       Avoid BGP Best Path Transitions from One External to Another

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.

Abstract

   In this document, we propose an extension to the BGP route selection
   rules that would avoid unnecessary best path transitions between
   external paths under certain conditions. The proposed extension
   would help the overall network stability, and more importantly, would
   eliminate certain BGP route oscillations in which more than one
   external path from one BGP speaker contributes to the churn.

1. Introduction

   The last two steps of the BGP route selection (Section 9.1.2.2,
   [BGP]) involve comparing the BGP identifiers and the peering
   addresses. The BGP identifier (treated either as an IP address or
   just an integer [BGP-ID]) for a BGP speaker is allocated by the
   Autonomous System (AS) to which the speaker belongs. As a result,
   for a local BGP speaker, the BGP identifier of a route received from
   an external peer is just a random number. When routes under
   consideration are from external peers, the result from the last two
   steps of the route selection is therefore "random" as far as the
   local BGP speaker is concerned.

   It is based on this observation that we propose an extension to the
   BGP route selection rules that would avoid unnecessary best-path
   transitions between external paths under certain conditions. The
   proposed extension would help the overall network stability, and more
   importantly, would eliminate certain BGP route oscillations in which
   more than one external path from one BGP speaker contributes to the
   churn.

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 [RFC2119].

3. The Algorithm

   Consider the case in which the existing best path A is from an
   external peer, and another external path B is then selected as the
   new best path by the route selection algorithm described in [BGP].
   When comparing all the paths in route selection, if neither Path A
   nor Path B is eliminated by the route selection algorithm prior to
   Step f) -- BGP identifier comparison (Section 9.1.2.2, [BGP]) -- we
   propose that the existing best path (Path A) be kept as the best path
   (thus avoiding switching the best path to Path B).

   This algorithm SHOULD NOT be applied when either path is from a BGP
   Confederation peer.

   In addition, the algorithm SHOULD NOT be applied when both paths are
   from peers with an identical BGP identifier (i.e., there exist
   parallel BGP sessions between two BGP speakers). As the peering
   addresses for the parallel sessions are typically allocated by one AS
   (possibly with route selection considerations), the algorithm (if
   applied) could impact the existing routing setup. Furthermore, by
   not applying the algorithm, the allocation of peering addresses would
   remain as a simple and effective tool in influencing route selection
   when parallel BGP sessions exist.

4. The Benefits

   The proposed extension to the BGP route selection rules avoids
   unnecessary best-path transitions between external paths under
   certain conditions. Clearly, the extension would help reduce routing
   and forwarding changes in a network, thus helping the overall network
   stability.

   More importantly, as shown in the following example, the proposed
   extension can be used to eliminate certain BGP route oscillations in
   which more than one external path from one BGP speaker contributes to
   the churn. Note however, that there are permanent BGP route
   oscillation scenarios [RFC3345] that the mechanism described in this
   document does not eliminate.

Consider the example in Figure 1 where

    o R1, R2, R3, and R4 belong to one AS.
    o R1 is a route reflector with R3 as its client.
    o R2 is a route reflector with R4 as its client.
    o The IGP metrics are as listed.
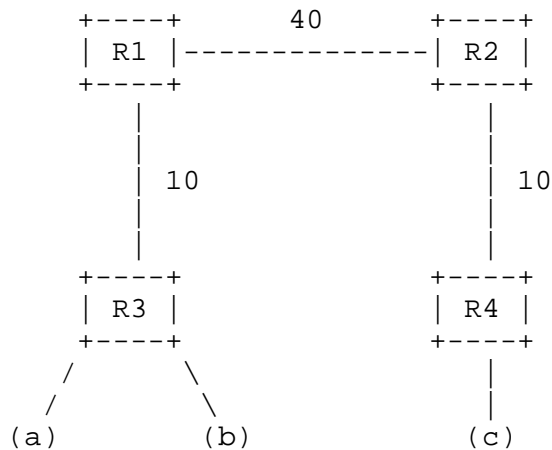   o External paths (a), (b), and (c) are as described in Figure 2.

```
          +----+        40       +----+
          | R1 |-------------| R2 |
          +----+              +----+
            |                    |
            |                    |
            | 10                 | 10
            |                    |
            |                    |
          +----+              +----+
          | R3 |              | R4 |
          +----+              +----+
         /       \               |
        /         \              |
      (a)        (b)           (c)
```

                      Figure 1

```
        Path    AS      MED    Identifier
         a       1        0         2
         b       2       20         1
         c       2       10         5
```

                      Figure 2

Due to the interaction of the route reflection [BGP-RR] and the
MULTI_EXIT_DISC (MED) attribute, the best path on R1 keeps churning
between (a) and (c), and the best path on R3 keeps churning between
(a) and (b).

With the proposed algorithm, R3 would not switch the best path from
(a) to (b) even after R1 withdraws (c) toward its clients, and that
is enough to stop the route oscillation.

Although this type of route oscillation can also be eliminated by
other route reflection enhancements being developed, the proposed
algorithm is extremely simple and can be implemented and deployed
immediately without introducing any backward compatibility issues.

5. Remarks

   The proposed algorithm is backward-compatible, and can be deployed on
   a per-BGP-speaker basis. The deployment of the algorithm is highly
   recommended on a BGP speaker with multiple external BGP peers
   (especially the ones connecting to an inter-exchange point).

   Compared to the existing behavior, the proposed algorithm may
   introduce some "non-determinism" in the BGP route selection --
   although one can argue that the BGP Identifier comparison in the
   existing route selection has already introduced some "randomness" as
   described in the introduction section. Such "non-determinism" has
   not been shown to be detrimental in practice and can be completely
   eliminated by using the existing mechanisms (such as setting
   LOCAL_PREF or MED) if so desired.

6. Security Considerations

   This extension does not introduce any security issues.

7. Acknowledgments

   The idea presented was inspired by a route oscillation case observed
   in the BBN/Genuity network in 1998. The algorithm was also
   implemented and deployed at that time.

   The authors would like to thank Yakov Rekhter and Ravi Chandra for
   their comments on the initial idea.

8. Normative References

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

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

   [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
              Requirement Levels", BCP 14, RFC 2119, March 1997.

9. Informative References

   [BGP-ID] Chen, E. and J. Yuan, "AS-wide Unique BGP Identifier for
             BGP-4", Work in Progress, November 2006.

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

Author Information

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

   EMail: enkechen@cisco.com

   Srihari R. Sangli
   Cisco Systems, Inc.
   170 W. Tasman Dr.
   San Jose, CA 95134

   EMail: rsrihari@cisco.com