATH	MED	NEXT_HOP IGP Cost
Re	*	200 400	1	2
		300 400		3

Re sends an UPDATE message to Rd advertising its new best path '200 400, 1'.

2) The '200 400, 0' path was advertised from Ra to Rb, and then from Rb to Rc. Rd learns the '200 400, 1' path from Re.

Router		AS_PATH	MED	NEXT_HOP IGP Cost
Rc	*	200 400	0	50
Rd	*	200 400	1	42
Re		300 400		3
	*	200 400	1	2

3) Rc and Rd advertise their best paths to each other; Rd selects '200 400, 0' because of the MED.

Router		AS_PATH	MED	NEXT_HOP IGP Cost
Rc	*	200 400	0	50
		200 400	1	44
Rd	*	200 400	0	52
		200 400	1	42
Re		300 400		3
	*	200 400	1	2

Rd has a new best path so it sends an UPDATE to to Re, announcing the new path and an UPDATE/withdraw for '200 400, 1' to Rc.

4) Re now selects '300 400' (with no MED) because '200 400, 0' beats '200 400, 1' based on MED and '300 400' beats '200 400, 0' because of IGP metric.

Router		AS_PATH	MED	NEXT_HOP IGP Cost
Rc	*	200 400	0	50
Rd	*	200 400	0	52
		200 400	1	42
Re	*	300 400		3
		200 400	0	92

Re has a new best path and sends an UPDATE to Rd for '300 400'.

5) Rd selects the '300 400' path because of IGP metric.

Router		AS_PATH	MED	NEXT_HOP IGP Cost
Rc	*	200 400	0	50
Rd		200 400	0	52
	*	300 400		43
Re	*	300 400		3
		200 400	0	92
		200 400	1	2

Rd has a new best path so it sends an UPDATE to Rc and a UPDATE/withdraw to Re for '200 400, 0'.

6) Rc selects '300 400' because of the IGP metric. Re selects '200 400, 1' because of the IGP metric.

Router	AS_PATH	MED	NEXT_HOP IGP Cost
Rc	200 400	0	50
	* 300 400		45
Rd	200 400	0	52
	* 300 400		43
Re	300 400		3
	* 200 400	1	2

Rc sends an UPDATE/withdraw for '200 400, 0' to Rd. Re sends an UPDATE for '200 400, 1' to Rd.

7) Rd selects '200 400, 1' as its new best path based on the IGP metric.

Router	AS_PATH	MED	NEXT_HOP IGP Cost
Rc	200 400	0	50
	* 300 400		45
Rd	* 200 400	1	42
Re	300 400		3
	* 200 400	1	2

Rd sends an UPDATE to Rc, announcing '200 400, 1' and implicitly withdraws '300 400'.

8) Rc selects '200 400, 0'.

Router	AS_PATH	MED	NEXT_HOP IGP Cost
Rc	* 200 400	0	50
	200 400	1	44
Rd	* 200 400	1	42
Re	300 400		3
	* 200 400	1	2

At this point we are back to Step 2 and are in a loop.

3.2. Potential Workarounds for Type II Churn

1) Do not accept MEDs from peers (this may not be a feasible alternative).

2) Utilize other BGP attributes higher in the decision process so that the BGP decision algorithm selects a single AS before it reaches the MED step. For example, if local-pref were set based on the advertising AS, then you first eliminate all routes except those in a single AS. In the example, router Re would pick either X or Y based on your local-pref and never change selections.

   This leaves two simple workarounds for the two types of problems.

   Type I: Make inter-cluster or inter-sub-AS link metrics higher than intra-cluster or intra-sub-AS metrics.

   Type II: Make route selections based on local-pref assigned to the advertising AS first and then use IGP cost and MED to make selection among routes from the same AS.

   Note that this requires per-prefix policies, as well as near intimate knowledge of other networks by the network operator. The authors are not aware of ANY [large] provider today that performs per-prefix policies on routes learned from peers. Implicitly removing this dynamic portion of route selection does not appear to be a viable option in today's networks. The main point is that an available workaround using local-pref so that no two AS's advertise a given prefix at the same local-pref solves type II churn.

3) Always compare BGP MEDs, regardless of whether or not they were obtained from a single AS. This is probably a bad idea since MEDs may be derived in a number of ways, and are typically done so as a matter of operator-specific policy and largely a function of available metric space provided by the employed IGP. As such, comparing MED values for a single prefix learned from multiple ASs is ill-advised. This mostly defeats the purpose of MEDs; Option 1 may be a more viable alternative.

4) Do not use more than one tier of Route Reflection or Sub-ASs in the network. The risk of route oscillation should be considered when designing networks that might use a multi-tiered routing isolation architecture.

5) In a RR topology, mesh the clients. For confederations, mesh the border routers at each level in the hierarchy. In Figure 3, for example, if Rb and Re are peers, then there's no churn.

4. Future Work

It should be stated that protocol enhancements regarding this problem must be pursued. Imposing network design requirements, such as those outlined above, are clearly an unreasonable long-term solution. Problems such as this should not occur under 'default' protocol configurations.

5. Security Considerations

This discussion introduces no new security concerns to BGP or other specifications referenced in this document.

6. Acknowledgments

The authors would like to thank Curtis Villamizar, Tim Griffin, John Scudder, Ron Da Silva, Jeffrey Haas and Bill Fenner.

7. References

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

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

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

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

8. Authors' Addresses

Danny McPherson
TCB
EMail: danny@tcb.net

Vijay Gill
AOL Time Warner, Inc.
12100 Sunrise Valley Drive
Reston, VA 20191
EMail: vijay@umbc.edu

Daniel Walton
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
EMail: dwalton@cisco.com

Alvaro Retana
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
EMail: aretana@cisco.com

9. Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.