Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 7, 2017

DL. Duan
K. Patel
Cisco Systems
J. Hass
Juniper Networks

July 6, 2016

Fixing Persistent Route Oscillation conditions in BGP RT-Constrain
draft-idr-bgp-rt-oscillation-00.txt

Abstract

[RFC4684] defines Multi-Protocol BGP (MP-BGP) procedures that allow BGP speakers to exchange Route Target reachability information (RT-Constrain) to restrict the propagation of Virtual Private Network (VPN) routes. In network scenarios where hierarchical route reflection (RR) is used, the existing RT-Constrain mechanism may result in persistent routing oscillations within RRs. This document describes the problem scenario and proposes solutions to address persistent routing oscillations in hierarchical RR scenario.

This document updates RFC 4684 by proposing solutions to avoid the persistent routing oscillations in hierarchical RR scenario.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on January 7, 2017.

Table of Contents

1. Introduction

[RFC4684] defines Multi-Protocol BGP (MP-BGP) procedures that allow BGP speakers to exchange Route Target reachability information to restrict the propagation of Virtual Private Network (VPN) routes.

[RFC4684 section 3.2] defines a new route advertisement rule for Route Target membership information. When advertising a RT membership NLRI to a non-client peer, if the best path as selected by the path selection procedure described in Section 9.1 of the base BGP specification [4] is a route received from a non-client peer, and if there is an alternative path to the same destination from a client peer, then the attributes of the client path are advertised to the

peer. [RFC4684] does not clarify which path to choose in case there are multiple client paths to the same destination.

In network scenarios where hierarchical route reflection (RR) is used, in case there are multiple such client paths existing, persistent routing oscillations might be formed based on which client path attributes are advertised to the non-client peers.

1.1. Requirements Language

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].

2. Problem Statement - Persistent Routing Oscillations

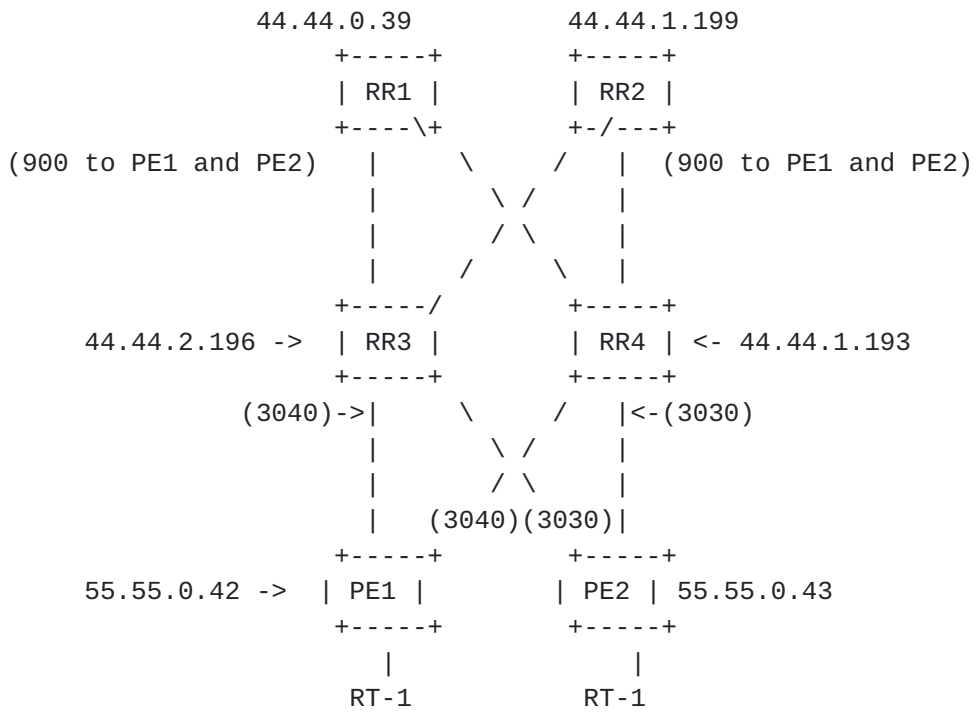

Figure 1. RT-Constrain with Hierarchical Route-reflector

In Figure 1, Hierarchical RRs are deployed. RR3 and RR4 are first level Router Reflectors and RR1 and RR2 are the second level Route Reflectors. PE1 and PE2 are Route Reflector clients of RR3 and RR4 while RR3 and RR4 are Route Reflector clients of RR1 and RR2. Both PE1 and PE2 are advertising route-target information of RT-1 to first level Router Reflectors RR3 and RR4. RR3 and RR4 are also

advertising route-target information to second level Router Reflectors RR1 and RR2. The numbers in the parentheses are metric to the nexthop.

At step #0 on RR3, RT-1 has two paths one path from PE1 and the other path from PE2. The path from PE1 has next hop metric 3040 and the path from PE2 has nexthop metric 3030. The path from PE2 is selected as a best path (lower metric) on RR3 and RR4. RR3 and RR4 advertise their best path for RT-1 to the second level router reflectors RR1 and RR2.

At step #1 on RR1, RT-1 has two paths one path from RR3 and the other path from RR4. The next hop metric to reach PE1 and PE2 are same 900. The path from RR4 get selected as best path because of lower originator id. RR1 advertise the RT-1 back to RR3. On RR1 and RR2 if originator is same, then the lower peering address will be used to select the best path.

At step #2 RR3 now has three or (four) paths: one from PE1, second one from PE2, and the third (and possibly forth) from RR1 (and/or RR2). The non-client path from RR1 to PE2 is selected as best path because of lower router id (or originator id). Since there are client path available to reach Route-target of RT-1, RR3 advertises the path attribute of a client path to RR1 according [RFC4684 section 3.2]. RR3 choose path attribute of RT-1 from PE1 randomly.

At step #3 RR1 receives the updates and recalculates the best path. The path from RR3 is selected as best path because of the lower peering address. RR1 updates the originator and send it back to RR3.

At step #4 RR3 receives the updates from RR1 and drop the updates since its own cluster-id is in the cluster list. Now RR-3's route state goes back to step #0 with 2 paths from clients and the whole cycle starts again.

Same thing happens on RR4 as on RR3 and same thing happens on RR2 as on RR1.

These iterations results in a persistent route oscillation for RT-1 prefix of RT-Constrain address-family on RR1, RR2, RR3 and RR4.

3. Potential Solution

This Draft specifies three solutions to the persistent routing oscillation issue described above.

First solution is that the attributes of client's best path should be advertised to the non-client peer. If the best path is from a non-

client peer, then select a client best path among all available and valid client paths. In such a case, at step #2 the path from PE2 will be selected as client best path. This is the same path as the best path at Step #0. So at step #3 on RR1 there is no best path change. The oscillation terminates.

Second solution is that Route Reflector always prefers the client paths when selecting a best path. So at Step #3 on RR1, the best path is still the path from PE2. The oscillation terminates with PE2's path. We could give client path the same priority as the local originated path. So client path will be preferred comparing to non-client path when calculate the best path.

Note that the scenario can not happen if RR1 and RR2 are part of the same cluster domain. So at step #2 RR3 only has two client paths. The update from top level Route Reflector will be dropped because of cluster id check. The oscillation never happens with such a topology.

4. IANA Considerations

This draft makes no request of IANA.

5. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4271] and [RFC4271].

5.1. Acknowledgements

The authors would like to thank Shyam Sethuram, Nitin Kumar, Sameer Gulrajani, Mohammed Mirza and Mike Dubrovskiy.

6. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <http://www.rfc-editor.org/info/rfc4684>.

Authors' Addresses

Dongling Duan
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: duan@cisco.com

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

Jeffrey Hass
Juniper Networks
1194 N. Mathida Ave
Sunnyvale, CA 94089
USA

Email: jhaas@juniper.net