

Internet Engineering Task Force
Internet Draft
Expiration Date: May 1999

Rohit Dube
Bell Labs, Lucent Technologies
John G. Scudder
Internet Engineering Group, LLC

Route Reflection Considered Harmful

[draft-dube-route-reflection-harmful-00.txt](#)

1. Status of this Memo

This document is an Internet-Draft. Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as ``work in progress.''

To view the entire list of current Internet-Drafts, please check the "lid-abstracts.txt" listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa), ftp.nordu.net (Northern Europe), ftp.nis.garr.it (Southern Europe), munnari.oz.au (Pacific Rim), ftp.ietf.org (US East Coast), or ftp.isi.edu (US West Coast).

2. Abstract

Route reflection as defined by [2] is a popular way of reducing the full-mesh IBGP peering required by routers running the Border Gateway Protocol [1]. There are cases where a topology built using route reflectors produces persistent loops or does not produce the same results as what one would expect with a full IBGP mesh. This document describes these problems.

3. Introduction

Route reflectors by design are selective as to which routes they forward to their peers (i.e. reflect). Specifically, if many routes to the same NLRI are available, a route reflector will reflect only the route it has selected for its own use. Typically this reduces the number of routes each peer in the AS must store in its RIB as well as the volume of BGP update traffic. By this very nature of route reflection, every peer in the network doesn't have a full view of all the routes to a prefix to choose from. This coupled with the

specifics of BGP causes problems as we now describe.

4. Persistent Loops

Consider the topology in Figure 1.

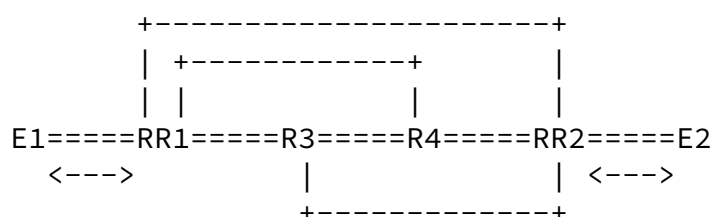


Figure 1

RR1, RR2, R3 and R4 are bgp routers in the same AS. E1 and E2 are BGP routers in some other AS peering with RR1 and RR2 respectively via EBG. RR1 is configured as a route reflector with R4 as a client and RR2 is configured as a reflector in a different cluster with R3 as a client. The IBGP sessions are denoted in the diagram above by +---+ and the EBG sessions by <--->. For simplicity, assume that all the physical links (denoted by ==) have the same IGP cost.

Now if both E1 and E2 advertise the same prefix to RR1 and RR2 respectively, all other things being equal, RR1 picks the route through E1 for this prefix on account of lower IGP cost. RR1 then reflects this route to R4 which now routes to the prefix in question through R3 and RR1. Similarly RR2 picks the route through E2 and reflects it to R3 which now routes to the prefix in question through R4 and RR2. Clearly a data packet for this prefix will loop between R3 and R4.

Note that the problem would disappear if the topology is reverted to full-mesh IBGP - R3 would pick the route through RR1 and R4 would pick the route through RR2, both on account of lower IGP cost.

5. Incorrect Routing Decision

Consider the topology in Figure 2.

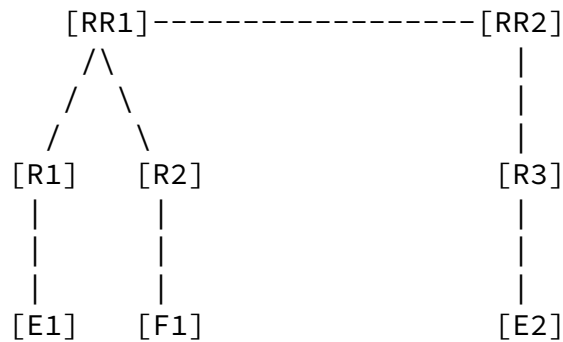


Figure 2

RR1, RR2, R1, R2, R3 are bgp routers in the same AS R. RR1 is a route reflector with clients R1 and R2 and RR2 is a route reflector in a different cluster with client R3. E1 and E2 are bgp routers in AS E and EBGP peer with R1 and R3 respectively. F1 is a bgp router in AS F which EBGP peers with R2. Assume that E1, E2 and F advertise the same prefix to R1, R2, R3 in accordance with the following table -

Router	AS	Router-id	MED

E1	E	3.3.3.3	50
F1	F	2.2.2.2	-
E2	E	1.1.1.1	100

All other attributes of the prefix in question are the same.

Further assume that RR1's IGP cost to R1 (and E1) is the same as its cost R2 (and F1) and RR2's IGP cost to R3 (and E2) is the same as its IGP cost to R1 (and E1) and R2 (F1). (The --- lines in Figure 2 denote both physical and BGP connectivity).

Now, RR1 chooses the route thru F1 on account of lower router-id as compared to the route through E1 (which wins over the route from E2 on account of MEDs). RR2 on the other hand chooses the route through E2 on account of lower router-id as compared to F. Note that RR1 sends only the route through F1 to RR2 and not the route through E1.

Instead if we had a full-mesh, RR2 would see all the 3 routes and pick the one thru F1 - the route through E1 wins over the route through E2 on MEDs and the route through F1 wins over the route

through E1 on account of lower router-id.

A network operator shifting from a topology without to reflectors to the one above with reflectors would have a problem. Packets destined for the prefix in question would flow from RR2 through E2 instead of the original F1.

6. Characterization

Problem 1 ([Section 4](#)) has two ingredients - a) the selective nature of route reflectors which prevents some routes from getting to some clients and b) The fact the some of the BGP decision process -- specifically the "prefer lowest IGP cost" rule -- depend on the router's location in the network. Thus the route reflector's decision can never perfectly mirror the decision its client would have made. Note that b) implies that reflector topologies can be out of sync with the physical topologies but bad things happen only when they get out of sync enough that clients would make decisions (in this case based on IGP cost) different from their servers if reflection was replaced by full-mesh.

Problem 2 ([Section 5](#)) has two components too - a) the selective nature of route reflectors as above and b) the partial order that MEDs impose upon competing routes (this is because MEDs can be compared only between routes from the same AS). If all decision criteria used by BGP imposed a total order on the routes (i.e all BGP routes for a prefix could be arranged in strict order of precedence), then b) would not be an issue and in-spite of a) this problem would not happen.

For both examples discussed, it is possible to come up with several other topologies which suffer from the problems described above.

7. Avoidance Guidelines

Since there are no protocol mechanisms currently available to detect the problems mentioned above, we provide guidelines to avoid situations where these problems could surface.

As noted in [section 6](#), problem 1 happens because the IBGP reflector topology doesn't follow the physical topology. A simple way of avoiding this problem would be to ensure that reflector clusters are constrained to follow the physical connectivity between the routers. It is always safe (at least with respect to this problem) to deploy route reflection such that no IBGP session between a pair of route reflectors will ever physically transit a reflector client. One common mode of deployment is to fully mesh all the routers in a "backbone" region, and to do route reflection to/from/between the routers in a POP, using one or more of the backbone routers as the reflector(s).

Problem 2 can be avoided by always making sure that reflectors are never forced to decide on the best BGP route based on MEDs. This can be achieved either by setting the local preference of a route at the border router to reflect the MED values or by configuring community based policies using which the reflector can decide on the best route.

[8](#). Acknowledgments

The First author would like to thank to Harry Mantakos, James Da Silva and Arvind Srivaths (all at Torrent Networking Technologies Corp.), Rob Coltun (Fore Systems) and Tony Przgyienda (Bell Labs, Lucent Technologies) for discussions on this topic. The second author would like to thank Ravi Chandra and Tony Bates (both at Cisco Systems) for similar discussions.

[9](#). References

- [1] Rekhter, Y., and Li, T., "A Border Gateway Protocol 4 (BGP-4)", [RFC 1771](#), March 1995.
- [2] Bates, T., and Chandra, R., "BGP Route Reflection An alternative to full mesh IBGP", [RFC 1966](#), June 1996.

10. Author Information

Rohit Dube
Bell Labs, Lucent Technologies Inc.
4C-508, 101 Crawfords Corner Road
Holmdel, NJ 07724
e-mail: rohitd@dnrc.bell-labs.com

John G. Scudder
Internet Engineering Group, LLC
122 S. Main, Suite 280
Ann Arbor, MI 48104
e-mail: jgs@ieng.com