Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 17 March 2026

S R. Mohanty
Zscaler
I. Means
P. Ramadenu
AT&T Labs, Inc.
M. Mishra
Cisco Systems, Inc.

13 September 2025

The Secondary Label and its applications
draft-mohanty-idr-secondary-label-02

Abstract

This draft utilizes the concept of a secondary label to solve few cases in L3VPN Deployments.In BGP VPN networks, BGP speakers associate a local MPLS label when the next-hop is reset and advertise that label to other peers. The receiving peer installs this "received" label in the forwarding and forwards traffic to the sending router using this label. In some deployments, there arises need where a different label is required to be sent. We illustrate with two use-cases.

This draft presents a method where this label is encoded in a newly defined attribute that is advertised with the BGP updates targeting these specified use-cases

Table of Contents

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 [RFC2119].

2. Introduction

In BGP L3VPN VPN networks, BGP speakers associate a local MPLS label with the next-hop is reset and advertise that label to other peers. The receiving peer installs this "received" label in the forwarding and directs traffic to the sending router using this label. This local label allocation is governed by the configured label allocation mode. Broadly, most vendors already offer different allocation modes like per-vhf, per-prefix, per-next-hop and per-nexthop-per-received-label.

In certain cases, the exclusive allocation of the local label is not sufficient. In this draft, we outline use-cases where the allocation of an additional label, hereby referred to as the secondary label, is necessary to be communicated to the BGP peer. Using this secondary label, the peer can impose forwarding decisions and solve some use-cases that are significantly non-trivial to achieve with the standard local-label allocation alone.

3. The per-Nexthop-received-label Mode

The standard behavior in case of option-B ASBR [RFC4364] is to allocate a per-prefix label for vpn prefixes. To conserve label space at the ASBR, many vendors implement a label allocation mode called per-nexthop-received- label. With per-nexthop-received-label, all prefixes received with the same next-hop and same received-label (both together constitute the label context) will be assigned the same local label. This approach conserves label space by avoiding the allocation is a unique label for each prefix. In case of Primary/backup, the context of the label allocation is the set of tuples {(Nexthop, recvd-label)} The above implementation (originally meant for the ASBR) also applies to RR with next-hop-self. In the below topology (representative of a tier 1 provider topology), RR1 and RR2 have the per-nexthop-received-label mode configuration and have next-hop-self towards each other. Both RRs receive the VPN prefix (RD 1:1: 2.2.2.2/32) from R1 with its connected address as the next-hop and advertise to the other RR on the cross-link after resetting the next-hop to self.

Although we will not explain here, a similar topology can be thought of in an dual Option-B deployment where the ASBRs will have each other as backup [RFC2119].

4. Problem Description

4.1. Problem Description 1 (PD#1)

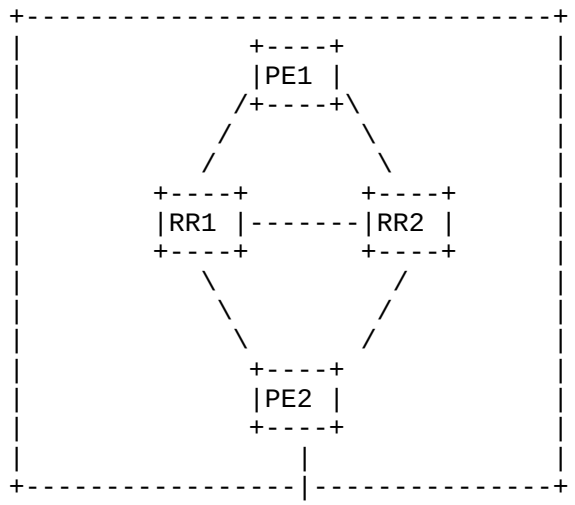

Figure 1 RR deployed with nexthop-self in a symmetric PIC Configuration

Figure 1

Figure 1 represents an all-IBGP network. PE1 is originating VPN routes and advertising them to RR1 and RR2. Both these RRs are also clients of each other and advertise VPN routes to each other with the next-hop set to the peering address. Each RR considers the path from PE1 as the best and the backup from the other RR (BGP PIC for VPNV4 and VPNV6 is configured). Label mode per-nexthop-received-label is configured.

a. This is how the issue gets manifested.

b. Initially, RR1 receives the primary path from PE1. Local Label allocation at RR1 has context [(PE1, LabelPE1)] and local label, LabelRR1 is allocated. This label is advertised to RR2.

c. Similarly, RR2 receives the primary path from PE1. Local Label allocation at RR2 has context [(PE1, LabelPE1)] and local label, LabelRR2 is allocated. This label is advertised to RR1

d. RR1 gets the update from RR2. It now sees the label context as [(PE1, LabelPE1),(RR2, LabelRR2)] and allocates local label LabelRR11. This label now becomes the received label at RR2.

e. RR2 now sees the label context as [(PE1, LabelPE1),(RR1, LabelRR11)] and allocates local label LabelRR21. This label now becomes the received label at RR1.

f. RR1 gets the update from RR2. It now sees the label context as [(PE1, LabelPE1),(RR2, LabelRR21)] and allocates local label LabelRR12. This label now becomes the received label at RR2 and this process continues.

The root cause of the label churn in is because the local label in RR1 (same for RR2) is an input to the label allocation context at RR2, and the resulting allocated local label at RR2 now serves as an input into the label allocation context at RR1. Because of this feedback loop the situation quickly results in the RRs getting out of label space very quickly.

Notice that if the RRs have the per-prefix label allocation mode configured, then this sort of oscillation will not happen. However, the per-prefix label allocation in an RR with next-hop-self configured will also mean a unique label for every unique prefix and that is not scalable.

4.2. Problem Description 2 (PD#2)

ISP1 and ISP2 are CE devices that establish an EBGP session with PE1 and PE2 respectively. Both ISPs advertise the same 700k prefixes/ routes to PE1 and PE2. Both PE1 and PE2 only send the default route to the remote PE, PE0.

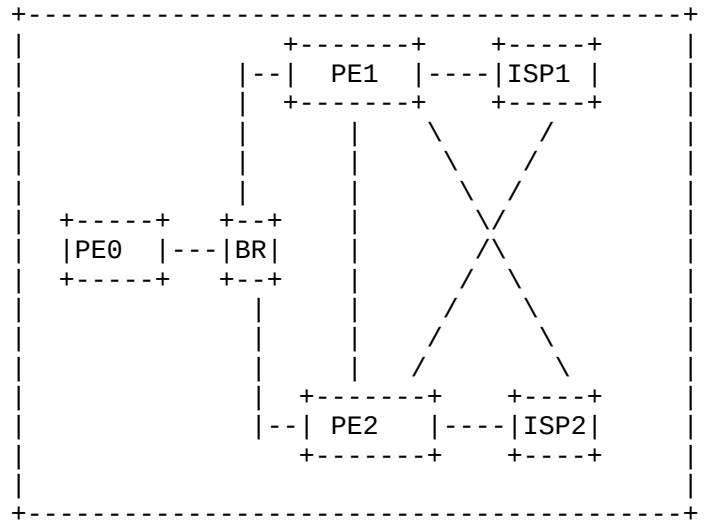

Figure 2 Dual Homed CE Setup

Figure 2

a. The PE devices have VPNv4 peering between them. BR is a P router.

b. Traffic from the remote PE, PE0, does ECMP forwarding to both PE1 and PE2.

c. The 400K routes prefer ISP1 as the egress NH and 300K routes prefer ISP2 as the egress NH by the policy configured on PE devices.

d. The Policy is a simple BGP policy that associates the highest Local Preference (LP) with the preferred EBGP path, the next highest local preference with the IBGP path and the lowest local preference with the least preferred EBGP path.

Failure scenario 1 (FS#1) When ISP1-PE1 link goes down , after BGP at PE1 converges, traffic will traverse the link through PE2, and then reach ISP1.

Failure scenario 2 (FS#2) The links from ISP1 to PE1 and PE2 are down at the same time; traffic will go to ISP2 after the BGP convergence at PE1 and PE2

FS#1 is a classic example when BGP PIC is appropriate . It is no wonder therefore that the convergence is good. However, in case of FS#2, with BGP PIC in place, this is what happens:

1. when the link between PE1 and ISP1 went down, the traffic that ingresses on PE1 gets diverted to PE2

2. Because the link between PE2 and ISP1 was torn down and the BGP withdraw from PE1 has not yet been received at PE2, this diverted traffic suffers a routing lookup at PE2 and will be sent back to PE1.

3. On PE1 it suffers a routing lookup again and diverted again to PE2.

4. This process repeats until the BGP withdraws corresponding to link breakages are received at the peer PEs

5. It is important to observe here that the particular label allocation mode (per-prefix or per-next-hop)has no bearing on the loop, it will still happen regardless.

6. FIB performance gets impacted due to the loop and new control plane state after convergence takes more time to be installed in the FIB.

The conclusion is BGP PIC by itself is not adequate to handle these kinds of convergence issues arising from double link-failures.

5. Proposed Solutions

5.1. Proposed Solution PS#1

To solve the issue of Section 4.1 the concept of a secondary label is introduced. At both the RRs, in addition to the local label another label, hereby referred to as the secondary label, is also allocated. This secondary label depends on the primary path exclusively i.e. only the path from PE1 and not on the path from RR2. This secondary label is encoded inside an attribute called the secondary label attribute that is advertised along with the the BGP best-path advertisement to RR2 and PE2. The format of the secondary label attribute is described in Section 6. Similar concept is described in [I-D.kaliraj-idr-multinexthop-attribute] but the next-hop and other fields etc. is not included in the secondary label attribute.

When RR2 receives the update from RR1 that it selects as its backup path and finds the secondary label attribute, it will only consider the label encoded in the secondary label attribute and ignore the received label in its local-label allocation decision. It will also program the label encoded in the secondary label attribute instead of the received label in the forwarding imposition. As the secondary label only depends on the primary path from PE1, it is unaffected by the advertisement from the other RR, and the continuous label churn is arrested immediately.

5.2. Proposed Solution PS#2

Without loss of generality, considering PE2 as the DUT, the main reason about the inability of BGP PIC (as described above) to help in this case is that the status of the primary link on the peer PE, PE1, the PE1-ISP1 link, is unknown to PE2 until it receives the corresponding BGP withdraw. Following is the main underlying idea of our proposed solution.

1. Allocate a Primary label with the primary path pointing to the directly connected preferred CE (best EBGP path) and the backup to the less preferred PE (IBGP Path).

2. Allocate a 2nd label with primary path to directly connected preferred CE (best EBGP path) and backup to the less preferred EBGP path. This second label is advertised in the Control Plane along with the primary label leveraging the idea of the secondary label. But the notion that this second (backup) label is also associated with a primary path and in case of failure also points to another backup path is what distinguishes this from the PD#1. Accordingly, secondary label needs to have a context

3. With the help of the Figure below, we explain our scheme with respect to the traffic of 400K, that prefers ISP1 (The explanation for the other 300k follows a symmetric reasoning).

4. Consider 10.10.1.0.0/24 as one such VPN prefix in the group of 400k.

5. Accordingly, PE1 allocates a primary label of 100 that points to the primary next-hop (NH), ISP1, and to the backup NH, PE2; and, a backup label of 200 (pointing to primary NH ISP1 and backup NH ISP2). Similarly, PE2 allocates a primary label of 300 (this primary label points to primary NH ISP1 and backup NH PE1) and a backup label of 400 (pointing to primary NH ISP1 and backup NH ISP2)

6. Traffic from the remote PEs always uses the primary label. Traffic sent from one peer PE to another is always sent using the backup label.

7. Therefore, traffic to 10.10.0.0/24 from the PE0 is received on PE1 with label 100. In the normal case, this traffic will be sent on the direct PE1-ISP1 link.

8. Now, if link PE1-IPS1 breaks, this traffic is diverted to PE2 with label 400.

9. When this traffic is received at PE2, if the PE2-ISP1 link is up, traffic will be forwarded to ISP1 on that link. But, now if the PE2-ISP1 goes down, the backup path for the label 400 which points to the NH ISP2, is activated immediately and the traffic is directed to ISP2 on the PE2-ISP2 link.

6. Secondary Label Attribute

A new Optional Transitive Attribute will be created for carrying the secondary label. This attribute will be referred as the secondary label attribute. The format is as specified below.

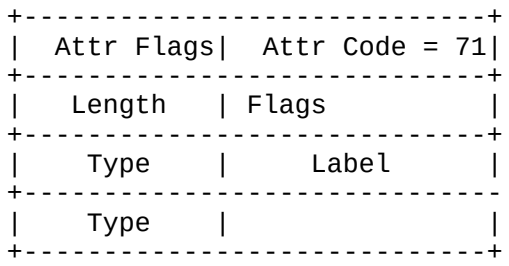

Figure 3 Secondary Label Attribute

Figure 3

The Secondary label attribute contains a flags field (1-byte) and a set of Type (1 byte) and Label (3 bytes). The flag bits will be specified in the future. The label type will denote the context, for PS1#, the type is 0, for PS#2, the type is 1. As we find more and more use-cases, types will be assigned appropriately.

We will request IANA assignment for the secondary label attribute

7. Conclusion

We have described two use-cases where the concept of a second label greatly helps in optimizing network resources and improve convergence at the potential cost of increasing the label allocation resources. However, the advantages of the solutions with the secondary label are the simplicity, the optimization and convergence improvements that it provides to the network. There can be many potential use-cases for this secondary label concept.

We will request IANA assignment for the secondary label attribute.

7.1. IANA Considerations

Request IANA assignment for the secondary label attribute with code-type 71

7.2. Operational Considerations

TBD.

7.3. Security Considerations

This document raises no new security issues for RT Constraints.

7.4. Acknowledgements

TBD.

8. Contributors

The following individuals made significant contributions to this document:

* Bhavik Patel (AT&T Labs, email: bp536y@att.com)

9. Normative References

[I-D.kaliraj-idr-multinexthop-attribute]
Vairavakkalai, K. and J. M. Jeganathan, "BGP MultiNexthop Attribute", Work in Progress, Internet-Draft, draft-kaliraj-idr-multinexthop-attribute-07, 5 July 2023, <https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-multinexthop-attribute-07>.

[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>.

[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>.

[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>.

Authors' Addresses

Satya Ranjan Mohanty
Zscaler
120 Holger Way
San Jose, CA 95134
United States of America
Email: smohanty@zscaler.com

Israel Means
AT&T Labs, Inc.
7337 Trade St
San Diego, CA 92121
United States of America
Email: im8327@att.com

Praveen Ramadenu
AT&T Labs, Inc.
3538 Torrance Blvd Unit 124
Torrance, CA 90503
United States of America
Email: pr9637@att.com

Mankamana Prasad Mishra
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
United States of America
Email: mankamis@cisco.com