draft-sajassi-bess-l3-optimized-irb-00.txt

A. Sajassi (Cisco), C. Wang (Cisco),
K. Ananthamurthy (Cisco)

IETF 118, November 2023
Prague

Problem Statement

- To alleviate MAC scale issue in CE bridges used to aggregate IP hosts into an EVPN network
  - i.e., to limit number of MAC addresses learned in CE bridges connected to EVPN PEs.
  - It also helps with MAC scale in PE devices (side benefit)
- To enable L3 only policy and QoS for both inter and intra subnet traffic of IP hosts when PEs operate in IRB mode while maintaining host mobility
  - i.e., to avoid turning on L2 features such as L2 QoS, L2 ACL, L2 Policy forwarding, etc. for intra-subnet traffic
  - To simplify operation by turn on L3 feature only!

EVPN L3-Optimized IRB

- **H1 & H2 are in one subnet, and H3 in a different subnet**
- **Per RFC 9135 communication between H1 & H2 is bridged**
- **Whereas communication between H1/H2 and H3 is routed**
- **With L3-Optimized IRB, communication between H1 & H2 is also routed**

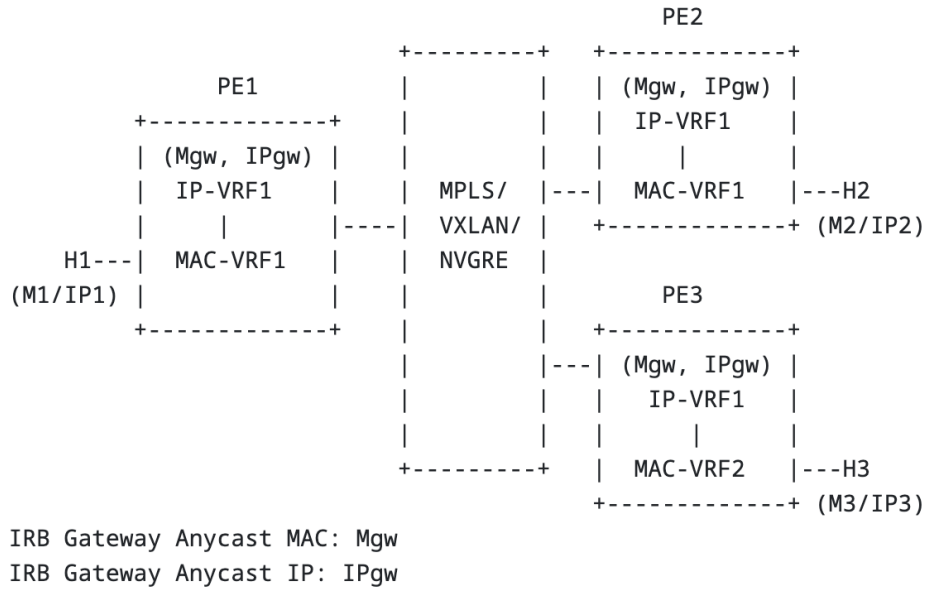


Figure 1: IRB Model with Distributed IRB Gateways

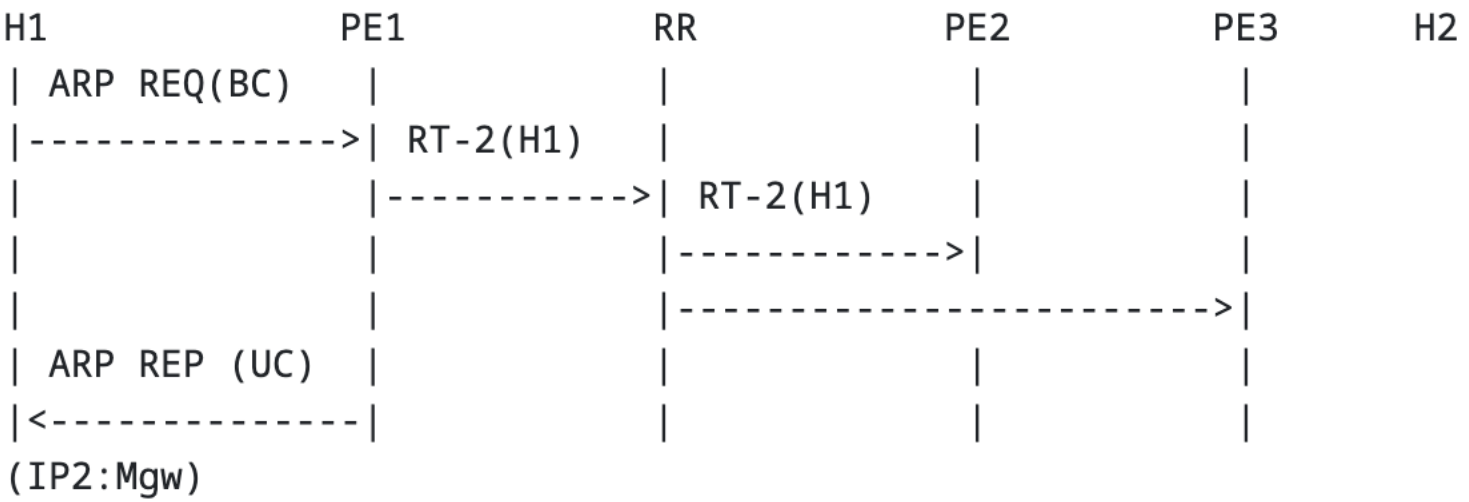
Caveats To Consider

- 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 or install a route.
- Static ARP configuration, or anything that avoids ARP process will not work.

Solution Overview

- This solution requires the ARP/ND messages to be terminated by the PE that receive these messages. I.e., PE acts a router for that subnet
  - It replies to the ARP Request message received from the locally connected host with its own anycast IRB MAC address as Sender MAC address in the ARP Reply message.
  - 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 for the missed IP DA.

ARP Message Handling



ARP REQ (BC): Broadcast ARP REQUEST
ARP REP (UC): Unicast ARP REPLY

Figure 2: ARP Request from an IP Host

ARP Message Handling – Cont.

- 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.
- PE1 generates an ARP Response message with the Anycast MAC address of its IRB interface as the Sender MAC address and sends the message to H1. PE1 also advertises H1 MAC and IP addresses in EVPN MAC/IP route with a flag indicating L3-Optimized IRB operation.
- 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.
- 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

First data packet from an IP host

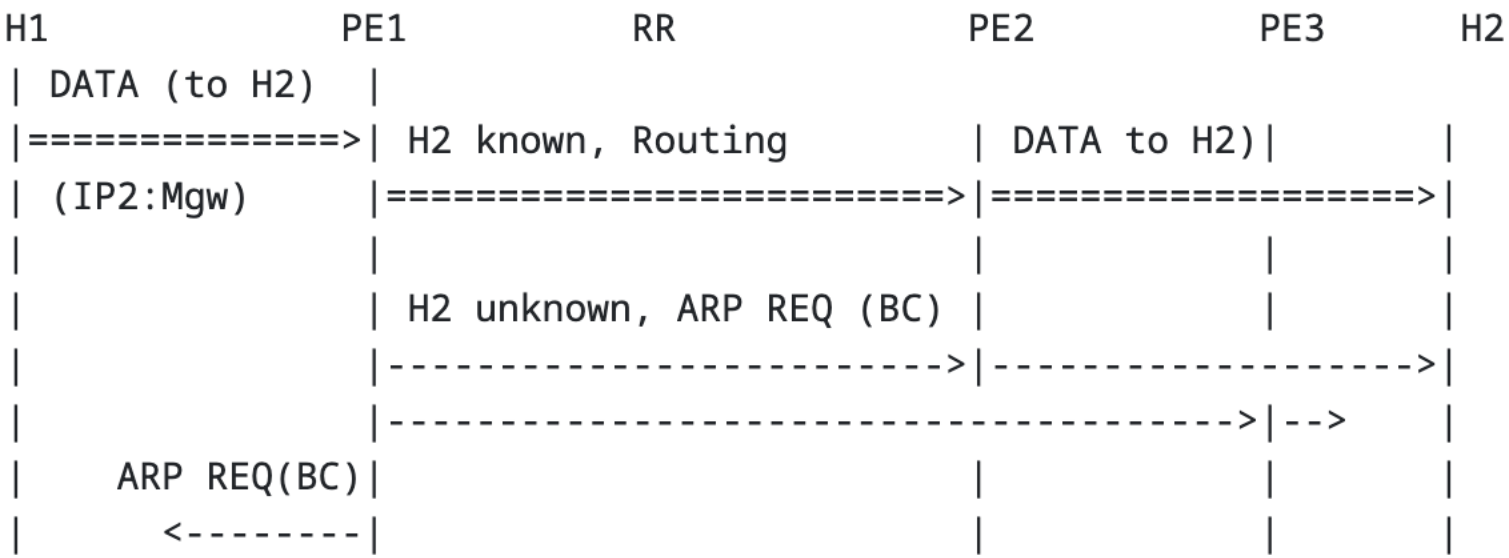


Figure 3: First data packet from an IP host

ARP Response Handling

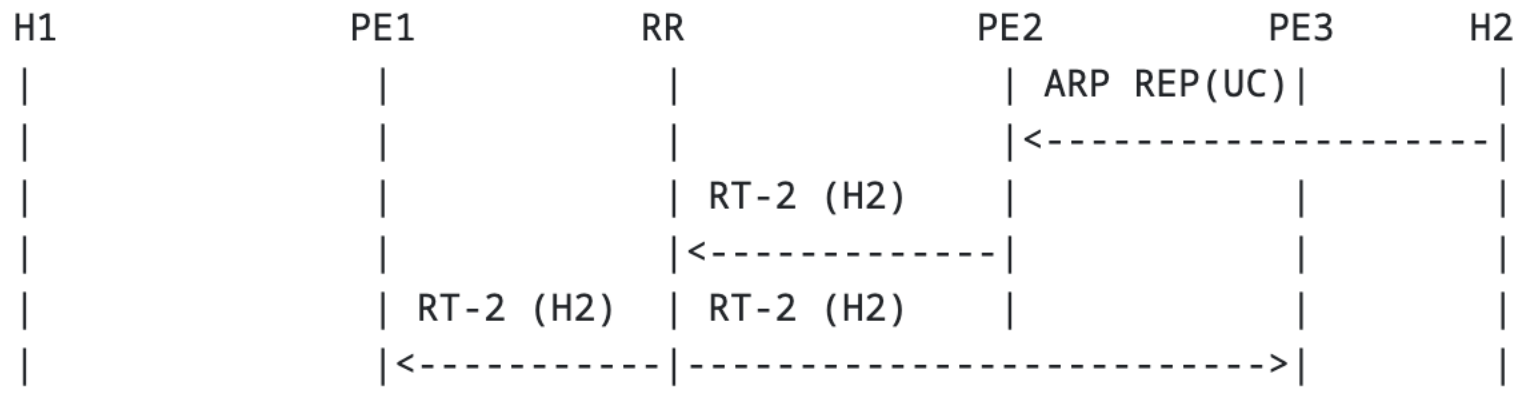


Figure 4: ARP Response from an IP host

Interop Scenario-1

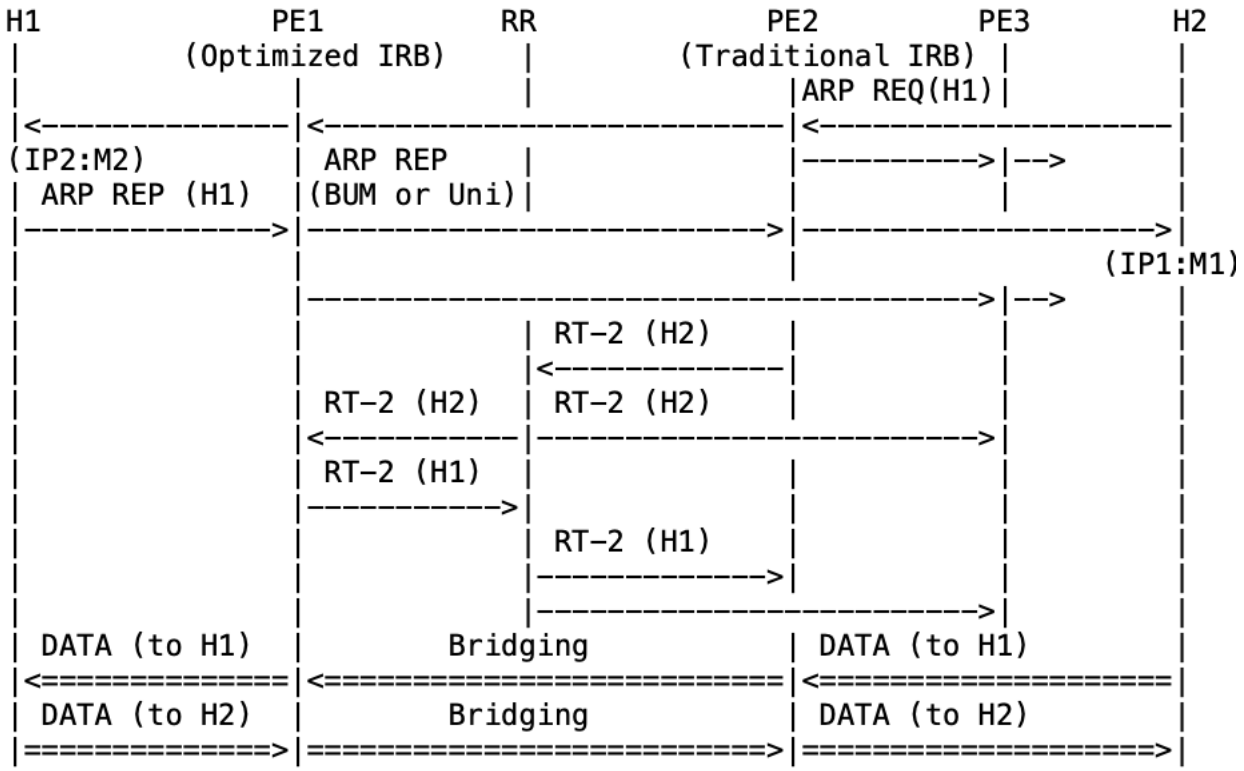


Figure 7: ARP Request received by a Traditional-IRB PE

Interop Scenario-2

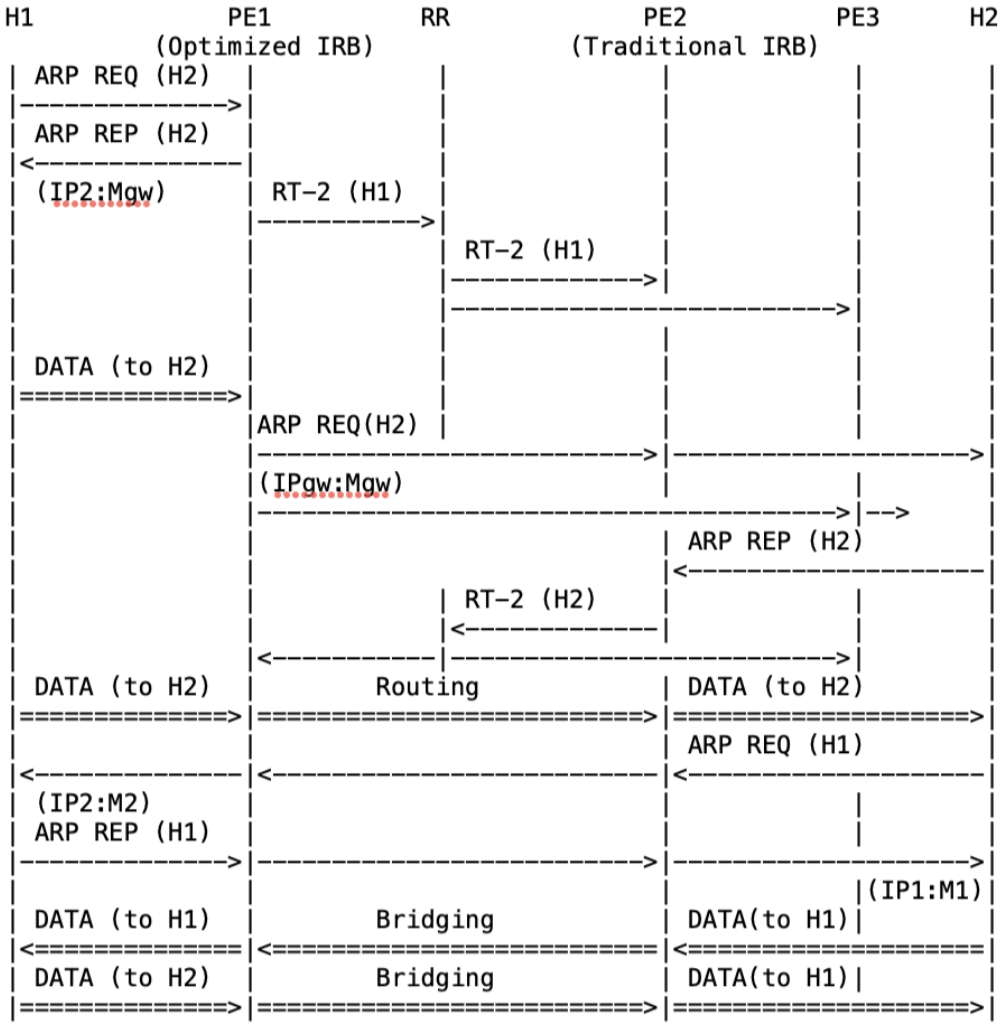


Figure 8: ARP Request received by a L3-Optimized-IRB PE

THANK YOU!