BESS WG
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2021

Y. Wang
Z. Zhang
ZTE Corporation

July 4, 2020

ARP/ND Synching And IP Aliasing without IRB
draft-wang-bess-evpn-arp-nd-synch-without-irb-06

Abstract

This document proposes an extension to [RFC7432] and [I-D.sajassi-bess-evpn-ip-aliasing] to do ARP synchronizing and IP aliasing for Layer 3 routes that is needed for EVPN signalled L3VPN to build a complete IP ECMP. The phrase "EVPN signalled L3VPN" means that there may be no MAC-VRF or IRB interface in the use case. When there are no MAC-VRF or IRB interface, EVPN signalled L3VPN is also called as "pure L3VPN instance" which is a different usecase from [I-D.sajassi-bess-evpn-ip-aliasing].

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 January 5, 2021.

Table of Contents

1. Introduction 2
1.1. EVPN signalled L3VPN 3
1.2. Integrated Routing and Cross-connecting 4
1.3. Terminology 5
2. ARP/ND Synching and IP Aliasing 6
2.1. Constructing MAC/IP Advertisement Route 7
2.2. Constructing IP-AD/EVI Route 8
2.3. Constructing IP-AD/ES Route 8
3. Fast Convergence for Routed Traffic 9
4. Determining Reach-ability to Unicast IP Addresses 9
5. Forwarding Unicast Packets 9
6. RT-5 Routes in EVPN signalled L3VPN 10
6.1. RT-5E Advertisement on Distributed L3 GW 10
6.2. Centerlized RT-5G Advertisement for Distributed L3 Forwarding 11
6.2.1. Centerlized CE-BGP 12
6.2.2. RT-2E Advertisement from PE1/PE2 to PE3 12
6.2.3. RT-5G Advertisement from PE3 to PE1/PE2 12
6.2.4. RT-2E Advertisement between PE1 and PE2 13
6.2.5. Egress ESI Link Protection between PE1 and PE2 . . . 13
6.2.6. Comparing with Distributed RT-5G Advertisement . . . 13
6.2.7. Mass-Withdraw by EAD/ES Route 14
6.2.8. On the Failure of PE3 Node 14
6.2.9. Floating GW-IP between R1 and R2 15
6.3. RT-5L Advertisement 15
7. Load Balancing of Unicast Packets 16
8. Special Considerations for Single-Active ESIs 16
9. Security Considerations 16
10. IANA Considerations 16
11. References 17
11.1. Normative References 17
11.2. Normative References 18
Authors' Addresses 18

1. Introduction

In [I-D.sajassi-bess-evpn-ip-aliasing], an extension to [RFC7432] to do aliasing for Layer 3 routes is proposed for symmetric IRB to build a complete IP ECMP. But typically there may be both IRB interfaces(to do EVPN IRB per-MAC-VRF basis) and VRF- ACs in the same IP-VRF instance. It is necessary to apply the EVPN control-plane to the VRF-ACs in order to support EVPN signalled L3VPN, including such

mixed situations, the pure L3VPN instance use case where maybe no IRB interfaces will be found in the IP-VRF instances.

There are also an Integrated Routing and Cross-connecting use case which is described in Section 1.2.

1.1. EVPN signalled L3VPN

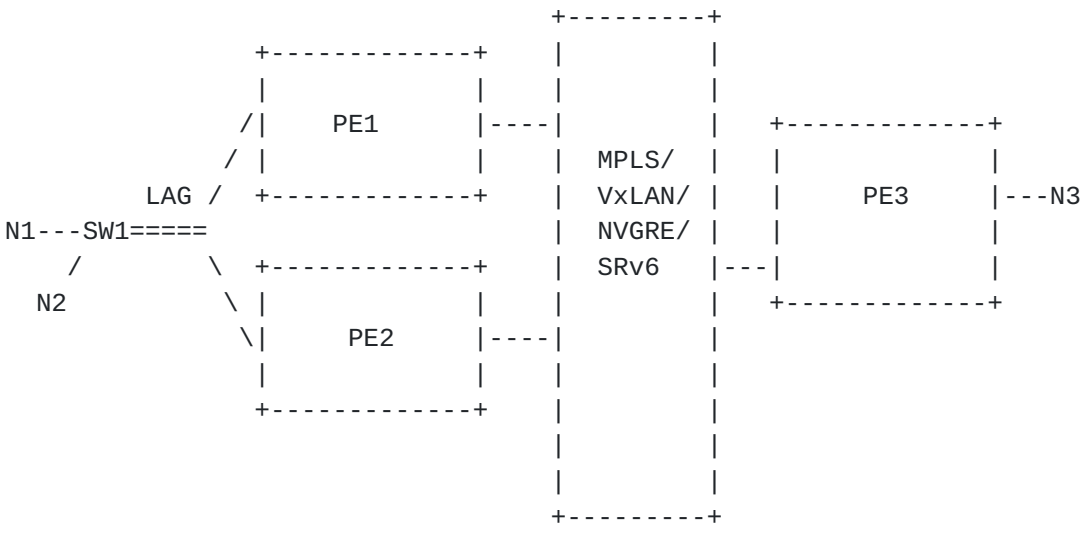

Figure 1: ARP/ND Synchronizing and IP Aliasing without IRB

There are three CE nodes named N1/N2/N3 in the above network. N1/N2/N3 may be a host or a IP router. When N1/N2/N3 is a host, it is also called H1/H2/H3 in this document. When N1/N2/N3 is a router, it is also called R1/R2/R3 in this document.

Consider a pair of multi-homed PEs PE1 and PE2. Let there be two hosts H1 and H2 attached to them via a L2 switch SW1. Consider another PE PE3 and a host H3 attached to it. The H1 and H2 represent subnet SN1 and the H3 represents subnet SN2.

Note that it is different from [I-D.sajassi-bess-evpn-ip-aliasing] in the following aspects: There may be no MAC-VRF or IRB interface on PE1/PE2/PE3. And it is the IP-VRFs that are called as EVPN instance instead. Such EVPN instance can be called pure L3 EVPN instance or L3 EVI for short. The anycast gateway of H1/H2 is configured on a sub-interface on PE1/PE2.

Note that the communication between H1 and H2 won't pass through any of the multi-homed PEs. So it is not necessary for PE1/PE2 keeping a Broadcast domain and its IRB for SN1.

Note that the SW1 multi-homing PE1 and PE2 via a LAG interface which maybe load-balance traffic to the PEs.

This draft proposes an extension to do ARP/ND synchronizing and IP aliasing for Layer 3 routes that is needed for L3 EVI to build a complete IP ECMP.

1.2. Integrated Routing and Cross-connecting

When an IP-VRF instance and an EVPN VPWS instance is connected by an virtual-interface, We call such scenarios as Integrated Routing and Cross-connecting (IRC) use-case where the EVPN VPWS is represent by the term "cross-connecting", and the IP-VRF is represent by the term "Routing". The virtual-interface connecting EVPN VPWS and IP-VRF is called as IRC interface.

The IRC use case is illustrated by the following figure:

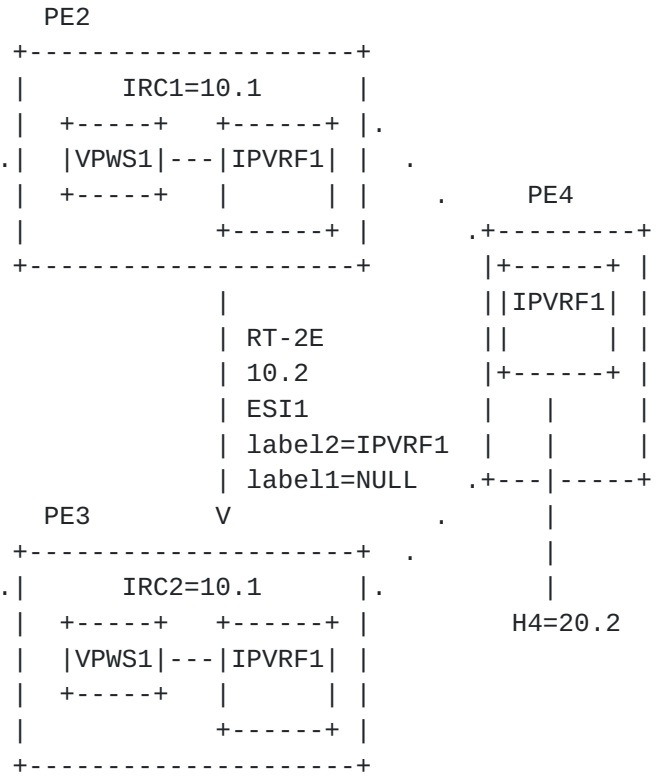

Figure 2: ARP/ND Synchronizing for IRC Interfaces

Note that the IRC interfaces are considered as AC interfaces in EVPN VPWS interface. At the same time, they are considered as VRF-ACs in IP-VRF instances.

When H1 sends an ARP packet P1, then PE1 will be forwarded by PE1 to either PE2 or PE3, not to the both. Both the IRC1 on PE2 and IRC2 on PE3 are H1's subnet-gateway(SNGW). But when H4 send an packet P2 to H1, then PE4 may load-balance P2 to either PE2 ore PE3, not to the both.

When P1 is load-balance to PE2, not to PE3, but PE4 load-balance P2 to PE3, The ARP entry of H1 will not be prepared on PE3 for P2. So the fowarding of P2 will be delayed due to ARP missing.

We use RT-2 routes to advertise the ARP entry of H1 from PE2 to PE3. But there SHOULD be no RT-2 advertisement in EVPN VPWS according to [RFC8214]. So the RT-2 routes from PE2 to PE3 SHOULD not carry any export-RTs of VPWS1, and the label1 of these RT-2 route will be set to NULL.

The NULL value of label1 in MPLS EVPN should be implicit-null. The NULL value of label1 in VXLAN EVPN should be 0.

Note that an ESI may be assigned to IRC1 and IRC2, Because the ESI of the RT-2 routes will be used to determine that to which the ARP entries should be installed.

1.3. Terminology

Most of the terminology used in this documents comes from [RFC7432] and [I-D.sajassi-bess-evpn-ip-aliasing] except for the following:

VRF AC: An Attachment Circuit (AC) that attaches a CE to an IP-VRF but is not an IRB interface.

IRC: Integrated Routing and Cross-connecting, thus a IRC interface is the virtual interface connecting an IP-VRF and an EVPN VPWS.

VRF Interface: An IRB interface or a VRF-AC or an IRC interface. Note that a VRF interface will be bound to the routing space of an IP-VRF.

L3 EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN which contains VRF ACs and maybe contains IRB interfaces or IRC interfaces.

IP-AD/EVI: Ethernet Auto-Discovery route per EVI, and the EVI here is an IP-VRF.

IP-AD/ES: Ethernet Auto-Discovery route per ES, and the EVI for one of its route targets is an IP-VRF.

CE-BGP: The BGP session between PE and CE. Note that CE-BGP route doesn't have a RD or Route-Target.

RMAC: Router's MAC, which is signaled in the Router's MAC extended community.

RT-2E: A MAC/IP Advertisement Route with a non-reserved ESI.

RT-5E: An EVPN Prefix Advertisement Route with a non-reserved ESI.

RT-5G: An EVPN Prefix Advertisement Route with a zero ESI and a non-zero GW-IP.

RT-5L: An EVPN Prefix Advertisement Route with both zero ESI and zero GW-IP.

2. ARP/ND Synching and IP Aliasing

Host IP and MAC routes are learnt by PEs on the access side via a control plane protocol like ARP. In case where a CE is multihomed to multiple PE nodes using a LAG and is running in All-Active Redundancy Mode, the Host IP will be learnt and advertised in the MAC/IP Advertisement only by the PE that receives the ARP packet. The MAC/IP Advertisement with non-zero ESI will be received by both PE2 and PE3.

As a result, after PE2 receives the MAC/IP Advertisement and imports it to the L3 EVI, PE2 installs an ARP entry to the VRF interface whose subnet matches the IP Address from the MAC/IP Advertisement. Such ARP entry is called remote synched ARP Entry in this document.

Note that the PEs follow [I-D.sajassi-bess-evpn-ip-aliasing] to achieve the ESI load balance except for the constructing of MAC/IP Advertisement Route and IP AD per EVI route.

When PE3 load balance the traffic towards the multihomed Ethernet Segment, both PE1 and PE2 would have been prepared with corresponding ARP entry yet because of the ARP synching procedures.

It is important to explain that typically there may be both IRB interface and VRF interface in an IP-VRF instance, which is called as the "VRF interface in EVPN IRB" use-case in this document. But each IRB/VRF interface is independent to each other in EVPN control plane. So the use-case here is constrained to a pure L3 EVPN schema, Because it is enough to describe all the control-plane updates for both the pure L3 EVPN use-case and the "VRF interface in EVPN IRB" use-case.

In current EVPN control-plane for "VRF interface in EVPN IRB" use-case, the VRF interface is considered as "external link" and it just inter-operates with the EVPN control-plane. But in this document it is assumed to be better if the EVPN control-plane directly applied to the VRF interfaces.

2.1. Constructing MAC/IP Advertisement Route

This draft introduces a new usage/construction of MAC/IP Advertisement route to enable Aliasing for IP addresses in pure L3 EVPN use-cases. The usage/construction of this route remains similar to that described in RFC 7432 with a few notable exceptions as below.

* The Route-Distinguisher should be set to the corresponding L3VPN context.

* The Ethernet Tag should be set to 0.

* The MAC/IP Advertisement SHOULD carry one or more IP VRF Route-Target (RT) attributes.

* The ESI SHOULD be set to the ESI of the VRF interface from which the ARP entry is learned.

Note that the ESI is used to install remote synched ARP entries to corresponding VRF interfaces on PE1/PE2. But it is only used to load balance traffic on PE3.

* The MPLS Label1 should be set to implicit-null in MPLS/SRv6 encapsulation. For VXLAN encapsulation, the MPLS label1 should be set to 0 instead. Note that in IRC use case, although there is a L2 EVPN instance (EVPN VPWS), the EVPN label and export-RT of that EVPN VPWS will not be carried in the MAC/IP route.

Note that there may be no MAC-VRF here, and this is outside the scope of RFC 7432.

* The MPLS Label2 should be set to the local label of the IP-VRF in MPLS or VXLAN EVPN. But it should be set to implicit-null in SRv6 EVPN.

Note that the label may be VNI label or MPLS label.

Note that in SRv6 EVPN an SRv6 L3 Service TLV MAY also be advertised along with the route following [I-D.dawra-bess-srv6-services]. But SRv6 L2 Service TLV won't be advertiseed along with the route. Because that no MAC-VRF exists in the use case.

* The RMAC Extended Community attribute SHOULD be carried in VXLAN EVPN.

2.2. Constructing IP-AD/EVI Route

Note that the IP-AD/EVI Advertisement is used for two reasons. It is used between PE1 and PE2 to do egress link protection for the subnet of the downlink VRF-interface. It is used between PE1/PE2 and PE3 to achieve the load balance to ES adjacent PEs.

The usage/construction of this route is similar to the IP-AD per EVI route described in [I-D.sajassi-bess-evpn-ip-aliasing] with a few notable exceptions as below.

Note that there may be no MAC-VRF here, and this is outside the scope of [RFC7432] and [I-D.sajassi-bess-evpn-ip-aliasing].

Note that the Encapsulation Sub-TLV of Tunnel Encapsulation Attribute per [I-D.ietf-idr-tunnel-encaps] may be used to emphasize that the RMAC in the Encapsulation Sub-TLV will be preferred.

Note that, in [I-D.ietf-idr-tunnel-encaps] setion 7, when the next hop of BGP UPDATE U1 is router X1 and the best path to router X1 is a BGP route that was advertised in UPDATE U2, and both U1 and U2 have a tunnel encapsulation attribute, the data packet will be carried through a pair of nested tunnels, each corresponding to a tunnel encapsulation attribute. But when U1 is a RT-2E route and U2 is an IP-AD/EVI route, the ESI in the recursion is not considered as a "next hop" of [I-D.ietf-idr-tunnel-encaps] setion 7. So only the tunnels in IP-AD/EVI route will be used, although both of the two EVPN routes have a Tunnel Encapsulation attribute.

Note that we have special considerations for single-active ESIs than [I-D.sajassi-bess-evpn-ip-aliasing], and it is detailed in Section 8.

Such Ethernet Auto-Discovery route is called Ethernet Auto-Discvoery route per IP-VRF which is abbreviated as EAD/IP-VRF in the old versions of this document.

2.3. Constructing IP-AD/ES Route

The usage/construction of this route remains similar to the IP AD per ES route described in [I-D.sajassi-bess-evpn-ip-aliasing] section 3.1 with a few notable exceptions as explained as below.

There may be no MAC-VRF RTs in the IP-AD/ES Route.

Such Ethernet Auto-Discovery route is called EAD/ES route in the old versions of this document.

3. Fast Convergence for Routed Traffic

The procedures for Fast Convergence do not change from [I-D.sajassi-bess-evpn-ip-aliasing] except for a few notable exceptions as explained as below.

The local ARP entries and remote synced ARP entries is installed/ learned on a VRF interface rather than an IRB interface.

There is no MAC entry.

4. Determining Reach-ability to Unicast IP Addresses

The procedures for local/remote host learning and MAC/IP Advertisement route constructing are described above. The procedures for Route Resolution do not change from [I-D.sajassi-bess-evpn-ip-aliasing] and/or [I-D.ietf-bess-evpn-prefix-advertisement].

5. Forwarding Unicast Packets

Because of the nature of the MPLS label or SRv6 SID for IP-VRF instance, when these IP-AD/EVI routes are referred in IP-VRF routing and forwarding procedures, the inner ethernet headers are absent on the corresponding packets transported following these IP-AD/EVI routes.

Note that in [I-D.sajassi-bess-evpn-ip-aliasing] the IP-AD per EVI route carries a "Router's MAC" extended community in case the RMAC is not the same among different PEs. In these cases, the inner destination MAC of the corresponding data packets from PE3 to PE1/PE2 must use the RMAC in IP-AD/EVI route instead, even if there is a RMAC in RT-2E route.

Note that this is a data-plane update of [I-D.ietf-bess-evpn-prefix-advertisement] for both EVPN signalled L3VPN and [I-D.sajassi-bess-evpn-ip-aliasing]. According to [I-D.ietf-bess-evpn-prefix-advertisement] section 4.3 or [I-D.ietf-bess-evpn-inter-subnet-forwarding] section 3.2.3, the inner destination MAC will follow the RMAC of RT-5E Route or RT-2E Route. Although PE3 SHOULD prefers the RMAC in the IP-AD/EVI routes following this document, we also suggest the RMAC being included in RT-2E or RT-5E route for compatibility.

When a packet is forwarded following the subnet route of a downlink VRF-interface, and the bypass tunnel is used, the ARP lookup is not needed because of the RMAC in the IP-AD/EVI route. But if the downlink VRF-interface is up at that time, the ARP lookup is used to encapsulated the destination MAC of the packet's ethernet header as usual.

Note that the packets received from a bypass tunnel can only be forwarded to a local downlink VRF-interface. In order to prevent the micro loop on R1's node failure, a few split-horizon filter rules should be introduced. In EVPN NVO3, the packet received from a tunnel is not allowed to forwarded to the same tunnel. In SRv6 EVPN, the packet received from a locator may be not allowed to forwarded to the same locator based on configurations. In MPLS EVPN, the packet may include an extra label to identify its ingress router as proposed in [I-D.wang-bess-evpn-context-label]. In MPLS EVPN, the packet may include an extra label to identify that it is forwarded on a bypass tunnel. And the extra label can be a extended special-purpose label or an ESI label.

6. RT-5 Routes in EVPN signalled L3VPN

EVPN signalled L3VPN can be deployed without EVPN IRB like what MPLS/BGP VPNs have done for a long time, but it can be combined with EVPN IRB. The EVPN siganlled L3VPN without EVPN IRB is not well defined yet, so we take the non-IRB usecase as an example. But the following routes and procedures can be used in EVPN IRB usecase too. Note that in EVPN IRB usecase, the IRB interfaces are VRF-interface too.

6.1. RT-5E Advertisement on Distributed L3 GW

Given that PE1/PE2 can install a synced ARP entry to its proper VRF-interface benefitting from the RT-2 route of section 2.1. So it is not necessary for PE1/PE2 to advertise per-host IP prefixes by RT-2 routes. It is recommended that PE1/PE2 advertise an RT-5 route per subnet to PE3 instead. The ESI of these RT-5E routes can be set to the ESI of the corresponding VRF interface. If the VRF interface fails, these subnets will achieve more faster convergency on PE3 by the withdraw of the corresponding IP-AD/EVI route.

Note that N1/N2 may be a host or a router, when it is a router, those subnets will be the subnets behind it. When N1 and N2 are hosts, those subnets will be the subnets of N1 and N2 whether they are different subnets or not.

6.2. Centerlized RT-5G Advertisement for Distributed L3 Forwarding

When N1/N2/N3 is a router, it is called R1/R2/R3 in the following figure. Note that figure 1 only illustrates the physical ethernet links, but figure 2 illustrates the logical L3 adjacencies between PE and CE as the following.

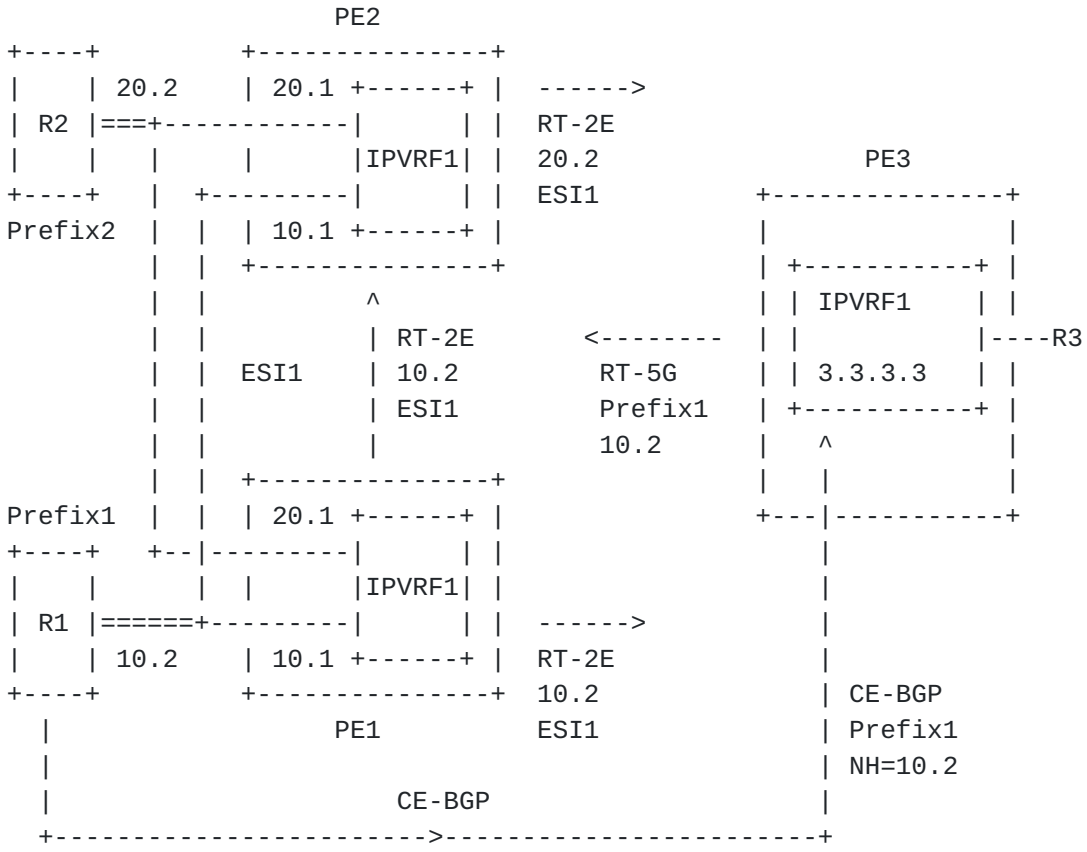

Figure 3: Centerlized RT-5G Advertisement

Note that R1/R2 should establish CE-BGP session with both PE1 and PE2 in case of one of them fails, PE1 and PE2 will advertise RT-5E route to PE3 for their prefixes learned from CE-BGP independently. If R1/R2 prefers to establish a single CE-BGP session, it can establish the CE-BGP session with PE3 instead. This CE-BGP session can be called the centerlized CE-BGP session. But when we use centerlized CE-BGP session, we should use RT-5G route instead.

Note that we just use centerlized CE-BGP session to do route advertisement, but we still expect a distributed Layer 3 forwarding framework.

6.2.1. Centerlized CE-BGP

The CE-BGP session between R1 and PE3 is established between 10.2 and 3.3.3.3. The CE-BGP session between R2 and PE3 is established between 20.2 and 3.3.3.3. The IP address 10.2/20.2 is called the uplink interface address of R1/R2 in this document. The IP address 3.3.3.3 is called the centerlized loopback address of IPVRF1 in this document. The IP address 10.1/20.1 is called the downlink VRF-interface address of PE1/PE2 in this document.

Note that the downlink VRF-interface is a Layer 3 link and it needn't attach an BD.

R1 advertises a BGP route for a prefix (say "Prefix1") behind it to PE3 via that CE-BGP session. The nexthop for Prefix1 is R1's uplink interface address (say 10.2).

The route advertisement of R2 is similar to the above advertisement.

Note that the packets from R1/R2 to the centerlized loopback address may be routed following the default route on R1/R2.

6.2.2. RT-2E Advertisement from PE1/PE2 to PE3

When PE1 learns the ARP entry of 10.2, it advertises a RT-2E route to PE3. The ESI value of the RT-2E route is ESI1, which is the ESI of PE1's downlink VRF-interface for R1. The RT-2E route is constructed following section 2.1.

Note that in [RFC7432], when the ESI is single-active, the MAC forwarding only use the label and the MPLS nexthop of the RT-2E route as long as they are valid for forwarding status. But in RT-5 routes we assume that the ESI is always preferred even if the ESI is single-active. This is similar to [I-D.ietf-bess-evpn-prefix-advertisement] section 3.2 Table 1. The ESI usage in IP forwarding is out of the [RFC7432]'s scope.

The RT-2E route advertisement of PE2 is similar to the above advertisement.

6.2.3. RT-5G Advertisement from PE3 to PE1/PE2

When PE3 receives the prefix1 from the CE-BGP session. The nexthop for Prefix1 is 10.2, and the ESI for 10.2 is ESI1. So PE3 advertises a RT-5G route to PE1/PE2 for Prefix1. The GW-IP value of the RT-5G route for Prefix1 is 10.2.

Note that PE3 can load-balance packets for Prefix1 via the IP-AD/EVI routes from PE1/PE2. Because ESI1 is the ESI for Prefix1's GW-IP.

The RT-5 route advertisement and packet forwarding for Prefix2 is similar to the above.

Note that the centerlized loopback address is advertised by PE3 via RT-5L route. The nexthop of the RT-5L route is PE3, and the GW-IP value of the RT-5L route is zero. The label of the RT-5L route is IPVRF1's label on PE3. The RMAC of the RT-5L route is PE3's MAC when the encapsulation is VXLAN.

Note that no Tunnel Encapsulation attribute should be carried in a RT-5G route, in order to avoid the nested tunnel encapsulation described in [I-D.ietf-idr-tunnel-encaps] setion 7.

6.2.4. RT-2E Advertisement between PE1 and PE2

The RT-2E routes advertisement between PE1 and PE2 is used to sync these ARP entries to each other in order to avoid ARP missing. The ESI Value of these two RT-2E routes is ESI1.

Note that we assume that the ARP entry for 10.2 will be learned on PE1 only, and 20.2 will be learned on PE2 only. Note that the two downlink VRF-interfaces for R1/R2 on PE1/PE2 are sub-interfaces of the same physical interface. So they have the same ESI.

6.2.5. Egress ESI Link Protection between PE1 and PE2

The IP-AD/EVI routes between PE1 and PE2 is used to do egress link protection. The egress link protection follows the second approach of the [RFC8679] section 6.

Note that although the ARP entry for 10.2 on PE2 is synced from PE1 via RT-2E route. The ARP entry on PE2 is installed to forward packets directly to the corresponding downlink VRF-interface primarily. The bypass tunnel following the IP-AD/EVI route is only activated when the downlink VRF-interface fails.

6.2.6. Comparing with Distributed RT-5G Advertisement

When R1/R2 establish CE-BGP sessions with both PE1 and PE2, The RT-5G routes can be used by PE1/PE2 instead of the RT-5E routes. But when R1 only establish just a single CE-BGP session with PE1, there will be some trouble when PE1 fails. Even if PE2/PE3 applies a delayed deletion when PE1 fails, the delay cann't be long enough when PE1 never comes up again.

Note that when there is only a single CE-BGP session, the RT-5E advertisement will face the same fact. In fact it is even worse when R1 uses different subnets to connect to PE1 and PE2 as described in [I-D.sajassi-bess-evpn-ip-aliasing] section 1.2. Because that RT-5E can only sync the prefixes, it can't sync the nexthops, so when PE2 receives a RT-5E route from PE1 the ARP entry for the other uplink interface that connects R1 to PE2 will not be resolved by PE2.

Note that when R1 uses different subnets to connect to PE1 and PE2 , it is not necessary to configure a BD for the two subnets connecting PE and CE like what is described in [I-D.sajassi-bess-evpn-ip-aliasing] section 1.2.

Note that we can make the RT-5E route carry the MAC address of its overlay nexthop (which is R1's uplink interface)'s ARP entry, so that when when PE2 receives a RT-5E route carrying such MAC address, these routes don't need to do ARP lookup. Such MAC address can be carried in a new extended community called as GW-MAC extended community. By doing so, when R1 uses different subnets to connect to PE1 and PE2, then the RT-5E can be used to sync the prefixes.

6.2.7. Mass-Withdraw by EAD/ES Route

We can assume that R1 and R2 are attached to different IP-VRFs(say IPVRF1 and IPVRF2 respectively), and the physical interface of the downlink VRF-interfaces on PE1 fails, PE1 will withdraw the IP-AD/ES route of ESI1, so PE3 will re-route 10.2 for Prefix1 in IPVRF1 and 20.2 for Prefix2 in IPVRF2 at the same time. Then data packets for Prefix1 and Prefix2 will be sent to PE2 instead.

6.2.8. On the Failure of PE3 Node

On the failure of PE3, PE1/PE2 should delay the deletion of the RT-5G route from PE3. PE3 can use a new BGP attribute to indicate the delayed-deletion requirement to PE1/PE2. Otherwise the L3 traffic between R1 and R2 will be interrupted. Fortunately, PE3 will typically have a redundant node (PE3' in Figure 3), and PE3' can be used to take PE3's place when PE3 fails.

Note that from the viewpoint of R1 and R2, the total of PE1, PE2, PE3, PE3' and the underlay network between them is regarded as the following logical router:

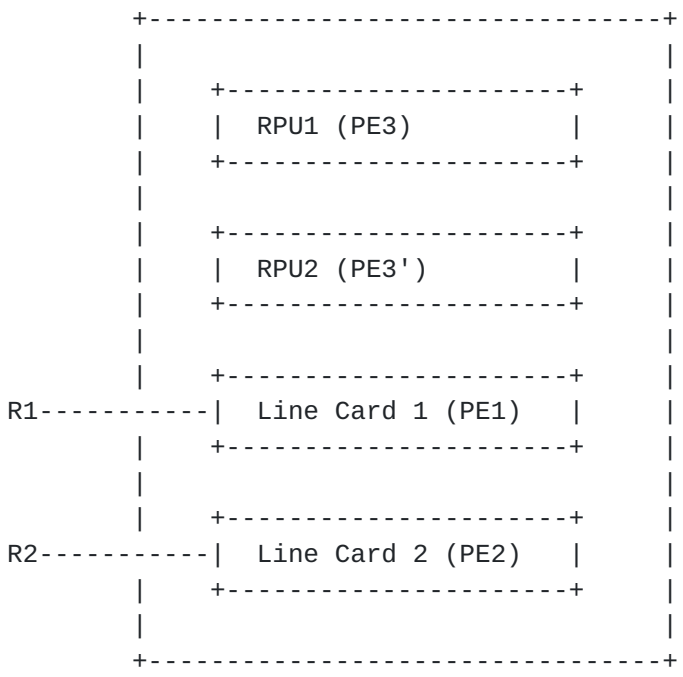

Figure 4: The Logical Router Framework

R1 and R2 connect to the line-cards of the logical router. and the data packets between R1 and R2 just pass through the line-cards, not through the RPUs(Routing Processing Units). But R1/R2 establish the BGP session with the RPUs, not the line-cards. When the RPU1(or actually PE3) fails, the line-cards(or actually PE1/PE2) will keep the forwarding state unchanged untill the RPU1 or RPU2 comes up. So the delayed deletion on PE1/PE2 for PE3's sake is apprehensible for the same reason.

6.2.9. Floating GW-IP between R1 and R2

It is similar to [I-D.ietf-bess-evpn-prefix-advertisement] section 4.2 except for a few notable differences as described in the following. There may be no BD in PE1/PE2/PE3. There is no need for a PE node that don't have an IP-VRF instance to advertise the RT-5G routes here.

6.3. RT-5L Advertisement

When R1/R2 establish CE-BGP sessions with both PE1 and PE2, it is enough for PE1/PE2 to advertise RT-5L routes to PE3. There is no need for RT-5G or RT-5E advertisement on PE1/PE2 in that usecase.

Note that when R1/R2 establish CE-BGP sessions with both PE1 and PE2, the downlink VRF-interface addresses on PE1 and PE2 may be different IP addresses of the same subnet.

Note that when centerlized CE-BGP session is used, the prefixes from R3 and the local loopback addresses on PE3 are advertised to PE1/PE2 using RT-5L too.

7. Load Balancing of Unicast Packets

It is similar to [I-D.sajassi-bess-evpn-ip-aliasing] except for a few notable exceptions as explained in section 6.2.3 and the following.

Note that when the encapsulation is VXLAN, PE3 will encapsulate the RMAC of the RT-2E route for corresponding GW-IP address. And the RMAC of PE1 MAY have the same value with the RMAC of PE2. This can be achieved by configuration. When a IP packet is encapsulated with a VNI label according to an IP-AD/EVI route, the packet SHOULD be encapsulated with a Destination-MAC according to the RMAC of the same IP-AD/EVI route, if and only if the IP-AD/EVI route have a RMAC of its own.

Note that PE1/PE2 just do egress link protection following IP-AD/EVI and EAD/ES route. Even if ESI1 is configured as all-active ESI, PE1/PE2 will not load-balance between local downlink VRF-interface and the bypass tunnel. The downlink VRF-interfaces will always have more higher priority than the bypass tunnel.

8. Special Considerations for Single-Active ESIs

When the R1 is an Ethernet Segment of MHD type, and the uplink interfaces of R1 operates in linux network-bonding mode type 1. So the Primary flag according to DF election may cause packet-drop on R1 because of the nature of linux bond1.

In the linux bond1 use case, we propose that the Layer 2 extended community should not be included. and on PE3 the single-active ESI have lower priority than the MAC/IP route's own MPLS nexthop, but at the same time the downlink VRF-interface on PE1/PE2 may still have higher priority than the bypass tunnel to make convergency faster.

9. Security Considerations

This document does not introduce any new security considerations other than already discussed in [RFC7432] and [RFC8365].

10. IANA Considerations

There is no IANA consideration.

11. References

11.1. Normative References

[I-D.dawra-bess-srv6-services]
Dawra, G., Filsfils, C., Brissette, P., Agrawal, S., Leddy, J., daniel.voyer@bell.ca, d., daniel.bernier@bell.ca, d., Steinberg, D., Raszuk, R., Decraene, B., Matsushima, S., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay services", draft-dawra-bess-srv6-services-02 (work in progress), July 2019.

[I-D.ietf-bess-evpn-inter-subnet-forwarding]
Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-08 (work in progress), March 2019.

[I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., and S. Ramachandra, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-15 (work in progress), December 2019.

[I-D.sajassi-bess-evpn-ip-aliasing]
Sajassi, A., Badoni, G., Warade, P., Pasupula, S., Drake, J., and J. Rabadan, "L3 Aliasing and Mass Withdrawal Support for EVPN", draft-sajassi-bess-evpn-ip-aliasing-01 (work in progress), March 2020.

[I-D.wang-bess-evpn-context-label]
Wang, Y. and B. Song, "Context Label for MPLS EVPN", draft-wang-bess-evpn-context-label-02 (work in progress), June 2020.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[RFC8679] Shen, Y., Jeganathan, M., Decraene, B., Gredler, H., Michel, C., and H. Chen, "MPLS Egress Protection Framework", RFC 8679, DOI 10.17487/RFC8679, December 2019, <https://www.rfc-editor.org/info/rfc8679>.

11.2. Normative References

[RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <https://www.rfc-editor.org/info/rfc8214>.

Authors' Addresses

Yubao(Bob) Wang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: yubao.wang2008@hotmail.com

Zheng(Sandy) Zhang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: zzhang_ietf@hotmail.com