                        BGP Route Reflection
                    An alternative to full mesh IBGP

Status of this Memo

Abstract

   The Border Gateway Protocol [1] is an inter-autonomous system routing
   protocol designed for TCP/IP internets. 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. This "full mesh" requirement clearly does not
   scale when there are a large number of IBGP speakers as is common in
   many of todays internet networks.

   For n BGP speakers within an AS you must maintain $n*(n-1)/2$ unique
   IBGP sessions. With finite resources in both bandwidth and router CPU
   this clearly does not scale.

   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". It represents a change in
   the commonly understood concept of IBGP and the addition of two new

   optional transitive BGP attributes.

2. Design Criteria

   Route Reflection was designed to satisfy the following criteria.

           o Simplicity

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

           o Easy Migration

             It must be possible to migrate 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].

           o 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 |
                     |       |
                     +-------+
```
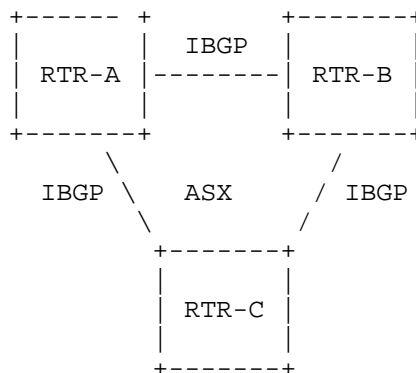
                    Figure 1: Full Mesh IBGP

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 reflect IBGP learned routes, 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 |
                  |       |
                  +-------+
```
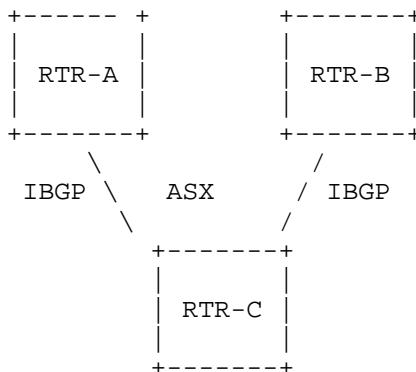
Figure 2: Route Reflection IBGP

The Route Reflection scheme is based upon this basic principle.

4. Terminology and Concepts

We use the term "Route Reflector" (RR) to represent an IBGP speaker that participates in the reflection. 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. 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. The Client peers should not peer with internal speakers outside of their cluster. 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 |
               +-------+        +-------+
```
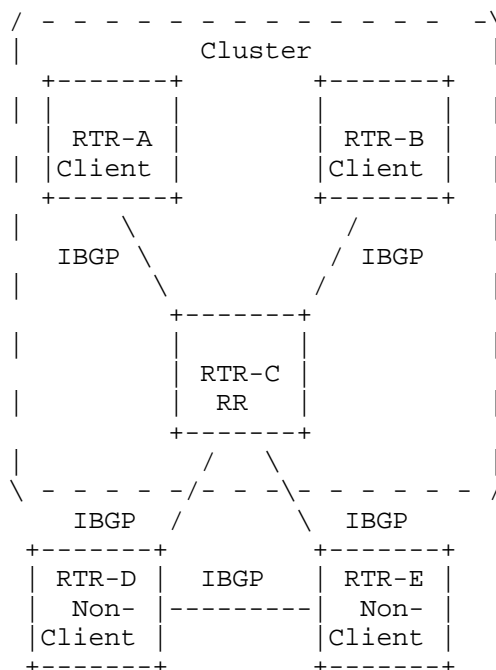
                     Figure 3: RR Components

5. Operation

   When a route is received by a RR, 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 peer

              Reflect to all other Clients.

           2) A Route from a Client peer

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

            3) Route from an EBGP peer

               Send to all the Client and Non-Client Peers.

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 normal 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 ould
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 must be
configured with a 4-byte CLUSTER_ID so that an RR can discern routes
from other RRs in the same cluster.
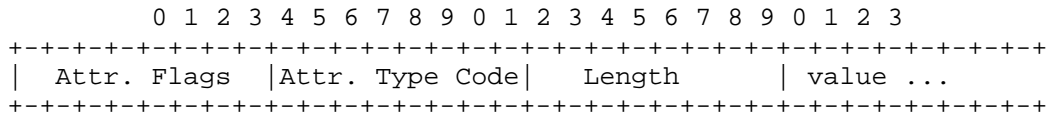
7.   Avoiding Routing Information Loops

As IBGP learned routes are 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. 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 route reflector
must never send routing information back to the router specified in
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 from its Clients to a Non-Client peer, 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 will be ignored.

8. Implementation and Configuration 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. This could result is looping of routes.

In some implementations, modification of the BGP path attribute,
NEXT_HOP is possible. For example, there could be a need for a RR to
modify NEXT_HOP for EBGP learned routes sent to its internal peers.
However, it must not be possible for an RR to set on reflected IBGP
routes as this breaks the basic principle of Route Reflection and
will result in potential black holeing of traffic.

An RR should not modify any AS-PATH attributes (i.e. LOCAL_PREF, MED,
DPA)that could change consistent route selection. This could result
in potential loops.

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

9. Security Considerations

Security issues are not discussed in this memo.

10. Acknowledgments

   The authors would like to thank Dennis Ferguson, Enke Chen, 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.

11. References

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

   [2] Haskin, D., "A BGP/IDRP Route Server alternative to a full mesh
        routing", RFC 1863, October 1995.

   [3] Traina, P., "Limited Autonomous System Confederations for BGP",
        RFC 1965, June 1996.

12. Authors' Addresses

   Tony Bates
   cisco Systems
   170 West Tasman Drive
   San Jose, CA 95134

   Phone: +1 408 527 2470
   EMail: tbates@cisco.com

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

   EMail: rchandra@cisco.com