INTERNET-DRAFT
<draft-ietf-idr-route-reflect-v2-00.txt>

Tony Bates
Ravi Chandra
Enke Chen
Cisco Systems
November 1998

BGP Route Reflection
An alternative to full mesh IBGP
<draft-ietf-idr-route-reflect-v2-00.txt>

Status of this Memo

Abstract

The Border Gateway Protocol [1] is an inter-autonomous system routing protocol designed for TCP/IP internets. Currently in the Internet BGP deployments are configured such that that all BGP speakers within a single AS must be fully meshed so that any external routing information must be re-distributed to all other routers within that AS. This represents a serious scaling problem that has been well documented with several alternatives proposed [2,3].

This document describes the use and design of a method known as 'Route Reflection' to alleviate the the need for 'full mesh' IBGP.

1. Introduction

Currently in the Internet, BGP deployments are configured such that that all BGP speakers within a single AS must be fully meshed and any external routing information must be re-distributed to all other routers within that AS. For n BGP speakers within an AS that requires to maintain n*(n-1)/2 unique IBGP sessions. This "full mesh" requirement clearly does not scale when there are a large number of IBGP speakers each exchanging a large volume of routing information, as is common in many of todays internet networks.

This scaling problem has been well documented and a number of proposals have been made to alleviate this [2,3]. This document represents another alternative in alleviating the need for a "full mesh" and is known as "Route Reflection". This approach allows a BGP speaker (known as "Route Reflector") to advertise IBGP learned routes to certain IBGP peers. It represents a change in the commonly understood concept of IBGP, and the addition of two new optional transitive BGP attributes to prevent loops in routing updates.

2. Design Criteria

Route Reflection was designed to satisfy the following criteria.

- Simplicity

  Any alternative must be both simple to configure as well as understand.

- Easy Transition

  It must be possible to transition from a full mesh configuration without the need to change either topology or AS. This is an unfortunate management overhead of the technique proposed in [3].

- Compatibility

  It must be possible for non compliant IBGP peers to continue be part of the original AS or domain without any loss of BGP routing information.

These criteria were motivated by operational experiences of a very large and topology rich network with many external connections.

3. Route Reflection

The basic idea of Route Reflection is very simple. Let us consider the simple example depicted in Figure 1 below.

```
+------ +        +-------+
|       |  IBGP  |       |
| RTR-A |--------| RTR-B |
|       |        |       |
+-------+        +-------+
      \            /
  IBGP \  ASX     / IBGP
        \        /
         +-------+
         |       |
         | RTR-C |
         |       |
         +-------+
```

Figure 1: Full Mesh IBGP

In ASX there are three IBGP speakers (routers RTR-A, RTR-B and RTR-C). With the existing BGP model, if RTR-A receives an external route and it is selected as the best path it must advertise the external route to both RTR-B and RTR-C. RTR-B and RTR-C (as IBGP speakers) will not re-advertise these IBGP learned routes to other IBGP speakers.

If this rule is relaxed and RTR-C is allowed to advertise IBGP learned routes to IBGP peers, then it could re-advertise (or reflect) the IBGP routes learned from RTR-A to RTR-B and vice versa. This would eliminate the need for the IBGP session between RTR-A and RTR-B as shown in Figure 2 below.

```
+------ +        +-------+
|       |        |       |
| RTR-A |        | RTR-B |
|       |        |       |
+-------+        +-------+
      \            /
  IBGP \  ASX     / IBGP
        \        /
         +-------+
         |       |
         | RTR-C |
         |       |
         +-------+
```

Figure 2: Route Reflection IBGP

The Route Reflection scheme is based upon this basic principle.

4. Terminology and Concepts

We use the term "Route Reflection" to describe the operation of a BGP speaker advertising an IBGP learned route to another IBGP peer. Such a BGP speaker is said to be a "Route Reflector" (RR), and such a route is said to be a reflected route.

The internal peers of a RR are divided into two groups:

1) Client Peers

2) Non-Client Peers

A RR reflects routes between these groups, and may reflect routes among client peers. A RR along with its client peers form a Cluster. The Non-Client peer must be fully meshed but the Client peers need not be fully meshed. Figure 3 depicts a simple example outlining the basic RR components using the terminology noted above.

```
                      / - - - - - - - - - - - - -  -
                      |           Cluster           |
                        +-------+        +-------+
                      | |       |        |       |  |
                        | RTR-A |        | RTR-B |
                      | |Client |        |Client |  |
                        +-------+        +-------+
                      |      \             /        |
                         IBGP \          / IBGP
                      |        \        /           |
                                +-------+
                      |         |       |           |
                                | RTR-C |
                      |         |  RR   |           |
                                +-------+
                      |           /   \             |
                       - - - - - /- - -\- - - - - - /
                          IBGP  /       \ IBGP
                       +-------+         +-------+
                       | RTR-D |  IBGP   | RTR-E |
                       |  Non- |---------|  Non- |
                       |Client |         |Client |
                       +-------+         +-------+
```

Figure 3: RR Components

5. Operation

When a RR receives a route from an IBGP peer, it selects the best path based on its path selection rule. After the best path is selected, it must do the following depending on the type of the peer it is receiving the best path from:

1) A Route from a Non-Client IBGP peer

   Reflect to all the Clients.

2) A Route from a Client peer

   Reflect to all the Non-Client peers and also to the Client peers. (Hence the Client peers are not required to be fully meshed.)

An Autonomous System could have many RRs. A RR treats other RRs just like any other internal BGP speakers. A RR could be configured to have other RRs in a Client group or Non-client group.

In a simple configuration the backbone could be divided into many clusters. Each RR would be configured with other RRs as Non-Client peers (thus all the RRs will be fully meshed.). The Clients will be configured to maintain IBGP session only with the RR in their cluster. Due to route reflection, all the IBGP speakers will receive reflected routing information.

It is possible in a Autonomous System to have BGP speakers that do not understand the concept of Route-Reflectors (let us call them conventional BGP speakers). The Route-Reflector Scheme allows such conventional BGP speakers to co-exist. Conventional BGP speakers could be either members of a Non-Client group or a Client group. This allows for an easy and gradual migration from the current IBGP model to the Route Reflection model. One could start creating clusters by configuring a single router as the designated RR and configuring other RRs and their clients as normal IBGP peers. Additional clusters can be created gradually.

6. Redundant RRs

Usually a cluster of clients will have a single RR. In that case, the cluster will be identified by the ROUTER_ID of the RR. However, this represents a single point of failure so to make it possible to have multiple RRs in the same cluster, all RRs in the same cluster can be configured with a 4-byte CLUSTER_ID so that an RR can discard routes

from other RRs in the same cluster.

7. Avoiding Routing Information Loops

When a route is reflected, it is possible through mis-configuration to form route re-distribution loops. The Route Reflection method defines the following attributes to detect and avoid routing information loops:

ORIGINATOR_ID

ORIGINATOR_ID is a new optional, non-transitive BGP attribute of Type code 9. This attribute is 4 bytes long and it will be created by a RR in reflecting a route. This attribute will carry the ROUTER_ID of the originator of the route in the local AS. A BGP speaker should not create an ORIGINATOR_ID attribute if one already exists. A router should ignore a route received with its ROUTER_ID as the ORIGINATOR_ID.

CLUSTER_LIST

Cluster-list is a new optional, non-transitive BGP attribute of Type code 10. It is a sequence of CLUSTER_ID values representing the reflection path that the route has passed. It is encoded as follows:

```
           0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |  Attr. Flags  |Attr. Type Code|   Length      | value ...
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Where Length is the number of octets.

When a RR reflects a route, it must append the local CLUSTER_ID to the CLUSTER_LIST. If the CLUSTER_LIST is empty, it must create a new one. Using this attribute an RR can identify if the routing information is looped back to the same cluster due to mis-configuration. If the local CLUSTER_ID is found in the cluster-list, the advertisement received will be ignored.

8. Implementation Considerations

Care should be taken to make sure that none of the BGP path attributes defined above can be modified through configuration when exchanging internal routing information between RRs and Clients and

Non-Clients. Their modification could potential result in routing loops.

In addition, when a RR reflects a route, it should not modify the following path attributes: NEXT_HOP, AS_PATH, LOCAL_PREF, and MED. Their modification could potential result in routing loops.

9. Configuration and Deployment Considerations

The BGP protocol provides no way for a Client to identify itself dynamically as a Client of an RR. The simplest way to achieve this is by manual configuration.

One of the key component of the route reflection approach in addressing the scaling issue is that the RR summarizes routing information and only reflects its best path.

Both MEDs and IGP metrics may impact the BGP route selection. Because MEDs are not always comparable and the IGP metric may differ for each router, with certain route reflection topologies the route reflection approach may not yield the same route selection result as that of the full IBGP mesh approach. A way to make route selection the same as it would be with the full IBGP mesh approach is to make sure that route reflectors are never forced to perform the BGP route selection based on IGP metrics which are significantly different from the IGP metrics of their clients, or based on incomparable MEDs. The former can be achieved by configuring the intra-cluster IGP metrics to be better than the inter-cluster IGP metrics, and maintaining full mesh within the cluster. The latter can be achieved by:

- o setting the local preference of a route at the border router to reflect the MED values.

- o or by making sure the AS-path lengths from different ASs are different when the AS-path length is used as a route selection criteria.

- o or by configuring community based policies using which the reflector can decide on the best route.

One could argue though that the latter requirement is overly restrictive, and perhaps impractical in some cases. One could further argue that as long as there are no routing loops, there are no compelling reasons to force route selection with route reflectors to be the same as it would be with the full IBGP mesh approach.

To prevent routing loops and maintain consistent routing view, it is

essential that the network topology be carefully considered in designing a route reflection topology. In general, the route reflection topology should congruent with the network topology when there exist multiple paths for a prefix. One commonly used approach is the POP-based reflection, in which each POP maintains its own route reflectors serving clients in the POP, and all route reflectors are fully meshed. In addition, clients of the reflectors in each POP are often fully meshed for the purpose of optimal intra-POP routing, and the intra-POP IGP metrics are configured to be better than the inter-POP IGP metrics.

10. Security

Security considerations are not discussed in this memo.

11. Acknowledgments

The authors would like to thank Dennis Ferguson, John Scudder, Paul Traina and Tony Li for the many discussions resulting in this work. This idea was developed from an earlier discussion between Tony Li and Dimitri Haskin.

In addition, the authors would like to acknowledge valuable review and suggestions from Yakov Rekhter on this document, and helpful comments from Tony Li, Rohit Dube, and John Scudder on Section 9.

12. References

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

[2] Haskin, D., "A BGP/IDRP Route Server alternative to a full mesh routing", RFC1863, October 1995.

[3] Traina, P. "Limited Autonomous System Confederations for BGP", RFC1965, June 1996.

13. Author's Addresses

Tony Bates
Cisco Systems
170 West Tasman Drive

email: tbates@cisco.com

Ravishanker Chandrasekeran
(Ravi Chandra)
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134

email: rchandra@cisco.com

Enke Chen
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134

email: enkechen@cisco.com