Network Working Group P. Marques
Internet-Draft
Expires: July 6, 2012 R. Fernando
E. Chen
P. Mohapatra
Cisco Systems
H. Gredler
Juniper Networks
January 3, 2012

Advertisement of the best external route in BGP
draft-ietf-idr-best-external-05

Abstract

The current BGP-4 protocol specification [RFC4271] states that the selection process chooses the best path for a given route which is added to the Loc-Rib and advertised to all peers.

Previous versions [RFC1771] of the specification defined a different rule for Internal BGP Updates. Given that Internal paths are not re-advertised to Internal peers, it was specified that the best of the external paths, as determined by the path selection tie breaking algorithm, would be advertised to Internal peers.

This document extends that procedure to operate in environments where Route Reflection [RFC4456] or Confederations [RFC5065] are used and explains why advertising the additional routing information can improve convergence time without causing routing loops.

Additional benefits include reduction of inter-domain churn and avoidance of permanent route oscillation.

This Internet-Draft will expire on July 6, 2012.

Copyright Notice

Table of Contents

1. Introduction

Earlier versions of the BGP-4 protocol specification [RFC1771] prescribed different route advertisement rules for Internal and External peers. While the overall best path would be advertised to External peers, Internal peers are advertised the best of the externally received paths.

This Internal advertisement rule was never implemented as specified and was latter dropped from the protocol. There is a trade-off in advertising the "best-external" route versus the behavior that became common standard of not advertising the route when the selected best path is received from an Internal peer. By not advertising information in this case it is possible to reduce state both in the local BGP speaker as well as in the network overall. Early BGP implementations where very concerned with reducing state as they where limited to relatively low memory footprints (e.g. 16 MB). There is also the possible concern regarding advertising a path different than the path that has been selected for forwarding.

However, 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:

- o 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.

- o Reducing inter-domain churn and traffic black-holing due to the readily available alternate path.

- o Reducing the potential for situations of permanent IBGP route oscillation [RFC3345].

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

This document defines procedures to select the best external route for each peer. It also describes how above benefits are realized with best external route announcement with the help of certain scenarios.

2. 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].

3. Generalization

The BGP-4 protocol [RFC1771] has extended with two alternative mechanisms that provide ways to reduce the operational complexity of route distribution within an AS: Route Reflection [RFC4456] and Confederations [RFC5065]. It is important to be able to express route advertisement rules in the context of both of these mechanisms.

When Route Reflection is used, Internal peers are further classified depending of the reflection cluster they belong to. Non-client internal peers form one BGP peering mesh. Each set of RR clients with the same "cluster-id" configuration forms a separate mesh.

When selecting the path to add to the Adj-RIB-OUT, this document specifies that the path that originate from the same mesh MAY be excluded from consideration. This results in an Adj-RIB-OUT selection per mesh (the set of non-client peers or a specific cluster).

Similarly, when BGP Confederations are used, each confederated AS is a BGP mesh. As with the Route Reflection scenario, when selecting the path to add to the Adj-RIB-OUT, routes from the same mesh MAY be excluded.

4. Algorithm for selection of the Adj-RIB-OUT path

The objective of this protocol extension is to improve the quality of the routing information known to a particular BGP mesh with minimum additional cost in terms of processing and state.

Towards that goal, it is useful to define a total order between the Adj-RIB-In routes which provides both the same overall best path as the algorithm defined in the current BGP-4 specification [RFC4271] as well as an ordering of alternate routes. Using this total order it is then computationally efficient to select the path for a specific Adj-RIB-OUT by excluding the routes that have been received from the BGP mesh corresponding to the peer (or set of peers).

In order to achieve this, it is helpful to introduce the concept of path group. A group is the set of paths that compare as equal through all the steps prior to the MED comparison step (as defined in section 9.1.2.2 of RFC 4271 [RFC4271] and have been received from the same neighbor AS.

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

In pseudo-code:

```
function compare(path_1, path_2) {
    cmp_result cmp = selection_steps_before_med(path_1, path_2);
    if (cmp != cmp_result.equal) {
        return cmp;
    }
    if (neighbor_as(path_1) == neighbor_as(path_2)) {
        return selection_steps_after_med(path_1, path_2);
    }

    if (is_group_best(path_1)) {
        if (!is_group_best(path_2)) {
            return cmp_result.greater_than;
        }
        return selection_steps_after_med(path_1, path_2);
    } else {
        if (is_group_best(path_2)) {
            return cmp_result.less_than;
        }
        /* Compare the best paths of respective groups */
        return compare(group_best(path_1), group_best(path_2));
    }
}
```

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

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 path within each group provides the sequence (b < c < f). The remaining paths are ordered in relation to their respective group best.

The first path in the ordering above is the best overall path for a given NLRI. When selecting a path for a particular Adj-RIB-Out (or set of RIB-Outs) an implementation MAY choose to select the first path in the global order which was not received from the same BGP mesh (as defined above) as the target peer (or peers).

5. Advertisement Rules

1. When advertising a route to a non-client Internal peer, a BGP speaker MAY choose to select the first path in order that did not originate from the same BGP mesh (i.e. the set of non-client Internal peers) whenever the best overall path has been received from this mesh and would be suppressed by the Internal BGP non-readvertisement rule.

2. When advertising a route to a Route Reflection client peer, in case the overall best path has been received from the same cluster, a BGP speaker MUST be able to advertise the best overall path to all the members of the cluster other than the originator, unless "client-to-client" reflection is disabled. The implementation MAY choose to advertise an alternate path to the specific peer that originates the best overall path by excluding from consideration all paths with the same originator-id.

3. When "client-to-client" reflection is disabled and the cluster is operating as a mesh, a Route Reflector MAY opt to advertise to the cluster the preferred path from the set of paths not received from the cluster. While this deployment mode is currently uncommon, it can be a practical way to guarantee path diversity inside the cluster.

4. A confederation border route MAY choose to advertise an alternate path towards its Internal BGP mesh or towards a con-fed member AS following the same procedure as defined above.

6. Consistency between routing and forwarding

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 reach-ability 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.

```
                AS 1 (c)
                  |
                +----+          +----+
                | r1 |..........| r2 |
                +----+          +----+
                  .
                  .
                  .
                  .
                  .
                  .
                +----+          +----+
                | r3 |..........| r4 | --- ebgp --- AS X
                +----+          +----+
               /      \
              /        \
           AS 1 (a)  AS 2 (b)
```

Inconsistency in Routing

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

Path Attribute Table

Following the rules as specified in RFC 1771 [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 IBGP 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 IBGP advertisement is being made for an AS 1 path.

Routing policies are typically specified in terms of neighboring AS-es. 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 GP speaker chooses to advertise an external route into IBGP that is different from the overall best route and its overall best is external.

7. Applications

8. 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 recompute the backup route and perinatal 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 perinatal the secondary path and switch to it on failure.

9. 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.

10. 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-AS-es such that at least one cluster has more than one path that can potentially contribute to the dependency.

11. Deployment Considerations

The mechanism specified in the draft allows a BGP speaker to advertise a route that is not the best route used for forwarding. This is a departure from the current behavior. However, consistency in the path selection process across the AS is still guaranteed since the ingress routers will not choose the best-external route as the best route for a destination in steady state (for the same reason that the BGP speaker announcing the best-external route chose an IBGP route as best instead of the externally learnt route). Though it is possible to alter this assurance by defining route policies on IBGP sessions, use of such policies in IBGP is not recommended, especially with best-external announcement turned on in the network. It is also worth noting that such inconsistency in routing and forwarding is mitigated in a tunneled network.

12. Acknowledgments

This document greatly benefits from the comments of Yakov Rekhter, John Scudder, Eric Rosen, Jenny Yuan, Jay Borkenhagen, Salkat Ray and Jakob Heitz.

13. IANA Considerations

This document has no actions for IANA.

14. Security Considerations

There are no additional security risks introduced by this design.

15. 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

Email: pedro.r.marques@gmail.com

Rex Fernando
Cisco Systems
170 W. Tasman Dr.
San Jose, CA 95134
US

Email: rex@cisco.com

Enke Chen
Cisco Systems
170 W. Tasman Dr.
San Jose, CA 95134
US

Email: enkechen@cisco.com

Pradosh Mohapatra
Cisco Systems
170 W. Tasman Dr.
San Jose, CA 95134
US

Email: pmohapat@cisco.com

Hannes Gredler
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: hannes@juniper.net