Network Working Group
Internet-Draft
Intended status: Experimental
Expires: 23 February 2026

K. Vairavakkalai, Ed.
N. Venkataraman, Ed.
Juniper Networks, Inc.

22 August 2025

BGP Route Reflector with Next Hop Self
draft-ietf-idr-bgp-fwd-rr-04

Abstract

The procedures in BGP Route Reflection (RR) spec RFC4456 primarily deal with scenarios where the RR is reflecting BGP routes with next hop unchanged. In some deployments like Inter-AS Option C (Section 10, RFC4364), the ABRs may perform RR functionality with nexthop set to self. If adequate precautions are not taken, the RFC4456 procedures can result in traffic forwarding loop in such deployments.

This document illustrates one such looping scenario, and specifies approaches to minimize possiblity of traffic forwarding loop in such deployments. An example with Inter-AS Option C (Section 10, RFC4364) deployment is used, where RR with next hop self is used at redundant ABRs when they re-advertise BGP transport family routes between multiple IGP domains.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 https://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 23 February 2026.

Table of Contents

1. Introduction 2
2. Terminology 3
2.1. Definitions and Notations 3
3. Problem Description 4
4. Solution Approaches 5
4.1. Using Same Cluster ID at the ABRs 6
4.2. Using IGP Metric Management 6
4.3. Using AIGP Cost Management 6
4.4. Using BGP CT Management 7
5. IANA Considerations 7
6. Security Considerations 7
7. References 7
7.1. Normative References 7
7.2. Informative References 8
Appendix A. Appendix 8
A.1. Document History 8
Acknowledgements 9
Authors' Addresses 9

1. Introduction

The procedures in BGP Route Reflection (RR) spec RFC4456 primarily deal with scenarios where the RR is reflecting BGP routes with next hop unchanged. In some deployments like Inter-AS Option C (Section 10, RFC4364), the ABRs may perform RR functionality with nexthop set to self. If adequate precautions are not taken, the RFC4456 procedures can result in traffic forwarding loop in such deployments.

This document illustrates one such looping scenario, and specifies approaches to minimize possiblity of traffic forwarding loop in such deployments. An example with Inter-AS Option C (Section 10, RFC4364) deployment is used, where RR with next hop self is used at redundant ABRs when they re-advertise BGP transport family routes between multiple IGP domains.

2. Terminology

ABR: Area Border Router

AS: Autonomous System

AFI: Address Family Identifier

BN: Border Node

EP: Endpoint, e.g. a loopback address in the network

MPLS: Multi Protocol Label Switching

PE: Provider Edge

SAFI: Subsequent Address Family Identifier

2.1. Definitions and Notations

Service Family: A BGP address family used for advertising routes for destinations in "data traffic". For example, AFI/SAFIs 1/1 or 1/128.

Transport Family: A BGP address family used for advertising tunnels, which are in turn used by service routes for resolution. For example, BGP LU (AFI/SAFI : 1/4) or BGP CT (AFI/SAFI : 1/76).

Transport Tunnel : A tunnel over which a service may place traffic. Such a tunnel can be provisioned or signaled using a variety of means. For example, Generic Routing Encapsulation (GRE), UDP, LDP, RSVP-TE, IGP FLEX-ALGO or SRTE.

Tunnel Route: A Route to Tunnel Destination/Endpoint that is installed at the headend (ingress) of the tunnel.

Tunnel Domain: A domain of the network under a single administrative control, containing transport tunnels between Service Nodes (SNs) and Border Nodes (BNs).

3. Problem Description

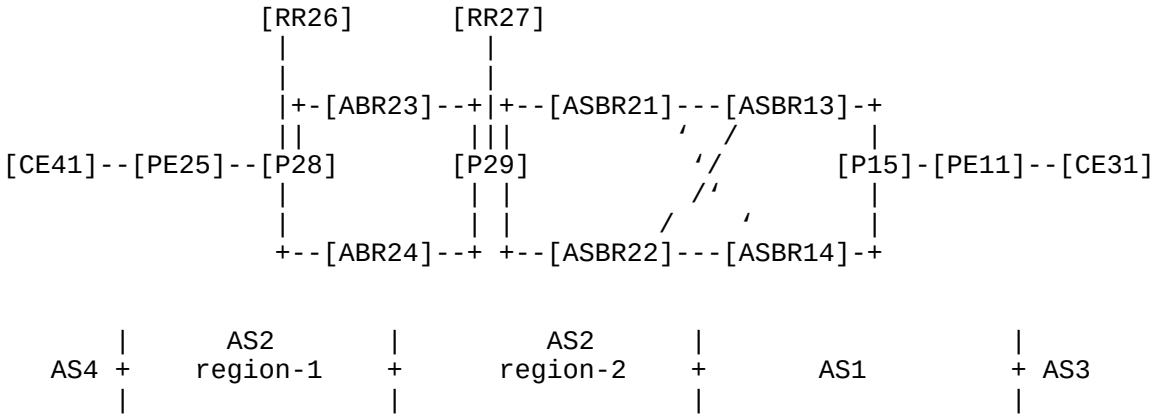

Figure 1: Reference Topology: Inter-domain BGP Transport Network

This topology shows an Inter-AS option C (Section 10, [RFC4364]) provider MPLS network that consists of two ASes, AS1 and AS2. They are serving customer networks AS3, AS4 respectively. Traffic direction being described is CE41 to CE31.

AS2 is further divided into two regions. There are three tunnel domains in provider's network: The two regions in AS2 use RSVP intra-domain tunnel. AS1 also uses RSVP-TE intra-domain tunnels. MPLS forwarding is used within these domains and on inter-domain links. BGP LU (AFI/SAFI: 1/4) is the transport family providing reachability between PE loopbacks PE25 and PE11.

Forwarding of PE25 to PE11 BGP LU traffic in AS2 region-2 is the focus of this discussion.

The following RSVP-TE tunnels exist in region-2.

- ABR23_to_ASBR21 - metric 40
- ABR23_to_ASBR22 - metric 30
- ABR24_to_ASBR21 - metric 40
- ABR24_to_ASBR22 - metric 30
- ABR23_to_ABR24 - metric 30

- ABR24_to_ABR23 - metric 30

The Router-ID of ASBR21 is better than ASBR22 from perspective of the BGP path selection.

The problem is that the pair of redundant ABRs (ABR23, ABR24 in Figure 1), each acting as an RR with next hop self, may choose each other as best path towards egress PE11, instead of the upstream ASBR (ASBR21 or ASBR22), causing a traffic forwarding loop.

This happens because of following the path selection rule specified in Section 9 of BGP RR [RFC4456] that tie-breaks on ORIGINATOR_ID before CLUSTER_LIST. RFC4456 considers pure RR functionality which leaves next hop unchanged.

This problem is more probable to happen for routes of BGP transport address families in Inter-AS Option C (Section 10 [RFC4364]) networks, like BGP LU (1/4 or 2/4) and BGP CT (AFI/ SAFIs: 1/76 or 2/76), because the ABRs perform RR with nexthop self functionality for these families.

Summarising, the necessary conditions for this problem are:

- Redundant ABRs perform RR with nexthop self
- The redundant ABRs using distinct CLUSTER_ID
- Addpath send enabled in Region 1, from RR26 to the redundant ABRs ABR23, ABR24
- ABR23, ABR24 using per-prefix label allocation mode for the transport layered families.
- IGP metric situations in Region 2, as explained above.
- Existance of Inter-ABR tunnels.
- RFC4456 tie-breaks on ORIGINATOR_ID before CLUSTER_LIST
- Router-ID values for upstream ASBRs.

4. Solution Approaches

Using one or more of the following approaches softens the possibility of such loops in an Inter-AS Option C network with redundant ABRs. These approaches manage one of the above necessary conditions.

4.1. Using Same Cluster ID at the ABRs

Configuring the same CLUSTER_ID at the redundant ABR nodes.

CLUSTER_ID Loop check will make routes reflected by an ABR unusable at the redundant ABRs.

This approach provides a stable way to avoid this loop, and is not affected by network churn.

However this approach does not allow the ABR-ABR tunnels to be used as backup path, in the event where an ABR looses all tunnels to upstream ASBR.

4.2. Using IGP Metric Management

Assign IGP metrics, such that "ABR to redundant ABR" cost is inferior to "ABR to upstream ASBR" cost.

Then ’IGP metric’ based tie-breaker will make an ABR chose the ASBRs as best path, instead of redundant ABR.

Since IGP metrics may change during network churn caused by events like link down, this approach needs careful planning to handle all possible IGP metric change scenarios. Debugging any loops caused by such transient situations may be much harder.

This approach allows using the ABR-ABR tunnels to be used as backup path, in the event where an ABR looses reachability to upstream ASBR. But there is a possibility of transient forwarding loop until BGP withdrawals are received, in situations where the redundant ABRs simultanously loose tunnel to upstream ASBR (like upstream ASBR failure event). Some mechanism like the one described in [MNH] Sec A.9 may be needed to handle the transient forwarding loop problem.

4.3. Using AIGP Cost Management

Using AIGP Cost in the network accumulates the IGP metric at "each next hop self" re-advertisement. This provides a better accumulated metric for the path.

Then ’AIGP cost’ based tie-breaker will make an ABR chose the ASBRs as best path, instead of redundant ABR.

This approach also needs careful IGP metric planning because it depends on the underlying IGP metric view of each node.

However this approach allows using the ABR-ABR tunnels to be used as backup path, in the event where an ABR looses all tunnels to upstream ASBR.

This approach allows using the ABR-ABR tunnels to be used as backup path, in the event where an ABR looses reachability to upstream ASBR. But there is a possibility of transient forwarding loop until BGP withdrawals are received, in situations where the redundant ABRs simultanously loose tunnel to upstream ASBR (like upstream ASBR failure event). Some mechanism like the one described in MNH Sec A.9 may be needed to handle the transient forwarding loop problem.

4.4. Using BGP CT Management

In a BGP CT network, using procedures described in [BGP-CT] , tunnels belonging to a certain Transport Class (TC) may not be provisioned between the redundant ABRs, or may not be included in the customized Resolution Scheme used to resolve BGP CT routes with that TC.

This will ensure that the BGP CT route received with redundant ABR as next hop will be Unusable at the receiving ABR, because it will fail resolving the next hop.

This approach needs Transport Class and Resolution Scheme planning in the BGP CT network, and provides a stable way to avoid this loop, and is not affected by network churn.

However this approach does not allow the ABR-ABR TC tunnels to be used as backup path, in the event where an ABR looses all tunnels for that TC to upstream ASBR.

5. IANA Considerations

This document makes no new requests of IANA.

6. Security Considerations

This document does not change the underlying security issues inherent in the existing BGP protocol, such as those described in [RFC4271], [RFC4272] and [RFC4456].

Mehanisms described in this document reduce possibility of loops within an IBGP domain. They do not affect routing across EBGP sessions.

7. References

7.1. 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, <https://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, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

7.2. Informative References

[BGP-CT] Vairavakkalai, Ed. and Venkataraman, Ed., "BGP Classful Transport Planes", 17 March 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-ct-28>.

[MNH] Vairavakkalai, Ed., "BGP MultiNexthop Attribute", 17 March 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-multinexthop-attribute-00>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

Appendix A. Appendix

A.1. Document History

The content in this document was introduced as part of [BGP-CT]. But because the described problem is not specific to BGP CT and is useful for other BGP families also, it is being extracted out to this separate document.

Acknowledgements

The authors thank Jeff Haas, Jon Hardwick, Keyur Patel, Igor Malyushkin, Robert Raszuk, Susan Hares for the discussions and review comments.

Authors' Addresses

Kaliraj Vairavakkalai (editor)
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: kaliraj@juniper.net

Natrajan Venkataraman (editor)
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America
Email: natv@juniper.net