                      BGP Route Reflection:
         An Alternative to Full Mesh Internal BGP (IBGP)

Status of This Memo

   This document specifies an Internet standards track protocol for the
   Internet community, and requests discussion and suggestions for
   improvements. Please refer to the current edition of the "Internet
   Official Protocol Standards" (STD 1) for the standardization state
   and status of this protocol. Distribution of this memo is unlimited.

Copyright Notice

Abstract

   The Border Gateway Protocol (BGP) is an inter-autonomous system
   routing protocol designed for TCP/IP internets. Typically, 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 Autonomous System (AS). This represents a serious
   scaling problem that has been well documented with several
   alternatives proposed.

   This document describes the use and design of a method known as
   "route reflection" to alleviate the need for "full mesh" Internal BGP
   (IBGP).

   This document obsoletes RFC 2796 and RFC 1966.

Table of Contents

1. Introduction

   Typically, 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 Internal BGP (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 today's 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 a "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
   non-transitive BGP attributes to prevent loops in routing updates.

   This document obsoletes RFC 2796 [6] and RFC 1966 [4].

2. Specification of Requirements

   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 [7].

[3](#). **Design Criteria**

   Route reflection was designed to satisfy the following criteria.

      o Simplicity

         Any alternative must be simple to configure and easy to
         understand.

      o 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](#)].

      o Compatibility

         It must be possible for noncompliant IBGP peers to continue to
         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.

[4](#). **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 |
                   |       |
                   +-------+
```
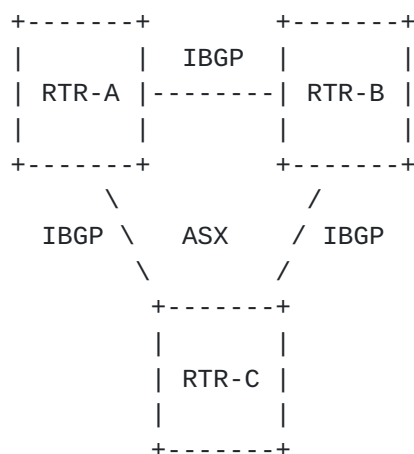
                 Figure 1: Full-Mesh IBGP

   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 |
                     |       |
                     +-------+
```
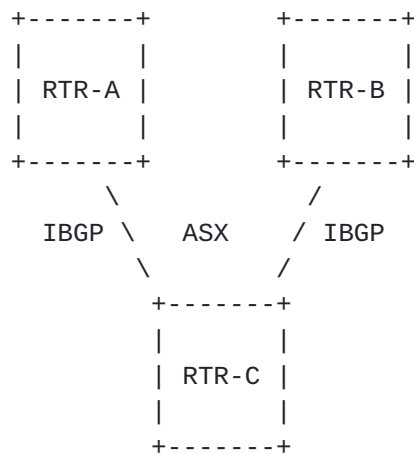
                 Figure 2: Route Reflection IBGP

   The route reflection scheme is based upon this basic principle.

5. 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 an RR are divided into two groups:

      1) Client peers

      2) Non-Client peers

   An RR reflects routes between these groups, and may reflect routes
   among client peers. An 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
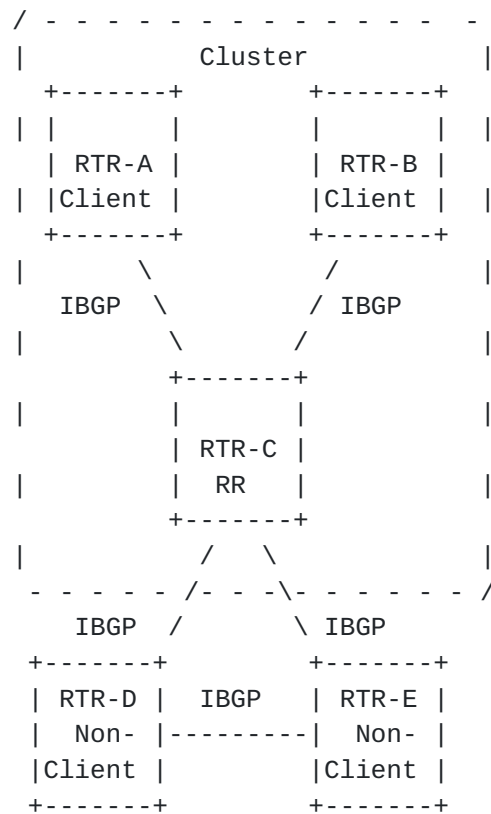   outlining the basic RR components using the terminology noted above.

```
            / - - - - - - - - - - - - - -  -
            |              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

[6](#). Operation

   When an 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 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. An RR treats other RRs
   just like any other internal BGP speakers. An 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 an 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 coexist. Conventional BGP speakers
could be members of either 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.

[7](#). Redundant RRs

Usually, a cluster of clients will have a single RR. In that case,
the cluster will be identified by the BGP Identifier 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.

[8](#). Avoiding Routing Information Loops

When a route is reflected, it is possible through misconfiguration 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 an
RR in reflecting a route. This attribute will carry the BGP
Identifier 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 that recognizes the ORIGINATOR_ID attribute SHOULD
ignore a route received with its BGP Identifier 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.

   When an RR reflects a route, it MUST prepend 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 has looped back to the same cluster due to
   misconfiguration. If the local CLUSTER_ID is found in the
   CLUSTER_LIST, the advertisement received SHOULD be ignored.

[9](#). **Impact on Route Selection**

   The BGP Decision Process Tie Breaking rules (Sect. 9.1.2.2, [[1](#)]) are
   modified as follows:

      If a route carries the ORIGINATOR_ID attribute, then in Step f)
      the ORIGINATOR_ID SHOULD be treated as the BGP Identifier of the
      BGP speaker that has advertised the route.

      In addition, the following rule SHOULD be inserted between Steps
      f) and g): a BGP Speaker SHOULD prefer a route with the shorter
      CLUSTER_LIST length. The CLUSTER_LIST length is zero if a route
      does not carry the CLUSTER_LIST attribute.

[10](#). **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 potentially 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 potentially result in routing loops.

[11](#). **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 Multi-Exit Discriminators (MEDs) and Interior Gateway Protocol
(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 that 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, or

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

   o configuring community-based policies to influence the route
      selection.

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 be congruent with the network topology
when there exist multiple paths for a prefix. One commonly used
approach is the reflection based on Point of Presence (POP), 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.

[12](#). Security Considerations

This extension to BGP does not change the underlying security issues
inherent in the existing IBGP [[1](#), [5](#)].

13. Acknowledgements

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, John Scudder, and Bruce Cole.

14. References

14.1. Normative References

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

14.2. Informative References

[2] Savola, P., "Reclassification of RFC 1863 to Historic", RFC
     4223, October 2005.

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

[4] Bates, T. and R. Chandra, "BGP Route Reflection An alternative
     to full mesh IBGP", RFC 1966, June 1996.

[5] Heffernan, A., "Protection of BGP Sessions via the TCP MD5
     Signature Option", RFC 2385, August 1998.

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

[7] Bradner, S., "Key words for use in RFCs to Indicate Requirement
     Levels", BCP 14, RFC 2119, March 1997.

Appendix A: Comparison with RFC 2796

   The impact on route selection is added.

   The pictorial description of the encoding of the CLUSTER_LIST
   attribute is removed as the description is redundant to the BGP
   specification, and the attribute length field is inadvertently
   described as one octet.

Appendix B: Comparison with RFC 1966

   All the changes listed in Appendix A, plus the following.

   Several terminologies related to route reflection are clarified, and
   the reference to EBGP routes/peers are removed.

   The handling of a routing information loop (due to route reflection)
   by a receiver is clarified and made more consistent.

   The addition of a CLUSTER_ID to the CLUSTER_LIST has been changed
   from "append" to "prepend" to reflect the deployed code.

   The section on "Configuration and Deployment Considerations" has been
   expanded to address several operational issues.

Authors' Addresses

    Tony Bates
    Cisco Systems, Inc.
    170 West Tasman Drive
    San Jose, CA 95134

    EMail: tbates@cisco.com

    Ravi Chandra
    Sonoa Systems, Inc.
    3255-7 Scott Blvd.
    Santa Clara, CA 95054

    EMail: rchandra@sonoasystems.com

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

    EMail: enkechen@cisco.com

Full Copyright Statement

Intellectual Property

Acknowledgement