BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: 2 September 2026

A. Sajassi
C. Wang
K. Ananthamurthy
Cisco
J. Rabadan
Nokia

1 March 2026

EVPN L3-Optimized IRB
draft-sajassi-bess-evpn-l3-optimized-irb-04

Abstract

Ethernet VPN Integrated Routing and Bridging (EVPN-IRB) provides dynamic and efficient intra and inter-subnet connectivity among Tenant Systems and end devices while maintaining very flexible multihoming capabilities. This document describes how EVPN-IRB can be optimized for IP hosts and devices such that PE devices only maintain MAC addresses for locally-connected IP hosts, thus improving MAC scalability of customer bridges and PE devices significantly. This document describes how such optimization is acheived while still supporting host moblity which is one of the fundamental features in EVPN and EVPN-IRB. With such optimization PE devices perform routing for both intra and inter-subnet traffic which results in some some caveats that operators and service providers need to be fully aware of.

Table of Contents

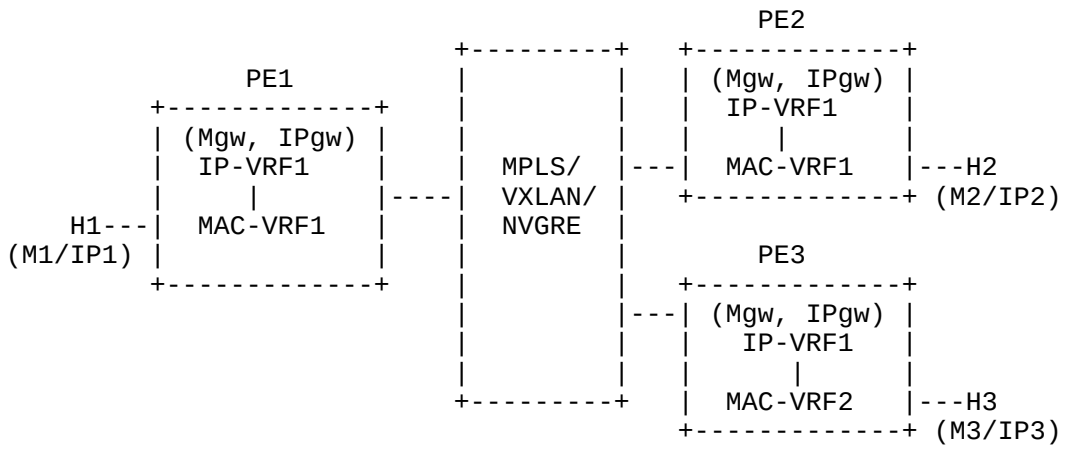
1. Introduction

Ethernet VPN Integrated Routing and Bridging (EVPN-IRB, [RFC9135] and [RFC9136]) provides dynamic and efficient intra and inter-subnet connectivity among Tenant Systems and end devices while maintaining very flexible multihoming capabilities. This document describes how EVPN-IRB can be optimized for IP hosts and devices such that PE devices only maintain MAC addresses for locally-connected IP hosts, thus improving MAC scalability of customer bridges and PE devices significantly. This document describes how such optimization is acheived while still supporting host moblity which is one of the fundamental features in EVPN ([RFC7432]) and EVPN-IRB ([RFC9135] and [RFC9136]). With such optimization PE devices perform routing for both intra and inter-subnet traffic which results in some some caveats that operators and service providers need to be fully aware of.

In some use cases, it is required to limit the number of MAC addresses learned in Customer Edge bridges connected to PE devices. These CE bridges can maintain a limited number of MAC addresses and thus when a subnet is stretched across one or more Enterprise or SP networks, the CE bridge needs to learn all MAC addresses in that stretched subnet for EVPN PE devices operating in Bridging mode or IRB mode. EVPN L3-Optimized IRB solution described in this document, limits the number of MAC addresses learned by CE bridges, connected to their local EVPN PEs, to a single MAC and that is the PE’s anycast MAC address associated with the IRB interface for that subnet or VLAN; therefore, significantly reducing the number of MAC addresses that are needed to be learned by a CE bridge. Of course, this assumes that most hosts aggregated by CE bridges are IP hosts.

In some other use cases, it is highly desireable to enable L3-only policy and QoS for both intra and inter-subnet traffic of IP hosts when PEs operate in EVPN IRB mode while maintaining host mobility - i.e., to avoid turning on L2 features such as L2 QoS, L2 ACL, L2 Policy forwarding, etc. The assumption is that by turning L3 features only, the operator can simplify the operation of their networks by avoiding enablement of both L2 and L3 features simultenously. In other words, with certain assumptions and caveats that are described later, PEs running in EVPN IRB mode, can run in Routed-only mode to enable L3-only features.

In the Figure 1 below, H1 and H2 are in the same MAC-VRF/subnet and H3 is in a different MAC-VRF/subnet. According to [RFC9135], the intra-subnet traffic between H1 and H2 is bridged and the inter-subnet traffic between H1 and H3 is routed. With L3-Optimized IRB solution, the intra-subnet traffic between H1 and H2 is routed as well.



IRB Gateway Anycast MAC: Mgw
IRB Gateway Anycast IP: IPgw

Figure 1: IRB Model with Distributed IRB Gateways

1.1. Terminology

* AC: Attachment Circuit

* DAD: Duplicate Address Detection

* EVPN: Ethernet VPN

* IRB: Integrated Routing and Bridging

* L2FIB: MAC-VRF Forwarding Table

* L2RIB: MAC-VRF Routing Table

* L3FIB: IP-VRF Forwarding Table

* L3RIB: IP-VRF Routing Table

* MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.

* PE: Provider Edge Device

1.2. Caveats to Consider

EVPN L3-Optimized IRB solution provides MAC scalability benefits for both CE bridges and PE devices as mentioned previously. Traffic from a connected host destined to a host in the same subnet will be routed at Layer-3 by the PE. This is made possible through the proxy ARP/ND action described in details in subsequent sections.

Packets that are forwarded in this manner undergo TTL decrements at the ingress and egress PE devices. They also undergo an outer MAC header rewrite such that the source address is the PE's IRB interface MAC address (i.e., overlay gateway anycast MAC address). This results in certain restrictions due to the changed forwarding semantics in supporting certain types of traffic and applications.

As a result of the changed forwarding semantics, following network traits are impacted for IP packets forwarded within a subnet:

* Multiple TTL decrements within subnet: Applications that depend on TTL=1 to control traffic to remain within subnet will not work with this mode of operation.

* Source MAC Rewrite: Due to the routing semantics, source address is rewritten with the PE's IRB interface MAC address (i.e., overlay gateway anycast MAC address). This breaks an assumption about the traffic within subnet: If an application depends on SMAC for some identification of a host then it might see a common MAC for many hosts within a subnet.

* Subnet broadcast will not work: In fact any unknown IP traffic is dropped or sent to CPU (glean) to trigger an ARP/ND or install a route.

* IPv6 link local and DaD: Both requires hosts within same subnet a layer2 reachability.

* Duplicate MAC detection within subnet will not work.

* Static ARP/ND configuration, or anything that avoids ARP/ND process will not work.

* IPv4 Address Conflict Detection as specified in [RFC5227], given that it uses ARP probes that are flooded. An ARP probe is an ARP Request constructed with an all-zero sender IP address that may be used by hosts for IPv4 Address Conflict Detection as specified.

* Neighbor Unreachability Detection, as per [RFC4861], in interoperability scenarios where the solicitation is received from a traditional-IRB PE.

Certain environments where nature of applications is well known can benefit from this mode of operation by gaining the scale offered by this solution. However, it is expected that the operator or the service provider is fully aware of these caveats and when enabling this functionality, they enable it on all relevant PE devices in order to get the correct and consistent behavior for the subnet on all PE devices across the fabric and the network.

2. Solution Overview

The solution described in this document is called EVPN L3-Optimized IRB and achieved by a simple modification to the existing EVPN IRB, i.e., by terminating ARP/ND messages received from locally connected hosts (e.g., local AC and not via EVPN network) which is also referred to as Local Proxy ARP/ND. In other words, when a PE is configured to operate in L3-Optimized IRB mode for a subnet (i.e., for a VLAN), then the PE acts as a router for that subnet by performing one of the following two options for the received ARP Request / Neighbor Solicitation message from an IP host:

Option A: Unconditional ARP Reply / Neighbor Advertisement and host discovery via data packet gleaning

1. It replies unconditionally to the ARP Request / Neighbor Solicitation message received from the locally connected host with its own anycast IRB MAC address as Sender MAC address in the ARP Reply / Neighbor Advertisement message.

2. It initiates a glean procedure upon receiving the first data packet with a miss IP destination address (DA) lookup by punting the packet to the control path (e.g., CPU) and generating a new ARP Request / Neighbor Solicitation for the missed IP DA.

Option B: Conditional ARP Reply / Neighbor Advertisement and host discovery via ARP Request / Neighbor Solicitation Optimization

1. It replies to the ARP Request / Neighbor Solicitation message received from the locally connected host with its own anycast IRB MAC address as Sender MAC address in the ARP Reply / Neighbor Advertisement message, ONLY IF the target host is known to the PE.

2. If the target host is unknown, the PE will NOT respond to the ARP Request / Neighbor Solicitation immediately, instead, the PE re-originates the ARP Request / Neighbor Solicitation with its own anycast IRB MAC and IP addresses as the Sender MAC and IP addresses to discover the target host. Once the target host is learned via EVPN RT-2 route, the PE can respond to the original ARP Request / Neighbor Solicitation or the next ARP Request / Neighbor Solicitation from host if the original one is expired on the PE.

Option C: Conditional ARP Reply and host discovery via BGP Control Plane

1. Option C follows the same procedures as described in Option B with one addition on the remote host discovery.

2. On the Fabric side, if the target host is unknown, the PE can originate a Host Discovery Route (as described later) and advertises it in BGP Control Plane to all the PEs in the same stretched Layer 2 topology to discover the target host.

These three options all have pros and cons. Option A leverages the regular Local Proxy ARP/ND functionality for ARP Reply / Neighbor Advertisement and host discovery, which is simple and straight forward to implement, but it may suffer performance impact due to the first data packet loss caused by data plane gleaning to discover the target host. Option B integrates the Local Proxy ARP/ND with EVPN control plane, which is relatively more complicated to implement, but it solves the first data packet loss issue, which might be critical for some applications. Option C utilizes BGP control plane for host discovery, so it gives the flexibility to turn off the Fabric Layer 2 Forwarding which would greatly improve the fabric port scalability, but it’s the most complicated option to implement.

Option B is the recommended approach which takes advantage of EVPN control plane to solve the first data packet loss issue. Choosing between Option A or Option B is a local matter on the PE which won’t introduce any interoperability problem, since eventually the PE would send an ARP Request / Neighbor Solicitation from its IRB interface to discover the target host, no matter it’s originated in Option A or re-originated in Option B.

The procedure described in this section for ingress PE is that of a typical router executes upon receiving an ARP Request / Neighbor Solicitation message with one additional enhancement with respect to the processing of a received EVPN MAC/IP route where the receiving PE does not populate MAC-VRF Forwarding Table (L2FIB), but it populates MAC-VRF Routing Table (L2RIB), IP-VRF Routing Tables (L3RIB) and IP-

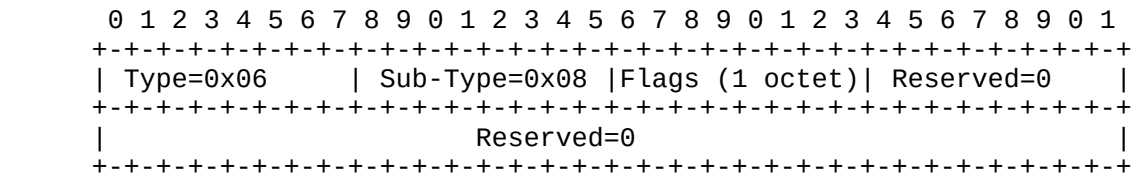
VRF Forwarding Tables (L3FIB). Since there is no L2 forwarding, there is no need for populating L2FIB; however, L2RIB needs to be populated for host mobility procedures because host mobility in EVPN is based on MAC mobility which is tracked in L2RIB.

When a PE operates in EVPN L3-Optimized IRB mode, it advertises a MAC/IP Advertisement route (aka route-type 2) along with a flag (via BGP extended community) to indicate this mode of operation so that the receiving PE adds the received MAC address to the L2RIB table but not the L2FIB. As it will be seen in Interop section, such flag is needed to ensure backward compatibility and seamless interoperability for brownfield deployment. If there is no such flag, then the received MAC address is added to both the L2RIB and the L2FIB.

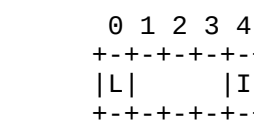
When operating in L3-Optimized IRB mode, the PE SHOULD NOT advertise an EVPN Type-2 route with MAC address only but instead it SHOULD wait and advertise an EVPN Type-2 route with both MAC & IP addresses. If the PE avertises two EVPN Type-2 routes (one with MAC address only and another with both MAC & IP addresses), then it MUST advertise both routes with the L3-Optimized flag.

When a PE receives an ARP/ND request and decides to discover the target host in BGP control plane (Option C), it advertises a "Host Discovery Route" with the Target Host IP Address and the MAC-VRF Route-Target in the stretched Layer 2 topology. The format of the "Host Discovery Route" will be defined in the future.

This "L3-Optimized IRB flag" can be carried in an extended flag field in "EVPN ARP/ND Extended Community" (RFC 9047).



Flags field:



R: Router flag (corresponds to Bit 23 of the EC)
O: Override flag (corresponds to Bit 22 of the EC)
I: Immutable ARP/ND Binding flag (corresponds to Bit 20 of the EC)

Proposed New flag (TBD)
L: L3-Optimized IRB flag (corresponds to Bit 16 of the EC)

Figure 2: EVPN ARP/ND Extended Community

EVPN L3-Optimized IRB shall operate seamlessly with all existing EVPN baseline features such as:

* All-Active and Single-Active Multi-Homing

* Aliasing

* Proper BUM filtering using DF election

* Host (MAC) Mobility

Furthermore, EVPN L3-Optimized IRB shall support the following services and deployment scenarios:

* EVPN IRB Multicast Service

* EVPN IRB E-Tree Service

* Greenfield deployment where all EVPN PEs operate in L3-Optimized mode

* Brownfield deployment where some EVPN PEs operate in L3-optimized IRB mode and the rest operate in EVPN Traditional IRB mode

2.1. ARP Message Handling

The following subsections describe ingress and egress PEs behaviors in details along with the corresponding ladder diagrams for the following:

* ARP Request from an IP host
* Host discovery via data packet gleaning (Option A)
* Host discovery via ARP Request Optimization (Option B)
* Host discovery via BGP Control Plane (Option C)
* ARP response from an IP host
* Gratuitous ARP from an IP host

2.1.1. ARP Request from an IP Host

The following steps in Figure 3 describe in detail the system behavior (procedures on ingress and egress PEs) upon receiving an ARP Request message from an IP host:

1. Host H1 ARP for host H2 MAC address.
2. PE1 receives the ARP Request broadcast message from H1, and it terminates it on its IRB interface associated with that subnet i.e., it punts the message to its CPU and does not flood the message in the Broadcast Domain. The punting should be done for ARP broadcast messages and not unicast messages. If ARP Request message is a unicast message with MAC DA different than that of IRB interface, then this ARP Request message should get bridged and not punted (and if punted then it needs to get forwarded as is). This ensures backward compatibility with traditional-IRB PEs as described in Interoperability section. If H1 MAC address and IP are learned for the first time, then PE1 populates L3RIB and L3FIB with the H1 IP address, L2RIB and L2FIB with H1 MAC address, and ARP table with H1 <MAC, IP> addresses. PE1 also advertises H1 MAC and IP addresses in EVPN MAC/IP route with a flag indicating L3-Optimized IRB operation.

3. With Option A, PE1 generates an unconditional ARP Response message with the Anycast MAC address of its IRB interface as the Sender MAC address and also in the outer Source MAC address. Then, it sends the message to H1. With Option B, PE1 only responds if the Target host (H2) is known on PE1. If the target host is unknown, PE1 will NOT respond the ARP Request and follow the ARP Request Optimization procedure as defined in the later section.

4. When PE2 receives the EVPN MAC/IP route, it populates its L3RIB and L3FIB. Then, it checks for the L3-Optimized-IRB flag, if the flag is set, then it populates the L2RIB (for new MAC address) but not the L2FIB. PE2 does NOT populate its L2FIB because the forwarding is performed in only L3 (packets are IP routed for both inter and intra subnet traffic). The reason L2RIB is populated is for mobility procedure as described before. However, if the flag is not present or is not set, then it populates both the L2RIB and L2FIB as for traditional IRB. In case of receiving multiple MAC/IP routes for the same MAC, the EVPN best path selection gets executed and the L3 optimized IRB flag is processed only in case the flag is set in the selected routes.

5. If PE2 realizes that this is not a new MAC (and IP) address but rather a MAC move because the received sequence number from EVPN MAC/IP route is higher than locally stored sequence number, then after sending an ARP probe to the host and ensuring that the host is no longer present locally, it performs mobility procedure and update the adjacency for that MAC in the L2RIB to point to the remote PE. It also deletes that MAC from its L2FIB if the MAC was learned locally. If the MAC is not advertised with the L3-Optimized IRB flag, then the adjacency for that MAC is also updated in the L2FIB as for traditional IRB since in such cases Intra-subnet forwarding is performed using bridging (as opposed to routing) to ensure backward compatibility.

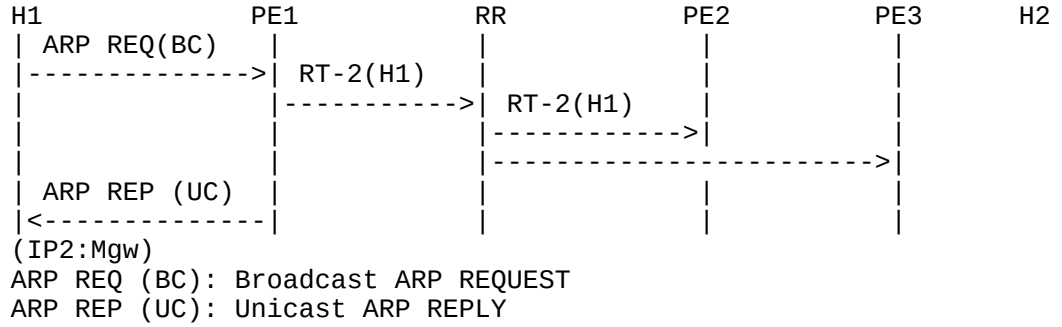


Figure 3: ARP Request from an IP Host

2.1.2. Host discovery via data packet gleaning (Option A)

The following steps in Figure 4 describe in detail the system behavior (procedures on ingress and egress PEs) upon receiving the first data packet from an IP host destined to another IP host with a miss IP destination lookup:

1. Host H1 sends its first data packet destined to host H2 with DMAC of Anycast-IRB-Interface MAC address.

2. PE1 does route lookup. If host H2’s IP address is known to PE1, then PE1 forwards the packet accordingly.

3. If host H2’s IP address is unknown to PE1 thus resulting in a lookup miss, then PE1 performs the longest-match prefix lookup for H2’s IP address which results in glean adjacency for that prefix and the packet is punted to the CPU.

4. PE1’s CPU for glean adjacency, initiates ARP procedure by generating an ARP Request message with its own Anycast IRB MAC and IP addresses as Sender MAC and IP addresses.

5. PE1 sends its ARP Request message over all the local interfaces for that bridge domain (BD), over its virtual PW interfaces (if any), and over its core-facing interface. Since the glean packet is received from a local interface, PE1 uses source-interface filtering to ensure that the ARP request packet is not sent back over the same interface from which it received the data packet.

6. When remote PEs (PE2 and PE3) receive this ARP Request message, they forward it over their physical or virtual (PW) interfaces. The ARP Request message is not punted to the CPU -- i.e., "punt" action is enabled on access interfaces (physical or virtual) but not on core-facing interface.

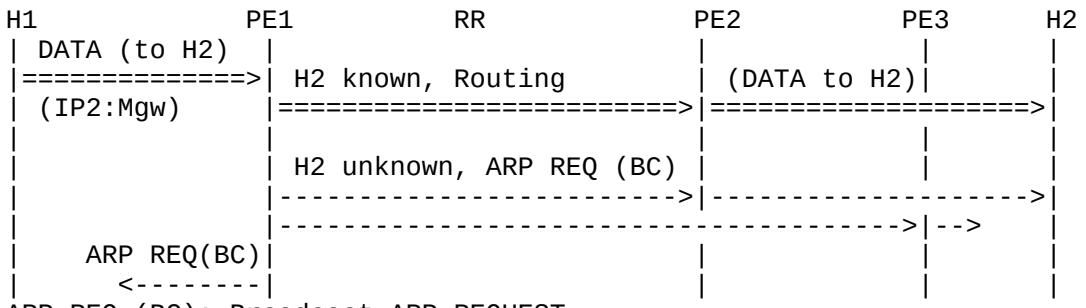


Figure 4: Host discovery via data packet gleaning (Option A)

2.1.3. Host discovery via ARP Request Optimization (Option B)

The following steps in Figure 5 describe in detail the system behavior (procedures on ingress and egress PEs) of host discovery via ARP Request Re-Origination if the target host is unknown to the ingress PE:

1. Since the target host (H2) is unknown to PE1, PE1 re-originates an ARP Request message with its own Anycast IRB MAC and IP addresses as Sender MAC and IP addresses and H2 as the Target IP Address.

2. PE1 sends its ARP Request message over all the local interfaces for that bridge domain (BD), over its virtual PW interfaces (if any), and over its L2-stretch (core-facing) interface. Since the original ARP Request packet is received from a local physical interface, PE1 uses source-interface filtering to ensure that the Re-originated ARP request packet is not sent back over the same interface.

3. When remote PEs (PE2 and PE3) receive this ARP Request message, they forward it over their physical or virtual (PW) interfaces. The ARP Request message is not punted to the CPU -- i.e., "punt" action is enabled on access interfaces (physical or virtual) but not on L2-stretch interface.

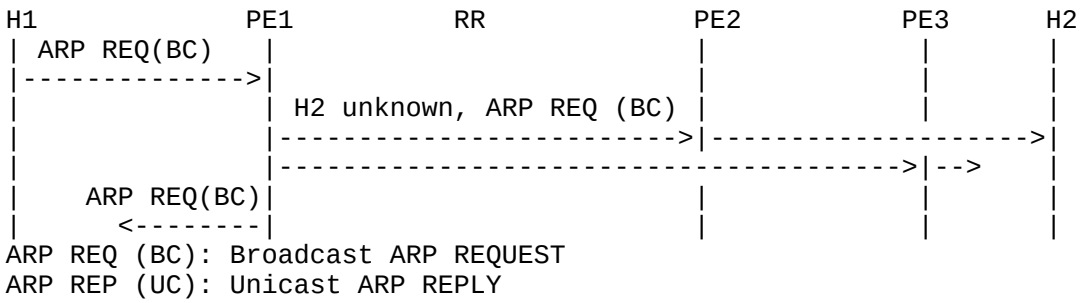


Figure 5: Host discovery via ARP Request Optimization (Option B)

2.1.4. Host discovery via BGP Control Plane (Option C)

The following steps in Figure 6 describe in detail the system behavior (procedures on ingress and egress PEs) of host discovery via BGP Control plane if the target host is unknown to the ingress PE:

1. Since the target host (H2) is unknown to PE1, PE1 re-originates an ARP Request message with its own Anycast IRB MAC and IP addresses as Sender MAC and IP addresses and H2 as the Target IP Address.

2. PE1 sends its ARP Request message over all the local interfaces for that bridge domain (BD), over its virtual PW interfaces (if any), and over its L2-stretch (core-facing) interface (If it’s not turned off). This is the same behavior as Option B.

3. On the fabric side, PE1 originates a "Host Discovery Route" with the Target Host IP Address and the corresponding MAC-VRF Route target.

4. PE1 starts a Host Discovery Timer for this advertised route with a default value of 30 seconds. If the timer expires and the target host is still not learned by PE1, the PE1 stops the timer and withdraws the route.

5. When remote PEs (PE2 and PE3) receive this "Host Discovery Route", they would import it into local MAC-VRF, retrieve the Target IP Address from the route, and initiate an ARP Request from their local IRB interfaces to discover the target host, and flood the ARP Request locally, if the ARP process is not initiated yet.

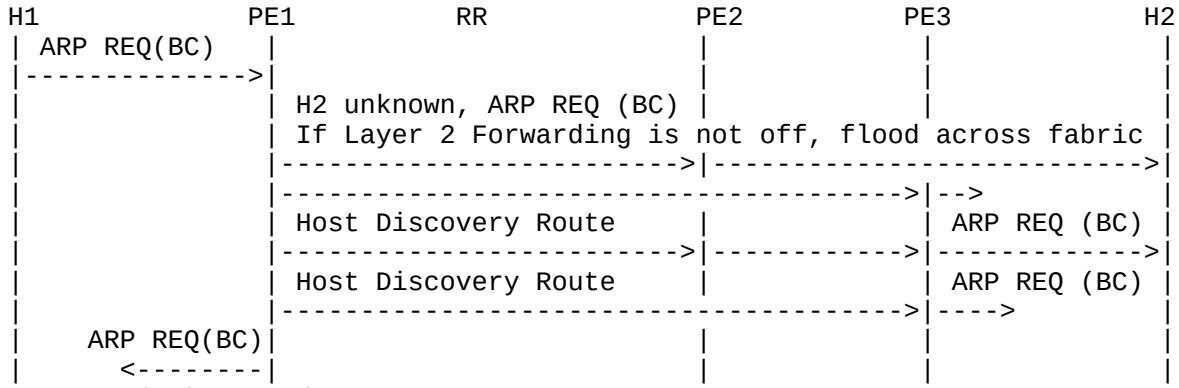


Figure 6: Host discovery via BGP Control Plane (Option C)

2.1.5. ARP Response from an IP host

The following steps in Figure 7 describe in detail the system behavior (procedures on ingress and egress PEs) upon receiving the ARP Response message from the remote host:

1. Host H2 sends its ARP Response message with Anycast-IRB-MAC and Anycast-IRB-IP addresses as its target addresses.

2. PE2 receives this message and if H2’s MAC and IP addresses are new, it populates its ARP cache table, its MAC & IP FIB tables, and its MAC & IP RIB tables. Next, it sends the corresponding EVPN MAC/IP Advertisement route along with a flag indicating L3-Optimized IRB mode.

3. When PE1 receives the EVPN MAC/IP route, it populates its L3RIB and L3FIB. Then, it checks for the L3-Optimized-IRB flag, if the flag is set, then it populates the L2RIB (for new MAC address) but not the L2FIB. However, if the flag is not present or is not set, then it populates both the L2RIB and L2FIB as for traditional IRB. PE1 does NOT populate its L2FIB because the forwarding is performed in only L3 (packets are IP routed for both inter and intra subnet traffic). The reason L2RIB is populated is for mobility procedure as described before.

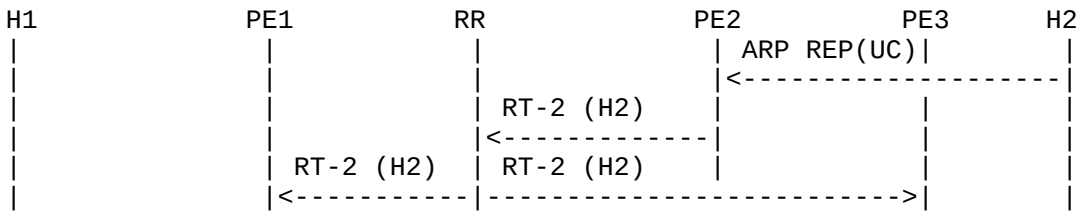


Figure 7: ARP Response from an IP host

Note: For Option B, once H2 is learned by PE1 via EVPN RT-2, PE1 may respond to the original ARP Request if it's not expired or respond to the next ARP request from H1. For Option C, once the target IP is learned, the Host Discovery Route is withdrawn by the advetising PE.

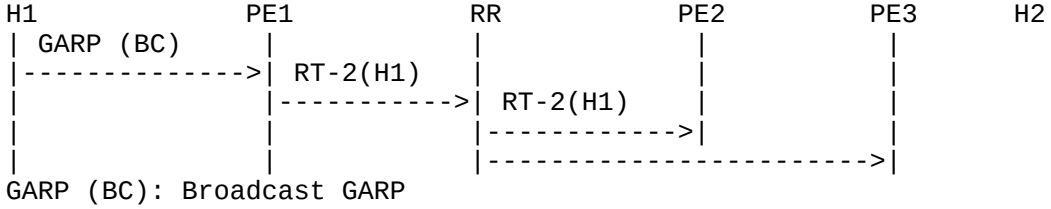
2.1.6. Gratuitous ARP from an IP host

The following steps in Figure 8 describe in detail the system behavior (procedures on ingress and egress PEs) upon receiving the Gratuitous ARP message from an IP host:

1. Host H1 sends a Gratuitous ARP broadcast message with target IP address of its own.

2. PE1 receives this message and if H1's MAC and IP addresses are new, it populates its ARP cache table as well as MAC and IP RIB and FIB tables accordingly and sends the corresponding EVPN MAC/IP Advertisement route along with a flag indicating L3-Optimized IRB mode. PE1 does not generate a Gratuitous ARP message with its Anycast-IRB addresses as sender's addresses.

3. When PE2 receives the EVPN MAC/IP route, it populates its L3RIB and L3FIB. Then, it checks for the L3-Optimized-IRB flag, if the flag is set, then it populates the L2RIB (for new MAC address) but not the L2FIB. However, if the flag is not present or is not set, then it populates both the L2RIB and L2FIB as for traditional IRB. PE2 does NOT populate its L2FIB because the forwarding is performed in only L3 (packets are IP routed for both inter and intra subnet traffic). The reason L2RIB is populated is for mobility procedure as described before.



GARP (BC): Broadcast GARP

Figure 8: Gratuitous ARP from an IP host

2.2. Neighbor Discovery Message Handling

In case of IPv6 (or dual stack) EVPN IRB, the ingress and egress PE behaviors are generally the same as the ones specified in Section 2.1 for ARP. However, there are some specific considerations for IPv6 Neighbor Discovery, as follows:

* The procedures for Neighbor Solicitation messages from an IP host follow the procedures for ARP Requests in Section 2.1.1, only that Neighbor Solicitation messages use multicast destination MAC addresses as opposed to broadcast in ARP Requests.

* The Neighbor Advertisement message generated by PE1 (with the Anycast MAC address) in response to a Neighbor Solicitation from H1 (for H2's IP address) always set to 1 the Flags R (Router Flag), O (Override Flag) and S (Solicited Flag), irrespective of the type of host H2 is.

* The discovery of IPv6 hosts follow the procedures in Section 2.1.2 and Section 2.1.3.

* The processing of the Neighbor Advertisement response from an IPv6 host, follows Section 2.1.5. In addition, when the egress PE generates an EVPN MAC/IP Advertisement route with the L3-Optimized IRB flag set, it MUST also set the R (Router) and O (Override) flags to zero or one, as per [RFC9047].

* The processing of unsolicited Neighbor Advertisement messages from hosts are handled as in Section 2.1.6 for the Gratuitous ARP messages from the hosts. In addition to setting the L3-Optimized IRB mode flag for the EVPN MAC/IP Advertisement routes, the PE will also modify the R and O flags as per [RFC9047].

3. Deployment Scenarios

The deployment scenarios in this section are specified for IPv4 Address Resolution Protocol. For IPv6 Neighbor Discovery, the DAD and NUD procedures will not work in hybrid deployments, however, the IPv6 address resolution aspects handled by multicast Neighbor Solicitation messages and their responses using solicited unicast Neighbor Advertisement messages will follow the procedures in this section.

When considering deployment scenarios, both greenfield and brownfield must be considered. For greenfield scenarios where all PEs are L3-Optimized-IRB PEs, procedures of Section 2.1 are applicable as-is and

the control-plane and data-plane flows are as depicted in that section. However, for brownfield deployment, there are some changes to control-plane and data-plane flows as described below and the following subsections concentrate on data-plane and control-plane flows for brownfield deployments.

EVPN IRB has been in deployment for many years; therefore, it is important to ensure backward compatibility with existing EVPN IRB PEs when L3-Optimized IRB is introduced into an existing network. Such backward compatibility and seamless interoperability with existing EVPN IRB, ensures gradual migration of PE devices in an EVPN IRB network with this feature.

As it will be seen, L3-Optimized IRB PEs can easily interoperate with existing IRB PEs as-is. In otherwords, L3-Optimized IRB PEs can be inserted into an existing network with traditional EVPN PEs (either IRB or just L2), and they can work seamlessly without the need for any gateway devices. Since no gateway devices are required for such interoperabilty, this can facilitate brownfield deployment of L3-Optimized-IRB PEs.

In traditional EVPN IRB, the intra-subnet traffic (traffic within the same subnet) is forwarded using bridging, whereas, in L3-Optimized IRB, the intra-subnet traffic is forwarded using routing. The following section describes in terms of control and data plane operations how this inter-operability works when for a given subnet some PE devices operate in L3-Optimized IRB while some other PE devices operate in traditional IRB.

3.1. Control-Plane Operation

As shown in the following figures, no changes to the control-plane are needed for this inter-operability. The traditional-IRB PEs operate as before and the new L3-Optimized-IRB PEs do not require any new functionality on top of what has already been described in the previous sections. The following just list some of the salient points for such interoperability.

1. ARP Request broadcast messages arriving from ACs (either physical or virtual) get punted to the CPU. The ARP Request broadcast messages from core-facing interface do not get punted to the CPU.

2. ARP Request unicast messages should not get punted to the CPU. If these messages get punted to the CPU, then the CPU should send them back to get bridged based on their MAC DA addresses.

3. When ARP Response message is generated by the CPU unconditionally, the sender MAC address is that of Anycast-IRB MAC address and the sender IP address is that of target IP address in ARP Request.

4. When ARP Request message is generated by the CPU as the result of glean procedure or re-originatation, both sender MAC and IP addresses are that of Anycast-IRB interface.

3.2. Data-Plane Operation

Intra-subnet traffic (traffic within a subnet or VLAN) among L3-Optimized-IRB PEs gets always routed and among traditional-IRB PEs gets always bridged. However, for such intra-subnet traffic exchanged between a L3-Optimized IRB PE and a traditional-IRB PE, majority of time it gets bridged except for the following case as listed below in Figure 9 and described in detail in interoperability section later.

1. Traffic is in the direction of Optimized-IRB PE toward traditional-IRB PE

2. Traditional-IRB PE operates with ARP suppression enabled; where it has MAC & IP addresses of a remote host in its ARP table so that when a local host sends an ARP Request for this remote host, the traditional-IRB PE can respond locally to this local host.

Under the above condition, the Optimized-IRB PE, attached to the remote host (H1), never receives and never forwards an ARP Request destined to the remote host and thus the remote host uses Anycast-IRB MAC address of the Optimized-IRB PE to send traffic to the local host (H2). And since Anycast-IRB MAC address is used, the traffic gets routed in that direction.

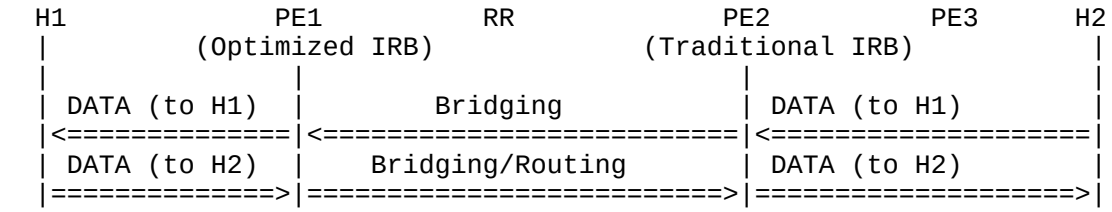


Figure 9: Traffic between Optimized and Traditional IRB PE

3.3. Interoperability Scenarios

When considering backward compatibility with EVPN IRB PEs, it is important to consider such interoperability with both traditional IRB PEs with and without ARP suppression since there can be deployments with such mixed of PEs. Since traditional IRB PEs can easily interoperate with IRB PEs with ARP suppression feature, when L3-Optimized-IRB PEs get inserted in such networks, these PEs must seamlessly interoperate with existing IRB PEs with and without ARP suppression feature.

Since L3-Optimized IRB support both routing and bridging for intra-subnet traffic and since traditional IRB PEs support only bridging for intra-subnet traffic, the traffic exchange from a traditional-IRB PE (with and without ARP suppression) to a L3-Optimized-IRB PE gets always settled in bridging mode (i.e., the common denominator forwarding mode). Furthermore, the traffic exchange from a L3-Optimized-IRB PE to a traditional-IRB PE without ARP suppression gets always bridged; whereas, the traffic exchange from a L3-Optimized-IRB PE to a traditional-IRB PE with ARP suppression can get routed if a host that sends an ARP request to its locally connected PE, gets an ARP response back right away because of ARP suppresion feature as shown in the use case for ARP suppression.

The following scenarios describe the interoperability between L3-Optimized-IRB PEs and traditional-IRB PEs. Furthermore, they illustrate when intra-subnet traffic is routed and when it is bridged.

1. ARP Request Originated by a L3-Optimized-IRB PE

2. ARP Request Originated by a Traditional IRB PE

3. Interop between Option C PE and PEs in other mode

3.3.1. ARP Request received by a Traditional-IRB PE

The following Figure 10 describes the scenario where an ARP Request message is first originated by a host connected to a traditional-IRB PE.

1. Host 2 sends an ARP Request broadcast message for host H1 MAC address.

2. Traditional-IRB PE2 receives the ARP Request broadcast message from H2, and it floods it over its local and core-facing interface. It also learns H2’s MAC address and advertises it in EVPN MAC/IP route.

3. PE1 and PE3 receive the ARP Request broadcast message over their core-facing interfaces and subsequently forward it over their local interfaces.

4. Host H1 receives this ARP Request message and adds H2 MAC & IP addresses to its ARP table and send an ARP Reply message to H2.

5. PE1 receives the ARP Reply from H1 and forwards it to PE2 (via either known unicast or unknown unicast packet). It also learns H1's MAC address and advertises it in EVPN MAC/IP route.

6. Host H2 upon receiving ARP Reply, updates its ARP table with MAC & IP addresses of H1.

7. Since both PE2 and PE1 have adjacency information for H1 and H2 MAC addresses, data traffic between H1 to H2 gets bridged via PE1 and PE2.

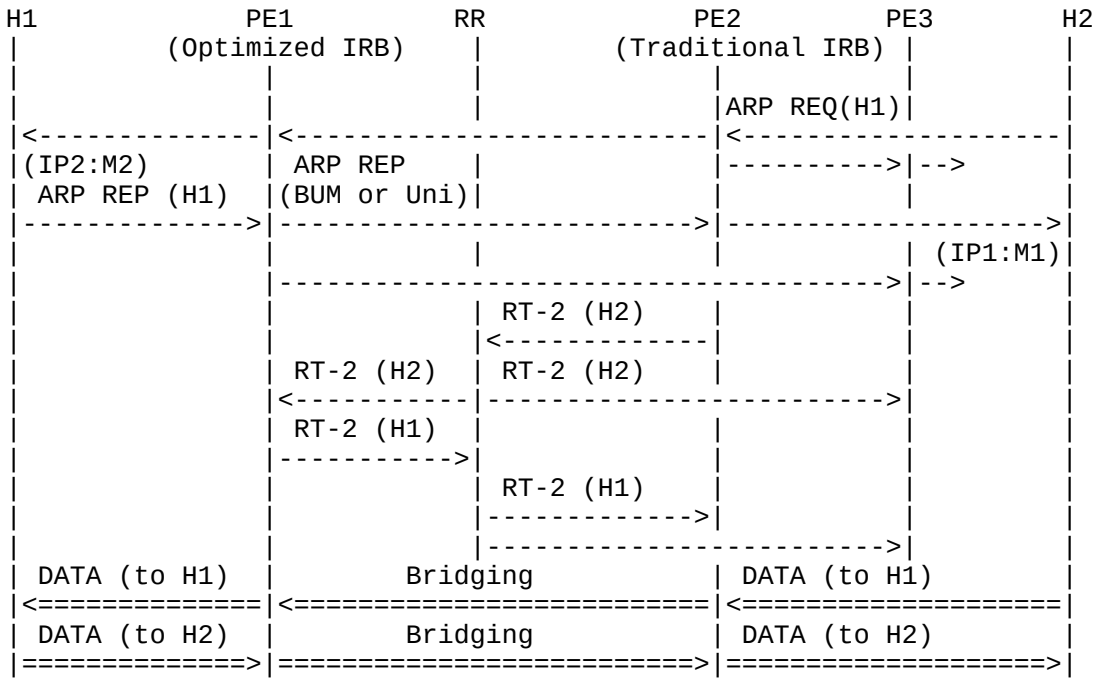


Figure 10: ARP Request received by a Traditional-IRB PE

3.3.2. ARP Request received by a L3-Optimized-IRB PE

The following Figure 11 describes the scenario where an ARP Request message is first originated by a host connected to a L3-Optimized-IRB PE with Option A (Unconditional ARP Response).

1. Host H1 ARP for host H2 MAC address.

2. L3-Optimized-IRB PE1 receives the ARP Request broadcast message from H1, and it terminates it on its IRB interface associated with that subnet and generates an unconditional ARP Response message with the anycast MAC address of its IRB interface as the sender MAC address and target IP address in ARP Request as the sender IP address.

3. L3-Optimized-IRB PE1 adds MAC & IP addresses of H1 to its ARP table, adds H1’s MAC to its L2 FIB and RIB table, and adds H1’s IP to its L3 FIB and RIB tables. It also advertises an EVPN MAC/IP route for H1’s MAC & IP addresses.

4. Host H1 receives this ARP response and adds H2 IP address along with anycast-IRB MAC address of PE1 to its ARP table.

5. When the PE1 receives the first data packet generated by H1 destined to H2, it performs an IP lookup for H2 which triggers the glean procedure and as the result PE1 generates an ARP Request message with its anycast-IRB MAC and IP addresses as sender MAC and IP and this message is forwarded in data-plane and it is received by H2.

6. H2, upon receiving this ARP Request, sends a reply to the anycast-IRB address which is received and terminated by the PE2. PE2 generates an EVPN MAC/IP Advertisement route for H2 MAC and IP addresses. When PE1 receives this advertisement, it adds H2 MAC and IP addresses to its RIBs and FIBs.

7. The next time H1 sends data traffic to H2, because H2 IP address is resolved in PE1, the packet is routed via PE1 and PE2 to H2.

8. When H2 wants to send data traffic to H1, it first sends an ARP Request for H1 which gets forwarded all the way to H1 as BUM traffic via PE2 and PE1.

9. Upon receiving this ARP Request message, H1 updates its ARP table to associate H2 MAC address (M2) with H2 IP address (IP2). This update overrides the previous association. H1 sends an ARP response which gets bridged by PE1 and PE2 all the way to H2.

10. All subsequent data traffic between H1 to H2 gets bridged via PE1 and PE2.

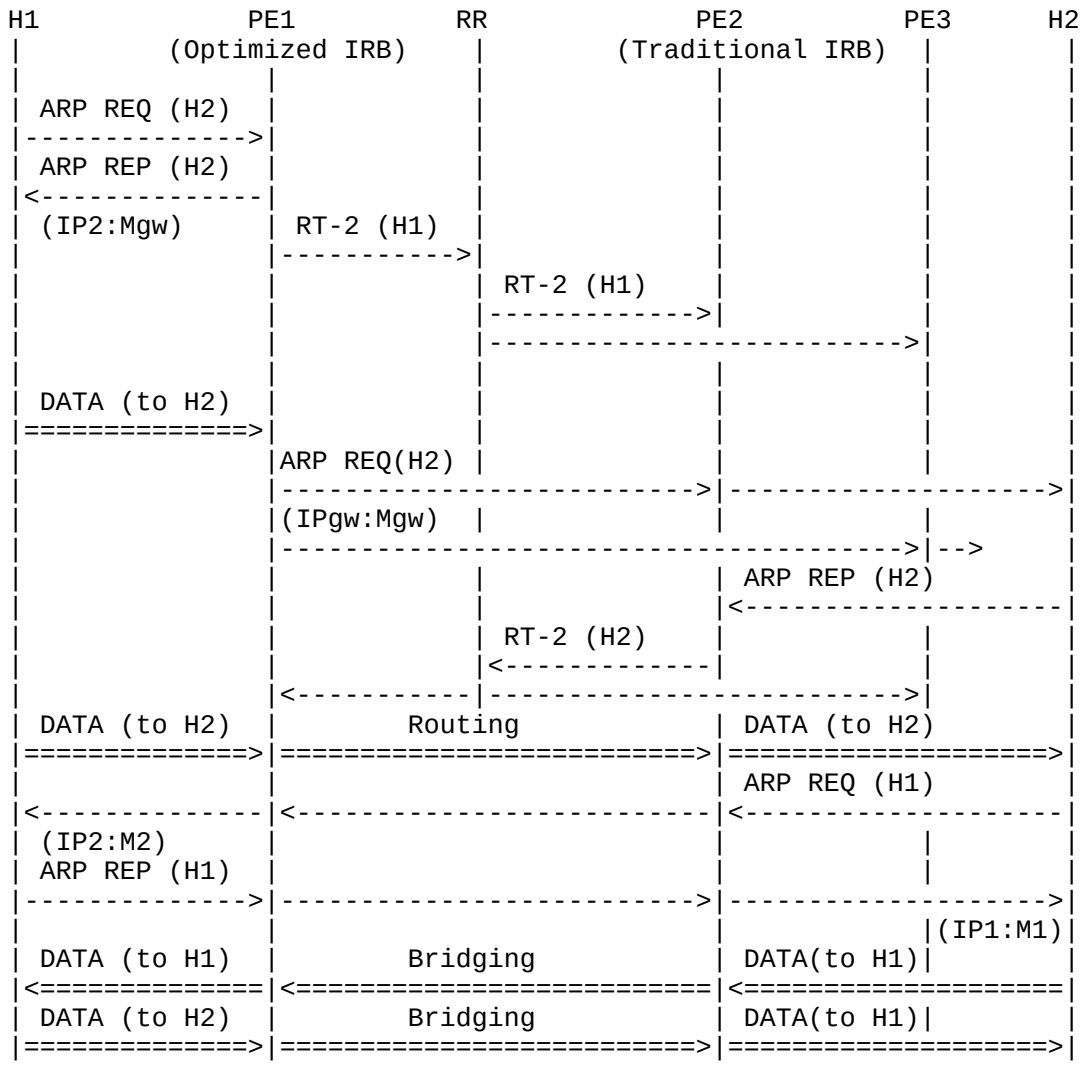


Figure 11: ARP Request received by a L3-Optimized-IRB PE

Note: The above procedure is described with Option A, but the end result will be the same for both options, since the only difference is whether the target host discovery is triggered from data packet gleaning or ARP Request Re-Origination.

Note: Due to the propagation delay of RT-2, the ARP messages might be exchanged earlier between the 2 hosts behind L3-Optimized IRB PE and Traditional IRB PE, which would end up the situation that the traffic between the 2 hosts get bridged only.

3.3.3. Interop between Option C PE and PEs in other mode

The following Figure 12 describes the scenario how an Option C PE interoperates with other PEs. When interoperating Non-Option-C PEs, the fabric Layer 2 Forwarding must be enabled.

1. Host H1 ARP for host H2 MAC address.

2. PE1 (Option C) does the same procedure as Option B, to re-originate the ARP Request (H2) and flood in the data plane locally and to the other PEs.

3. PE1 (Option C) originates a Host Discovery Route (H2) and floods in the BGP control plane to other PEs.

4. PE2 and PE3 receive the re-originated ARP Request from PE1 and flood it locally to discover H2.

5. PE2 receives the Host Discovery Route from PE1 and ignores it since it’s in Option B mode and it will rely on the ARP REQ in data plane to discover the host.

6. PE3 receives the Host Discovery Route from PE1 and ignores it as well, since it’s in Traditional IRB mode, and it will rely on the ARP REQ in data plane to discover the host.

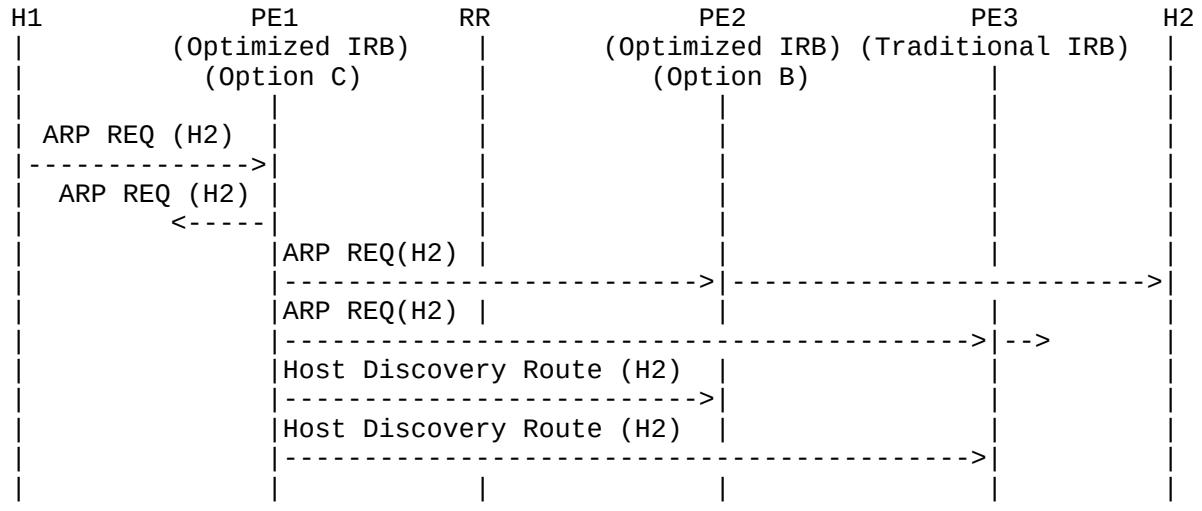


Figure 12: Interop between Option C PE and PEs in other mode

Note: An Option C PE may receive both ARP Request from Data Plane and Host Discovery Route from Control Plane. The receiving PE may check if the host discovery is already in progress to avoid concurrent local ARP handling or just to allow it.

4. Acknowledgements

The authors would like to thank Neeraj Malhotra, Mei Zhang, Lukas Krattiger, Ramchander Nadipally, and Rahul Kachalia for their reviews of this document and feedbacks.

5. Security Considerations

All the security considerations in [RFC7432] apply directly to this document because this document leverages the control and data plane procedures described in those documents.

This document does not introduce any new security considerations beyond that of [RFC7432] because advertisements and processing of Ethernet Segment route for vES in this document follows that of physical ES in those RFCs.

6. IANA Considerations

This document requests no actions from IANA.

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>.

[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>.

[RFC9047] Rabadan, J., Ed., Sathappan, S., Nagaraj, K., and W. Lin, "Propagation of ARP/ND Flags in an Ethernet Virtual Private Network (EVPN)", RFC 9047, DOI 10.17487/RFC9047, June 2021, <https://www.rfc-editor.org/info/rfc9047>.

[RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <https://www.rfc-editor.org/info/rfc9135>.

[RFC9136] Rabadan, J., Ed., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in Ethernet VPN (EVPN)", RFC 9136, DOI 10.17487/RFC9136, October 2021, <https://www.rfc-editor.org/info/rfc9136>.

7.2. Informative References

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <https://www.rfc-editor.org/info/rfc4861>.

[RFC5227] Cheshire, S., "IPv4 Address Conflict Detection", RFC 5227, DOI 10.17487/RFC5227, July 2008, <https://www.rfc-editor.org/info/rfc5227>.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Chuanfa Wang
Cisco
Email: chuanwan@cisco.com

Krishna Ananthamurthy
Cisco
Email: kriswamy@cisco.com

Jorge Rabadan
Nokia
Email: jorge.rabadan@nokia.com