Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 26, 2009

P. Marques
R. Fernando
Juniper Networks
E. Chen
P. Mohapatra
Cisco Systems
March 25, 2009

Advertisement of the best external route in BGP
draft-marques-idr-best-external-01.txt

Status of this Memo

Copyright Notice

Abstract

The base BGP specifications prevent a BGP speaker from advertising any route that is not the best route for a BGP destination. This document specifies a modification of this rule. Routes are divided into two categories, "external" and "internal". A specification is provided for choosing a "best external route" (for a particular value of the Network Layer Reachability Information). A BGP speaker is then allowed to advertise its "best external route" to its internal BGP peers, even if that is not the best route for the destination. The document explains why advertising the best external route can improve convergence time without causing routing loops. Additional benefits include reduction of inter-domain churn and avoidance of permanent route oscillation. The document also generalizes the notions of "internal" and "external" so that they can be applied to Route Reflector Clusters and Autonomous System Confederations.

Table of Contents

1. Introduction 4
1.1. Requirements Language 5
2. Algorithm for selection of best external route 5
3. Advertisement Rules 6
4. Consistency between routing and forwarding 6
5. Applications 8
5.1. Fast Connectivity Restoration 8
5.2. Inter-Domain Churn Reduction 9
5.3. Reducing Persistent IBGP oscillation 9
6. Acknowledgments 9
7. IANA Considerations 9
8. Security Considerations 9
9. Normative References 9
Authors' Addresses 10

1. Introduction

The base BGP specifications prevent a BGP speaker from advertising any route that is not the best route for a BGP destination. This document specifies a modification of this rule. Routes are divided into two categories, "external" and "internal". A specification is provided for choosing a "best external route" (for a particular value of the Network Layer Reachability Information). A BGP speaker is then allowed to advertise its "best external route" to its internal BGP peers, even if that is not the best route for the destination. The document explains why advertising the best external route can improve convergence time without causing routing loops. Additional benefits include reduction of inter-domain churn and avoidance of permanent route oscillation.

The document also generalizes the notions of "internal" and "external" so that they can be applied to Route Reflector Clusters [RFC4456] and Autonomous System Confederations [RFC5065]. More specifically, two routers in the same route reflector cluster having an IBGP session between them are defined to be "internal" peers, whereas two routers in different clusters having an IBGP session are defined to be "external" peers. Similarly, two routers in the same member AS of a confederation having an IBGP session between them are "internal" peers, whereas two routers in different member ASs of a confederation having a confed EBGP session between them are defined to be "external" peers. The definition of "best external route" ensues from this definition in that it is the most preferred route among those received from the "external" neighbors.

Advertising the best external route, when different from the best route, presents additional information into an IBGP mesh which may be of value for several purposes including:

- Faster restoration of connectivity, by providing additional paths, that may be used to fail over in case the primary path becomes invalid or is withdrawn.

- Reducing inter-domain churn and traffic blackholing due to the readily available alternate path.

- Reducing the potential for situations of permanent IBGP route oscillation, as discussed in some scenarios [RFC3345].

- Improving selection of lower MED routes from the same neighboring AS.

This document defines procedures to select the best external route for each destination. It also describes how above benefits are

realized with best external route announcement with the help of certain scenarios.

1.1. Requirements Language

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 [RFC2119].

2. Algorithm for selection of best external route

Given that the intent in advertising an external route, when the best route for the same destination is an internal route, is to provide additional information into the IBGP mesh into which a route is participating, it is desirable to take into account the routes received from internal neighbors in the selection process.

We propose a route selection algorithm that selects a total order between routes and which selects the same best route as the one currently specified [RFC4271].

In order to achieve this, we need to introduce the concept of route group. For a given NLRI, suppose the BGP decision process has run through all the steps prior to the MED comparison step (as defined in section 9.1.2.2 of [RFC4271]. Look at the set of routes that are still under consideration at that time. Now partition this set into a number of disjoint route groups, where two routes are in the same group if and only if the neighbor AS of each route is the same.

Routes are ordered within a group via MED or subsequent route selection rules.

The order of all routes for the same destination is determined by the order of the best route in each group.

As an example, the following set of received routes:

Path	AS	MED	rtr_id
a	1	10	10
b	2	5	1
c	1	5	5
d	2	20	20
e	2	30	30
f	3	10	20

Figure 1: Path Attribute Table

Would yield the following order (from the most to the least preferred):

b < d < e < c < a < f

In this example, comparison of the best route within each group provides the sequence (b < c < f). The remaining routes are ordered in relation to their respective group best.

The first route in the above ordering is indeed the best route for a given destination. Eliminating the best route and executing the above steps leads us to a new total order of the routes. The route to be advertised to a particular domain is selected by choosing the most preferred route that is external to that particular domain in the above order. Note that whenever the overall best route is external it will automatically be selected by this algorithm.

3. Advertisement Rules

1. In an AS domain, if a router has installed an internal route as best, it should advertise its "best external route" (as defined in the draft) to its internal neighbors.

2. In a Cluster domain, if a router (route reflector) has installed an external route as best, it should advertise its "best internal route" to its external neighbors. (Advertising to internal neighbors is unchanged.) Similarly, if the route reflector has installed an internal route as best, it should advertise its "best external route" to its internal (client) peers. In order for the reflector to be able to advertise the best external route into the cluster, it is necessary that client-to-client reflection be disabled, since its advertisement may otherwise contain the best route within the cluster domain.

3. In a Confederation Member domain, if a router (confederation border router) has installed an internal route as best, it advertises its best external route to its internal neighbors. However, if it has installed an external route as best, it advertises its best internal route to its external neighbors.

4. Consistency between routing and forwarding

The BGP protocol, as defined in [RFC1771], specifies that a BGP speaker shall advertise to its internal peers the route with the

highest degree of preference among routes to the same destination received from external neighbors.

This section discusses problems present with the approach described in [RFC1771] and the next section offers an alternative algorithm to select a best external route which can be advertised to an IBGP mesh.

The internal update advertisement rules contained in the original BGP-4 specification [RFC1771] can lead to situations where traffic is forwarded through a route other than the route advertised by BGP.

Inconsistencies between forwarding and routing are highly undesirable. Service providers use BGP with the dual objective of learning reachability information and expressing policy over network resources. The latter assumes that forwarding follows routing information.

Consider the Autonomous system presented in figure 1, where r1 ... r4 are members of a single IBGP mesh and routes a, b, and c are received from external peers.

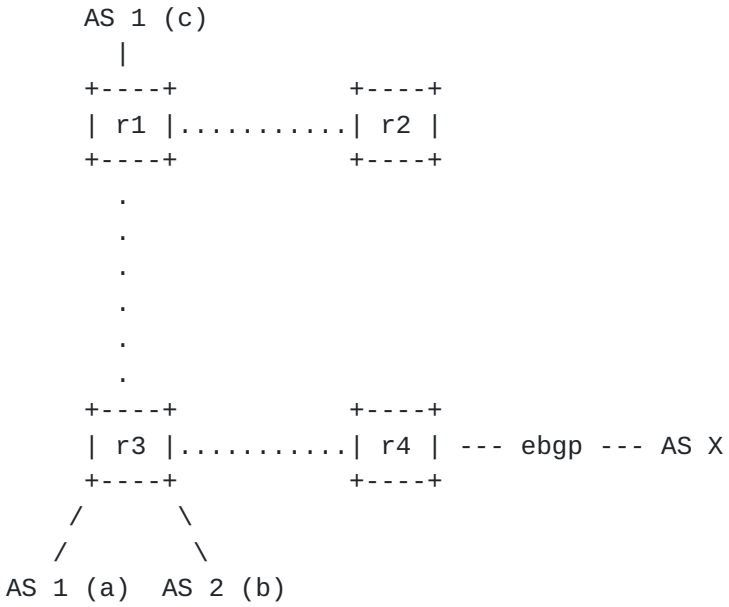

Figure 2: Inconsistency in Routing

Path	AS	MED	rtr_id
a	1	10	1
b	2	5	10
c	1	5	5

Figure 3: Path Attribute Table - 2

Following the rules as specified in [RFC1771], router r3 will select path (b) received from AS 2 as its overall best to install in the Loc-Rib, since path (b) is preferable to path (c), the lowest MED route from AS 1. However for the purposes of Internal Update route selection, it will ignore the presence of path (c), and elect (a) as its advertisement, via the router-id tie-breaking rule.

In this scenario, router r4 will receive (c) from r1 and (a) from r3. It will pick the lowest MED route (c) and advertise it out via ebgp to AS X. However at this point routing is inconsistent with forwarding as traffic received from AS X will be forwarded towards AS 2, while the ebgp advertisement is being made for an AS 1 path.

Routing policies are typically specified in terms of neighboring ASes. In the situation above, assuming that AS 1 is network for which this AS provides transit services while AS 2 and AS X are peer networks, one can easily see how the inconsistency between routing and forwarding would lead to transit being inadvertently provided between AS X and AS 2. This could lead to persistent forwarding loops.

Inconsistency between routing and forwarding may happen, whenever a bgp speaker chooses to advertise an external route into IBGP that is different from the overall best route and its overall best is external.

5. Applications

5.1. Fast Connectivity Restoration

When two exits are available to reach a particular destination and one is preferred over the other, the availability of an alternate path provides fast connectivity restoration when the primary path fails.

Restoration can be quick since the alternate path is already at hand. The border router could precompute the backup route and preinstall it in FIB ready to be switched when the primary goes away. Note that this requires the border router that's the backup to also preinstall the secondary path and switch to it on failure.

5.2. Inter-Domain Churn Reduction

Within an AS, the non availability of backup best leads to a border router sending a withdraw upstream when the primary fails. This leads to inter-domain churn and packet loss for the time the network takes to converge to the alternate path. Having the alternate path will reduces the churn and eliminates packet loss.

5.3. Reducing Persistent IBGP oscillation

Advertising the best external route, according to the algorithm described in this document will reduce the possibility of route oscillation by introducing additional information into the IBGP system.

For a permanent oscillation condition to occur, it is necessary that a circular dependency between paths occurs such that the selection of a new best path by a router, in response to a received IBGP advertisement, causes the withdrawal of information that another router depends on in order to generate the original event.

In vanilla BGP, when only the best overall route is advertised, as in most implementations, oscillation can occur whenever there are 2 or clusters/sub-ASes such that at least one cluster has more than one path that can potentially contribute to the dependency.

6. Acknowledgments

This document greatly benefits from the comments of Yakov Rekhter, John Scudder, Eric Rosen, and Jenny Yuan.

7. IANA Considerations

This document has no actions for IANA.

8. Security Considerations

There are no additional security risks introduced by this design.

9. 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.

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

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007.

Authors' Addresses

Pedro Marques
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Phone:
Email: roque@juniper.net

Rex Fernando
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Phone:
Email: rex@juniper.net

Enke Chen
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Phone:
Email: enkechen@cisco.com

Pradosh Mohapatra
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Phone:
Email: pmohapat@cisco.com