Internet Engineering Task Force (IETF)
Request for Comments: 9721
Category: Standards Track
ISSN: 2070-1721

N. Malhotra, Ed.
A. Sajassi
A. Pattekar
Cisco Systems
J. Rabadan
Nokia
A. Lingala
AT&T
J. Drake
Independent

April 2025

Extended Mobility Procedures for Ethernet VPN Integrated Routing and Bridging (EVPN-IRB)

Abstract

This document specifies extensions to the Ethernet VPN Integrated Routing and Bridging (EVPN-IRB) procedures specified in RFCs 7432 and 9135 to enhance the mobility mechanisms for networks based on EVPN-IRB. The proposed extensions improve the handling of host mobility and duplicate address detection in EVPN-IRB networks to cover a broader set of scenarios where a host's unicast IP address to Media Access Control (MAC) address bindings may change across moves. These enhancements address limitations in the existing EVPN-IRB mobility procedures by providing more efficient and scalable solutions. The extensions are backward compatible with existing EVPN-IRB implementations and aim to optimize network performance in scenarios involving frequent IP address mobility.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9721.

Table of Contents

1. Introduction

EVPN-IRB facilitates the advertisement of both MAC and IP routes via a single MAC+IP Route Type 2 (RT-2) advertisement. The MAC address is integrated into the local MAC Virtual Routing and Forwarding (MAC-VRF) bridge table, enabling Layer 2 (L2) bridged traffic across the network overlay. The IP address is incorporated into the local Address Resolution Protocol (ARP) / Neighbor Discovery Protocol (NDP) table in an asymmetric IRB design or into the IP Virtual Routing and Forwarding (IP-VRF) routing table in a symmetric IRB design. This facilitates routed traffic across the network overlay. For additional context on EVPN-IRB forwarding modes, refer to [RFC9135].

To support the EVPN mobility procedure, a single sequence number mobility attribute is advertised with the combined MAC+IP route. This approach, which resolves both MAC and IP reachability with a single sequence number, inherently assumes a fixed 1:1 mapping

between an IP address and MAC address. While this fixed 1:1 mapping is a common use case and is addressed via the existing mobility procedure defined in [RFC7432], there are additional IRB scenarios that do not adhere to this assumption. Such scenarios are prevalent in virtualized host environments where hosts connected to an EVPN network are Virtual Machines (VMs) or containerized workloads. The following IRB mobility scenarios are considered:

* A host move results in the host's IP address and MAC address moving together.

* A host move results in the host's IP address moving to a new MAC address association.

* A host move results in the host's MAC address moving to a new IP address association.

While the existing mobility procedure can manage the MAC+IP address move in the first scenario, the subsequent scenarios lead to new MAC-IP address associations. Therefore, a single sequence number assigned independently for each {MAC address, IP address} is insufficient to determine the most recent reachability for both MAC address and IP address, unless the sequence number assignment algorithm allows for changing MAC-IP address bindings across moves.

This document updates the sequence number assignment procedures defined in [RFC7432] to 1) adequately address mobility support across EVPN-IRB overlay use cases that permit MAC-IP address bindings to change across host moves and 2) support mobility for both MAC and IP route components carried in an EVPN RT-2 for these use cases.

Additionally, for hosts on an Ethernet Segment Identifier (ESI) that is multi-homed to multiple Provider Edge (PE) devices, additional procedures are specified to ensure synchronized sequence number assignments across the multi-homing devices.

This document addresses mobility for the following cases, independent of the overlay encapsulation (e.g., MPLS, Segment Routing over IPv6 (SRv6), and Network Virtualization Overlay (NVO) tunnel):

* Symmetric EVPN-IRB overlay

* Asymmetric EVPN-IRB overlay

* Routed EVPN overlay

1.1. Document Structure

The following sections of the document are informative:

* Section 3 provides the necessary background and problem statement being addressed in this document.

* Section 4 lists the resulting design considerations for the document.

* Section 5 lists the main solution components that are foundational for the specifications that follow in subsequent sections.

The following sections of the document are normative:

* Section 6 describes the mobility and sequence number assignment procedures in an EVPN-IRB overlay that are required to address the scenarios described in Section 4.

* Section 7 describes the mobility procedures for a routed overlay

network as opposed to an IRB overlay.

* Section 8 describes corresponding duplicate detection procedures for EVPN-IRB and routed overlays.

2. Requirements Language and Terminology

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

2.1. Abbreviations

ARP: Address Resolution Protocol [RFC0826]. ARP references in this document are equally applicable to both ARP and NDP.

CE: Customer Edge.

ES: Ethernet Segment. A physical ethernet or LAG port that connects an access device to an EVPN PE, as defined in [RFC7432].

EVPN PE: Ethernet VPN Provider Edge. A PE switch router in an EVPN-IRB network that runs overlay BGP-EVPN control planes and connects to overlay CE host devices. An EVPN PE may also be the first-hop L3 gateway for CE host devices. This document refers to EVPN PE as a logical function in an EVPN-IRB network. This EVPN PE function may be physically hosted on a ToR switching device or at layer(s) above the ToR in the Clos fabric. An EVPN PE is typically also an IP or MPLS tunnel endpoint for overlay VPN flows.

EVPN-IRB: Ethernet VPN Integrated Routing and Bridging. A BGP-EVPN distributed control plane that is based on the integrated routing and bridging fabric overlay discussed in [RFC9135].

L2: Layer 2.

L3: Layer 3.

LAG: Link Aggregation Group.

MC-LAG: Multi-Chassis Link Aggregation Group.

MPLS: Multiprotocol Label Switching (as specified in [RFC3031]).

NDP: Neighbor Discovery Protocol (for IPv6 [RFC4861]).

NVO: Network Virtualization Overlay.

NVO3: Network Virtualization over Layer 3 (as specified in [RFC8926]).

PE: Provider Edge.

RT-2: Route Type 2. EVPN RT-2 carrying both MAC address and IP address reachability as specified in [RFC7432].

RT-5: Route Type 5. EVPN RT-5 carrying IP prefix reachability as specified in [RFC9136].

SRv6: Segment Routing over IPv6 (as specified in [RFC8986]).

ToR: Top-of-Rack.

VM: Virtual Machine (or containerized workloads).

VXLAN: Virtual eXtensible Local Area Network (as specified in [RFC7348]).

2.2. Definitions

Asymmetric EVPN-IRB: A design approach used in EVPN-based networks [RFC9135] to handle L2 and L3 forwarding. In this approach, only the ingress PE router performs routing for inter-subnet traffic, while the egress PE router performs bridging.

EVPN all-active multi-homing: A redundancy and load-sharing mechanism used in EVPN networks. This method allows multiple PE devices to simultaneously provide L2 and L3 connectivity to a single CE device or network segment.

Host: In this document, generically refers to any user or CE endpoint attached to an EVPN-IRB network, which may be a VM, containerized workload, physical endpoint, or CE device.

MAC-IP address: The IPv4 and/or IPv6 address and MAC address binding for an overlay host.

Overlay: L2 and L3 VPNs that are enabled via NVO3, VXLAN, SRv6, or MPLS service-layer encapsulation.

Peer-Sync-Local MAC-IP route: The learned BGP EVPN MAC-IP route for a host that is directly attached to a local multi-homed ES.

Peer-Sync-Local MAC-IP sequence number: The sequence number received with a Peer-Sync-Local MAC-IP route.

Peer-Sync-Local MAC route: The learned BGP EVPN MAC route for a host that is directly attached to a local multi-homed ES.

Peer-Sync-Local MAC sequence number: The sequence number received with a Peer-Sync-Local MAC route.

Symmetric EVPN-IRB: A specific design approach used in EVPN-based networks [RFC9135] to handle both L2 and L3 forwarding within the same network infrastructure. The key characteristic of symmetric EVPN-IRB is that both ingress and egress PE routers perform routing for inter-subnet traffic.

Underlay: An IP, MPLS, or SRv6 fabric core network that provides routed reachability between EVPN PEs.

3. Background and Problem Statement

3.1. Optional MAC-Only RT-2

In an EVPN-IRB scenario where a single MAC+IP RT-2 advertisement carries both IP and MAC routes, a MAC-only RT-2 advertisement becomes redundant for host MAC addresses already advertised via MAC+IP RT-2. Consequently, the advertisement of a local MAC-only RT-2 is optional at an EVPN PE. This consideration is important for mobility scenarios discussed in subsequent sections. It is noteworthy that a local MAC route and its assigned sequence number are still maintained locally on a PE, and only the advertisement of this route to other PEs is optional.

MAC-only RT-2 advertisements may still be issued for non-IP host MAC addresses that are not included in MAC+IP RT-2 advertisements.

3.2. Mobility Use Cases

This section outlines the IRB mobility use cases addressed in this document. Detailed procedures to handle these scenarios are provided in Sections 6 and 7. The following IRB mobility scenarios are considered:

* A host move results in both the host's IP and MAC addresses moving together.

* A host move results in the host's IP address moving to a new MAC address association.

* A host move results in the host's MAC address moving to a new IP address association.

3.2.1. Host MAC+IP Address Move

This is the baseline scenario where a host move results in both the host's MAC and IP addresses moving together without altering the MAC-IP address binding. The existing MAC route mobility procedures defined in [RFC7432] can be leveraged to support this MAC+IP address mobility scenario.

3.2.2. Host IP Address Move to New MAC Address

This scenario involves a host move where the host's IP address is reassigned to a new MAC address.

3.2.2.1. Host Reload

A host reload or orchestrated move may cause a host to be re-spawned at the same or new PE location, resulting in a new MAC address assignment while retaining the existing IP address. This results in the host's IP address moving to a new MAC address binding, as shown below:

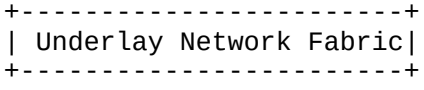
IP-a, MAC-a ---> IP-a, MAC-b

3.2.2.2. MAC Address Sharing

This scenario considers cases where multiple hosts, each with a unique IP address, share a common MAC address. A host move results in a new MAC address binding for the host IP address. For example, hosts running on a single physical server might share the same MAC address. Alternatively, an L2 access network behind a firewall may have all host IP addresses learned with a common firewall MAC address. In these "shared MAC" scenarios, multiple local MAC-IP ARP/NDP entries may be learned with the same MAC address. A host IP address move to a new physical server could result in a new MAC address association for the host IP.

3.2.2.3. Problem

In the aforementioned scenarios, a combined MAC+IP EVPN RT-2 advertised with a single sequence number attribute assumes a fixed IP-to-MAC address mapping. A host IP address move to a new MAC address breaks this assumption and results in a new MAC+IP route. If this new route is independently assigned a new sequence number, the sequence number can no longer determine the most recent host IP reachability in a symmetric EVPN-IRB design or the most recent IP-to-MAC address binding in an asymmetric EVPN-IRB design.



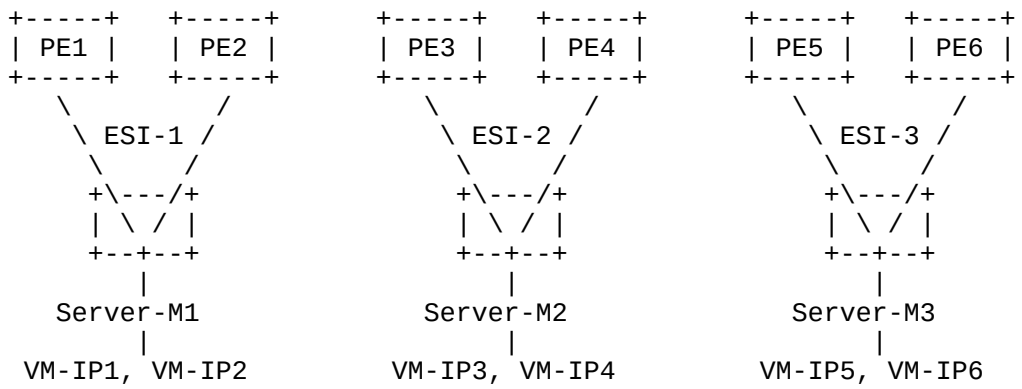


Figure 1

Figure 1 illustrates a topology with host VMs sharing the physical server MAC address. In steady state, the IP1-M1 route is learned at PE1 and PE2 and advertised to remote PEs with a sequence number N. If VM-IP1 moves to Server-M2, ARP or NDP-based local learning at PE3 and PE4 would result in a new IP1-M2 route. If this new route is assigned a sequence number of 0, the mobility procedure for VM-IP1 will not trigger across the overlay network.

A sequence number assignment procedure must be defined to unambiguously determine the most recent IP address reachability, IP-to-MAC address binding, and MAC address reachability for such MAC address sharing scenarios.

3.2.3. Host MAC Address Move to New IP Address

This is a scenario where a host move or re-provisioning behind the same or new PE location may result in the host getting a new IP address assigned while keeping the same MAC address.

3.2.3.1. Problem

The complication in this scenario arises because MAC address reachability can be carried via a combined MAC+IP route, whereas a MAC-only route may not be advertised. Associating a single sequence number with the MAC+IP route implicitly assumes a fixed MAC-to-IP mapping. A MAC address move that results in a new IP address association breaks this assumption and creates a new MAC+IP route. If this new route independently receives a new sequence number, the sequence number can no longer reliably indicate the most recent host MAC address reachability.

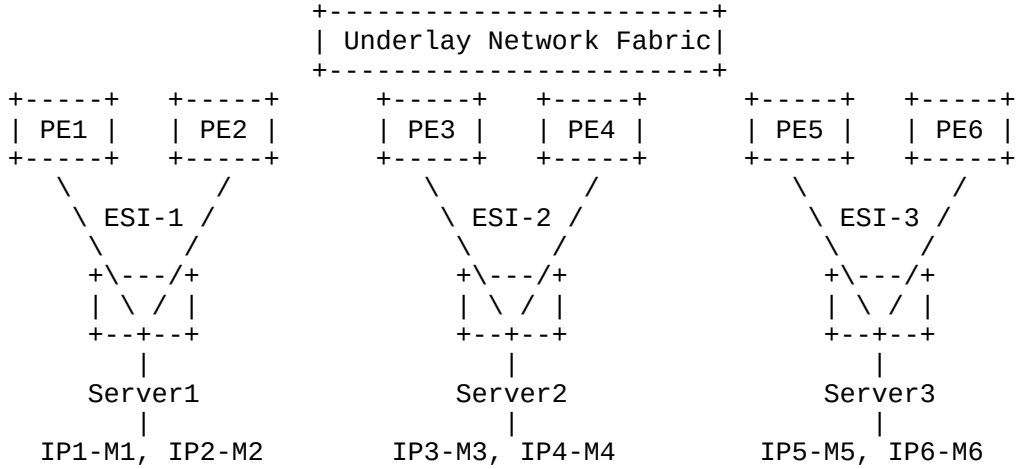


Figure 2

For instance, consider that host IP1-M1 is learned locally at PE1 and PE2 and advertised to remote hosts with sequence number N. If this host with MAC address M1 is re-provisioned at Server2 and assigned a different IP address (e.g., IP7), then the new IP7-M1 route learned at PE3 and PE4 would be advertised with sequence number 0. Consequently, L3 reachability to IP7 would be established across the overlay, but the MAC mobility procedure for M1 would not trigger due to the new MAC-IP route advertisement. Advertising an optional MAC-only route with its sequence number would trigger MAC mobility per [RFC7432]. However, without this additional advertisement, a single sequence number associated with a combined MAC+IP route may be insufficient to update MAC address reachability across the overlay.

A MAC-IP route sequence number assignment procedure is required to unambiguously determine the most recent MAC address reachability in the previous scenarios without advertising a MAC-only route.

Furthermore, upon learning new reachability for IP7-M1 via PE3 and PE4, PE1 and PE2 must probe and delete any local IPs associated with the MAC address M1, such as IP1-M1.

It could be argued that the MAC mobility sequence number defined in [RFC7432] applies only to the MAC route part of a MAC-IP route, thus covering this scenario. This interpretation could serve as a clarification to [RFC7432] and supports the need for a common sequence number assignment procedure across all MAC-IP mobility scenarios detailed in this document.

3.3. EVPN All Active Multi-Homed ES

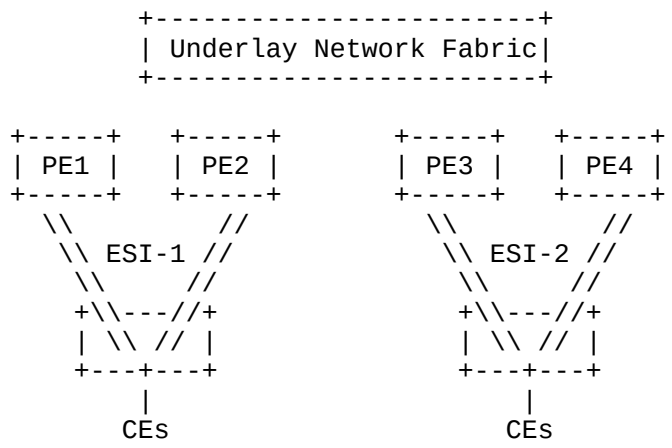


Figure 3

Consider an EVPN-IRB overlay network illustrated in Figure 3, where hosts are multi-homed to two or more PE devices via an all-active multi-homed ES. MAC and ARP/NDP entries learned on a local ES may also be synchronized across the multi-homing PE devices sharing this ES. This synchronization enables local switching of intra- and inter-subnet ECMP traffic flows from remote hosts. Thus, local MAC and ARP/NDP entries on a given ES may be learned through local learning and/or synchronization from another PE device sharing the same ES.

For a host that is multi-homed to multiple PE devices via an all-active ES interface, the local learning of the host MAC and MAC-IP routes at each PE device is an independent asynchronous event, dependent on traffic flow or an ARP/NDP response from the host hashing to a directly connected PE on the MC-LAG interface. Consequently, the sequence number mobility attribute value assigned to a locally learned MAC or MAC-IP route at each device may not always be the same, depending on transient states on the device at

the time of local learning.

For example, consider a host that is deleted from ESI-2 and moved to ESI-1. It is possible for the host to be learned on PE1 following the deletion of the remote route from PE3 and PE4 while being learned on PE2 prior to the deletion of the remote route from PE3 and PE4. In this case, PE1 would process local host route learning as a new route and assign a sequence number of 0, while PE2 would process local host route learning as a remote-to-local move and assign a sequence number of N+1, where N is the existing sequence number assigned at PE3 and PE4.

Inconsistent sequence numbers advertised from multi-homing devices:

* Create ambiguity regarding how remote PEs should handle paths with the same ESI but different sequence numbers. A remote PE might not program ECMP paths if it receives routes with different sequence numbers from a set of multi-homing PEs sharing the same ESI.

* Break consistent route versioning across the network overlay that is needed for EVPN mobility procedures to work.

For instance, in this inconsistent state, PE2 would drop a remote route received for the same host with sequence number N (since its local sequence number is N+1), while PE1 would install it as the best route (since its local sequence number is 0).

To support mobility for multi-homed hosts using the sequence number mobility attribute, local MAC and MAC-IP routes learned on a multi-homed ES must be advertised with the same sequence number by all PE devices to which the ES is multi-homed. There is a need for a mechanism to ensure the consistency of sequence numbers assigned across these PEs.

4. Design Considerations

To summarize, the sequence number assignment scheme and implementation must consider the following:

* Synchronization across multi-homing PE devices:

  MAC+IP routes may be learned on an ES that is multi-homed to multiple PE devices, requiring synchronized sequence numbers across these devices.

* Optional MAC-only RT-2:

  In an IRB scenario, MAC-only RT-2 is optional and may not be advertised alongside MAC+IP RT-2.

* Multiple IPs associated with a single MAC:

  A single MAC address may be linked to multiple IP addresses, indicating multiple host IPs sharing a common MAC address.

* Host IP movement:

  A host IP address move may result in a new MAC address association, necessitating a new IP address to MAC address association and a new MAC+IP route.

* Host MAC address movement:

  A host MAC address move may result in a new IP address association, requiring a new MAC address to IP address association

and a new MAC+IP route.

* Local MAC-IP route learning via ARP/NDP:

  Local MAC-IP route learning via ARP/NDP always accompanies a local MAC route learning event resulting from the ARP/NDP packet. However, MAC and MAC-IP route learning can occur in any order.

* Separate sequence numbers for MAC and IP routes:

  Use cases that do not maintain a constant 1:1 MAC-IP address mapping across moves could potentially be addressed by using separate sequence numbers for MAC and IP route components of the MAC+IP route. However, maintaining two separate sequence numbers adds significant complexity, debugging challenges, and backward compatibility issues. Therefore, this document addresses these requirements using a single sequence number attribute.

5. Solution Components

This section outlines the main components of the EVPN-IRB mobility solution specified in this document. Subsequent sections will detail the exact sequence number assignment procedures based on the concepts described here.

5.1. Sequence Number Inheritance

The key concept presented here is to treat a local MAC-IP route as a child of the corresponding local MAC route within the local context of a PE. This ensures that the local MAC-IP route inherits the sequence number attribute from the parent local MAC-only route. In terms of object dependencies, this could be represented as the MAC-IP route being a dependent child of the parent MAC route:

```
Mx-IPx -----> Mx (seq# = N)
```

Thus, both the parent MAC route and the child MAC-IP routes share a common sequence number associated with the parent MAC route. This ensures that a single sequence number attribute carried in a combined MAC+IP route represents the sequence number for both a MAC-only route and a MAC+IP route, making the advertisement of the MAC-only route truly optional. This enables a MAC address to assume a different IP address upon moving and still establish the most recent reachability to the MAC address across the overlay network via the mobility attribute associated with the MAC+IP route advertisement. For instance, when Mx moves to a new location, it would be assigned a higher sequence number at its new location per [RFC7432]. If this move results in Mx assuming a different IP address, IPz, the local Mx+IPz route would inherit the new sequence number from Mx.

Local MAC and local MAC-IP routes are typically sourced from data plane learning and ARP/NDP learning, respectively, and can be learned in the control plane in any order. Implementations can either replicate the inherited sequence number in each MAC-IP entry or maintain a single attribute in the parent MAC route by creating a forward reference local MAC object for cases where a local MAC-IP route is learned before the local MAC route.

5.2. MAC Address Sharing

For the shared MAC address scenario, multiple local MAC-IP sibling routes inherit the sequence number attribute from the common parent MAC route:

```
Mx-IP1 -----
  |          |
```

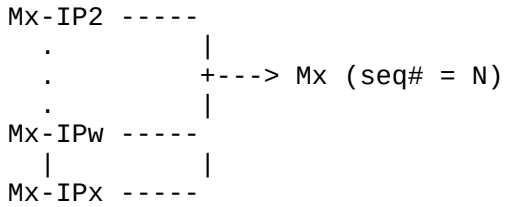


Figure 4

In such cases, a host-IP move to a different physical server results in the IP moving to a new MAC address binding. A new MAC-IP route resulting from this move must be advertised with a sequence number higher than the previous MAC-IP route for this IP, advertised from the prior location. For example, consider a route Mx-IPx currently advertised with sequence number N from PE1. If IPx moves to a new physical server behind PE2 and is associated with MAC Mz, the new local Mz-IPx route must be advertised with a sequence number higher than N and higher than the previous Mz sequence number M. This allows PE devices, including PE1, PE2, and other remote PE devices, to determine and program the most recent MAC address binding and reachability for the IP. PE1, upon receiving this new Mz-IPx route with sequence number N+1 or M+1 (whichever is greater), would update IPx reachability via PE2 for symmetric IRB and update IPx's ARP/NDP binding to Mz for asymmetric IRB while clearing and withdrawing the stale Mx-IPx route with the lower sequence number.

This implies that the sequence number associated with the local MAC route Mz and all local MAC-IP child routes of Mz at PE2 must be incremented to N+1 or M+1 if the previous Mz sequence number M is greater than N and is re-advertised across the overlay. While this re-advertisement of all local MAC-IP children routes affected by the parent MAC route adds overhead, it also avoids the need for maintaining and advertising two separate sequence number attributes for IP and MAC route components of MAC+IP RT-2. Implementations must be able to look up MAC-IP routes for a given IP and update the sequence number for its parent MAC route and for its MAC-IP route children.

5.3. Sequence Number Synchronization

To support mobility for multi-homed hosts, local MAC and MAC-IP routes learned on a shared ES must be advertised with the same sequence number by all PE devices to which the ES is multi-homed. This applies to local MAC-only routes as well. MAC and MAC-IP routes for a host that is attached to a local ES may be learned via data plane and ARP/NDP, respectively, as well as via BGP EVPN from another multi-homing PE to achieve local switching. MAC and MAC-IP routes learned via data plane and ARP/NDP are respectively referred to as local MAC routes and local MAC-IP routes. BGP EVPN learned MAC and MAC-IP routes for a host that is attached to a local ES are respectively referred to as Peer-Sync-Local MAC routes and Peer-Sync-Local MAC-IP routes as they are effectively local routes synchronized from a multi-homing peer. Local and Peer-Sync-Local route learning can occur in any order. Local MAC-IP routes advertised by all multi-homing PE devices sharing the ES must carry the same sequence number, independent of the order in which they are learned. This implies that:

* On local or Peer-Sync-Local MAC-IP route learning, the sequence number for the local MAC-IP route must be compared and updated to the higher value.

* On local or Peer-Sync-Local MAC route learning, the sequence number for the local MAC route must be compared and updated to the higher value.

If an update to the local MAC-IP route sequence number is required as a result of the comparison with the Peer-Sync-Local MAC-IP route, it essentially amounts to a sequence number update on the parent local MAC route, resulting in an inherited sequence number update on the local MAC-IP route.

6. Methods for Sequence Number Assignment

The following sections specify the sequence number assignment procedures required for local MAC, local MAC-IP, Peer-Sync-Local MAC, and Peer-Sync-Local MAC-IP route learning events to achieve the outlined objectives.

6.1. Local MAC-IP Learning

A local Mx-IPx learning via ARP or NDP should result in the computation or re-computation of the parent MAC route Mx's sequence number. After this, the MAC-IP route Mx-IPx inherits the parent MAC route's sequence number. The parent MAC route Mx sequence number MUST be computed as follows:

* It MUST be higher than any existing remote MAC route for Mx, as per [RFC7432].

* It MUST be at least equal to the corresponding Peer-Sync-Local MAC route sequence number, if present.

* It MUST be higher than the "Mz" sequence number if the IP is also associated with a different remote MAC "Mz".

Once the new sequence number for the MAC route Mx is computed as per the above criteria, all local MAC-IP routes associated with MAC Mx MUST inherit the updated sequence number.

6.2. Local MAC Learning

The local MAC route Mx sequence number MUST be computed as follows:

* It MUST be higher than any existing remote MAC route for Mx, as per [RFC7432].

* It MUST be at least equal to the corresponding Peer-Sync-Local MAC route sequence number if one is present.

  If the existing local MAC route sequence number is less than the Peer-Sync-Local MAC route sequence number, then the PE MUST update the local MAC route sequence number to be equal to the Peer-Sync-Local MAC route sequence number.

  If the existing local MAC route sequence number is equal to or greater than the Peer-Sync-Local MAC route sequence number, no update is required to the local MAC route sequence number.

Once the new sequence number for the MAC route Mx is computed as per the above criteria, all local MAC-IP routes associated with the MAC route Mx MUST inherit the updated sequence number. Note that the local MAC route sequence number might already be present if there was a local MAC-IP route learned prior to the local MAC route. In this case, the above may not result in any change in the local MAC route sequence number.

6.3. Remote MAC or MAC-IP Route Update

Upon receiving a remote MAC or MAC-IP route update associated with a MAC address Mx with a sequence number that is either:

* higher than the sequence number assigned to a local route for MAC Mx or

* equal to the sequence number assigned to a local route for MAC Mx, but the remote route is selected as best due to a lower VXLAN Tunnel End Point (VTEP) IP as per [RFC7432],

the following actions are REQUIRED on the receiving PE:

* The PE MUST trigger a probe and deletion procedure for all local MAC-IP routes associated with MAC Mx.

* The PE MUST trigger a deletion procedure for the local MAC route for Mx.

6.4. Peer-Sync-Local MAC Route Update

Upon receiving a Peer-Sync-Local MAC route, the corresponding local MAC route Mx sequence number (if present) should be re-computed as follows:

* If the current sequence number is less than the received Peer-Sync-Local MAC route sequence number, it MUST be increased to be equal to the received Peer-Sync-Local MAC route sequence number.

* If a local MAC route sequence number is updated as a result of the above, all local MAC-IP routes associated with MAC route Mx MUST inherit the updated sequence number.

6.5. Peer-Sync-Local MAC-IP Route Update

Because the MAC-only RT-2 advertisement is optional, receiving a Peer-Sync-Local MAC-IP route for a locally attached host results in a derived Peer-Sync-Local MAC Mx route entry. The corresponding local MAC Mx route sequence number (if present) should be re-computed as follows:

* If the current sequence number is less than the received Peer-Sync-Local MAC route sequence number, it MUST be increased to be equal to the received Peer-Sync-Local MAC route sequence number.

* If a local MAC route sequence number is updated as a result of the above, all local MAC-IP routes associated with MAC route Mx MUST inherit the updated sequence number.

6.6. Interoperability

Generally, if all PE nodes in the overlay network follow the above sequence number assignment procedures and the PE is advertising both MAC+IP and MAC routes, the sequence numbers advertised with the MAC and MAC+IP routes with the same MAC address would always be the same. However, an interoperability scenario with a different implementation could arise, where a non-compliant PE implementation assigns and advertises independent sequence numbers to MAC and MAC+IP routes. To handle this case, if different sequence numbers are received for remote MAC+IP routes and corresponding remote MAC routes from a remote PE, the sequence number associated with the remote MAC route MUST be:

* computed and interpreted as the highest of all received sequence numbers with remote MAC+IP and MAC routes with the same MAC address and

* re-computed on a MAC or MAC+IP route withdraw as per the above.

A MAC and/or IP address move to the local PE would then result in the MAC (and hence all MAC-IP) route sequence numbers being incremented from the above computed remote MAC route sequence number.

If MAC-only routes are not advertised at all, and different sequence numbers are received with multiple MAC+IP routes for a given MAC address, the sequence number associated with the derived remote MAC route should still be computed as the highest of all received MAC+IP route sequence numbers with the same MAC address.

Note that it is not required for a PE to maintain explicit knowledge of a remote PE being compliant or non-compliant with this specification as long as it implements the above logic to handle remote sequence numbers that are not synchronized between MAC route and MAC-IP route(s) for the same remote MAC address.

6.7. MAC Address Sharing Race Condition

In a MAC address sharing use case described in Section 5.2, a race condition is possible with simultaneous host moves between a pair of PEs. The example scenario below illustrates this race condition and its remediation:

* PE1 with locally attached host IPs I1 and I2 that share MAC address M1. As a result, PE1 has local MAC-IP routes I1-M1 and I2-M1.

* PE2 with locally attached host IPs I3 and I4 that share MAC address M2. As a result, PE2 has local MAC-IP routes I3-M2 and I4-M2.

* A simultaneous move of I1 from PE1 to PE2 and of I3 from PE2 to PE1 will cause I1’s MAC address to change from M1 to M2 and cause I3’s MAC address to change from M2 to M1.

* Route I3-M1 may be learned on PE1 before I1’s local entry I1-M1 has been probed out on PE1 and/or route I1-M2 may be learned on PE2 before I3’s local entry I3-M2 has been probed out on PE2.

* In such a scenario, MAC route sequence number assignment rules defined in Section 6.1 will cause new MAC-IP routes I1-M2 and I3-M1 to bounce between PE1 and PE2 with sequence number increments until stale entries I1-M1 and I3-M2 have been probed out from PE1 and PE2, respectively.

An implementation MUST ensure proper probing procedures to remove stale ARP, NDP, and local MAC entries, following a move, on learning remote routes as defined in Section 6.3 (and as per [RFC9135]) to minimize exposure to this race condition.

6.8. Mobility Convergence

This section is optional and details ARP and NDP probing procedures that MAY be implemented to achieve faster host relearning and convergence on mobility events. PE1 and PE2 are used as two example PEs in the network to illustrate the mobility convergence scenarios in this section.

* Following a host move from PE1 to PE2, the host’s MAC address is discovered at PE2 as a local MAC route via data frames received from the host. If PE2 has a prior remote MAC-IP host route for this MAC address from PE1, an ARP/NDP probe MAY be triggered at PE2 to learn the MAC-IP address as a local adjacency and trigger EVPN RT-2 advertisement for this MAC-IP address across the overlay with new reachability via PE2. This results in a reliable "event-based" host IP learning triggered by a "MAC address learning

event" across the overlay, and hence, a faster convergence of overlay routed flows to the host.

* Following a host move from PE1 to PE2, once PE1 receives a MAC or MAC-IP route from PE2 with a higher sequence number, an ARP/NDP probe MAY be triggered at PE1 to clear the stale local MAC-IP neighbor adjacency or to relearn the local MAC-IP in case the host has moved back or is duplicated.

* Following a local MAC route age-out, if there is a local IP adjacency with this MAC address, an ARP/NDP probe MAY be triggered for this IP to either relearn the local MAC route and maintain local L3 and L2 reachability to this host or to clear the ARP/NDP entry if the host is no longer local. This accomplishes the clearance of stale ARP/NDP entries triggered by a MAC age-out event even when the ARP/NDP refresh timer is longer than the MAC age-out timer. Clearing stale IP neighbor entries facilitates traffic convergence if the host was silent and not discovered at its new location. Once the stale neighbor entry for the host is cleared, routed traffic flow destined for the host can re-trigger ARP/NDP discovery for this host at the new location.

6.8.1. Generalized Probing Logic

The above probing logic may be generalized as probing for an IP neighbor anytime a resolving parent MAC route is inconsistent with the MAC-IP neighbor route, where inconsistency is defined as being not present or conflicting in terms of the route source being local or remote. The MAC-IP route to parent MAC route relationship described in Section 5.1 MAY be used to achieve this logic.

7. Routed Overlay

An additional use case involves traffic to an end host in the overlay being entirely IP routed. In such a purely routed overlay:

* A host MAC route is never advertised in the EVPN overlay control plane.

* Host /32 or /128 IP reachability is distributed across the overlay via EVPN Route Type 5 (RT-5) along with a zero or non-zero ESI.

* An overlay IP subnet may still be stretched across the underlay fabric. However, intra-subnet traffic across the stretched overlay is never bridged.

* Both inter-subnet and intra-subnet traffic in the overlay is IP routed at the EVPN PE.

Please refer to [RFC7814] for more details.

Host mobility within the stretched subnet still needs support. In the absence of host MAC routes, the sequence number mobility Extended Community specified in Section 7.7 of [RFC7432] MAY be associated with a /32 or /128 host IP prefix advertised via EVPN Route Type 5. MAC mobility procedures defined in [RFC7432] can be applied to host IP prefixes as follows:

* On local learning of a host IP on a new ESI, the host IP MUST be advertised with a sequence number higher than what is currently advertised with the old ESI.

* On receiving a host IP route advertisement with a higher sequence number, a PE MUST trigger ARP/NDP probe and deletion procedures on any local route for that IP with a lower sequence number. The PE will update the forwarding entry to point to the remote route with

a higher sequence number and send an ARP/NDP probe for the local IP route. If the IP has moved, the probe will time out, and the local IP host route will be deleted.

Note that there is only one sequence number associated with a host route at any time. For previous use cases where a host MAC address is advertised along with the host IP address, a sequence number is only associated with the MAC address. If the MAC is not advertised, as in this use case, a sequence number is associated with the host IP address.

This mobility procedure does not apply to "anycast" IPv6 hosts advertised via Neighbor Advertisement (NA) messages with the Override Flag (O Flag) set to 0. Refer to [RFC9161] for more details.

8. Duplicate Host Detection

Duplicate host detection scenarios across EVPN-IRB can be classified as follows:

* Scenario A: Two hosts have the same MAC address (host IPs may or may not be duplicates).

* Scenario B: Two hosts have the same IP address but different MAC addresses.

* Scenario C: Two hosts have the same IP address, and the host MAC address is not advertised.

As specified in [RFC9161], duplicate detection procedures for Scenarios B and C do not apply to "anycast" IPv6 hosts advertised via NA messages with the Override Flag (O Flag) set to 0.

8.1. Scenario A

In cases where duplicate hosts share the same MAC address, the MAC address is detected as duplicate using the duplicate MAC address detection procedure described in [RFC7432]. Corresponding MAC-IP routes with the same MAC address do not require separate duplicate detection and MUST inherit the duplicate property from the MAC route. If a MAC route is marked as duplicate, all associated MAC-IP routes MUST also be treated as duplicates. Duplicate detection procedures need only be applied to MAC routes.

8.2. Scenario B

Misconfigurations may lead to different MAC addresses being assigned the same IP address. This scenario is not detected by the duplicate MAC address detection procedures from [RFC7432] and can result in incorrect routing of traffic destined for the IP address.

Such situations, when detected locally, are identified as a move scenario through the local MAC route sequence number computation procedure described in Section 6.1:

* If the IP is associated with a different remote MAC "Mz", the sequence number MUST be higher than the "Mz" sequence number.

This move results in a sequence number increment for the local MAC route due to the remote MAC-IP route being associated with a different MAC address, which counts as an "IP move" against the IP, independent of the MAC. The duplicate detection procedure described in [RFC7432] can then be applied to the IP entity independent of the MAC. Once an IP is detected as duplicate, the corresponding MAC-IP route should be treated as duplicate. Associated MAC routes and any other MAC-IP routes related to this MAC should not be affected.

8.2.1. Duplicate IP Detection Procedure for Scenario B

The duplicate IP detection procedure for this scenario is specified in [RFC9161]. An "IP move" is further clarified as follows:

* Upon learning a local MAC-IP route Mx-IPx, check for existing remote or local routes for IPx with a different MAC address association (Mz-IPx). If found, count this as an "IP move" for IPx, independent of the MAC.

* Upon learning a remote MAC-IP route Mz-IPx, check for existing local routes for IPx with a different MAC address association (Mx-IPx). If found, count this as an "IP move" for IPx, independent of the MAC.

A MAC-IP route MUST be treated as duplicate if either:

* the corresponding MAC route is marked as duplicate via the existing detection procedure, or

* the corresponding IP is marked as duplicate via the extended procedure described above.

8.3. Scenario C

As described in Section 7, in a purely routed overlay scenario where only a host IP is advertised via EVPN RT-5 with a sequence number mobility attribute, procedures similar to the duplicate MAC address detection procedures specified in [RFC7432] can be applied to IP-only host routes for duplicate IP detection as follows:

* Upon learning a local host IP route IPx, check for existing remote or local routes for IPx with a different ESI association. If found, count this as an "IP move" for IPx.

* Upon learning a remote host IP route IPx, check for existing local routes for IPx with a different ESI association. If found, count this as an "IP move" for IPx.

* Using configurable parameters "N" and "M", if "N" IP moves are detected within "M" seconds for IPx, then IPx should be treated as duplicate.

8.4. Duplicate Host Recovery

Once a MAC or IP address is marked as duplicate and frozen, corrective action must be taken to un-provision one of the duplicate MAC or IP addresses. Un-provisioning refers to corrective action taken on the host side. Following this correction, normal operation will not resume until the duplicate MAC or IP address ages out, unless additional action is taken to expedite recovery.

Possible additional corrective actions for faster recovery are outlined in the following sections.

8.4.1. Route Unfreezing Configuration

Unfreezing the duplicate or frozen MAC or IP route via a command-line interface (CLI) can be used to recover from the duplicate and frozen state following corrective un-provisioning of the duplicate MAC or IP address. Unfreezing the MAC or IP route should result in advertising it with a sequence number higher than that advertised from the other location.

Two scenarios exist:

* Scenario A: The duplicate MAC or IP address is un-provisioned at the location where it was not marked as duplicate.

* Scenario B: The duplicate MAC or IP address is un-provisioned at the location where it was marked as duplicate.

Unfreezing the duplicate and frozen MAC or IP route will result in recovery to a steady state as follows:

* Scenario A: If the duplicate MAC or IP address is un-provisioned at the non-duplicate location, unfreezing the route at the frozen location results in advertising with a higher sequence number, leading to automatic clearing of the local route at the un-provisioned location via ARP/NDP PROBE and DELETE procedures.

* Scenario B: If the duplicate host is un-provisioned at the duplicate location, unfreezing the route triggers an advertisement with a higher sequence number to the other location, prompting relearning and clearing of the local route at the original location upon receiving the remote route advertisement.

The probes referred to in these scenarios are event-driven probes resulting from receiving a route with a higher sequence number. Periodic probes resulting from refresh timers may also occur independently.

8.4.2. Route Clearing Configuration

In addition to the above, route clearing CLIs may be used to clear the local MAC or IP route after the duplicate host is un-provisioned:

* Clear MAC CLI: Used to clear a duplicate MAC route.

* Clear ARP/NDP: Used to clear a duplicate IP route.

The route unfreeze CLI may still need to be executed if the route was un-provisioned and cleared from the non-duplicate location. Given that unfreezing the route via the CLI would result in auto-clearing from the un-provisioned location, as explained earlier, using a route clearing CLI for recovery from the duplicate state is optional.

9. Security Considerations

The security considerations discussed in [RFC7432] and [RFC9135] apply to this document. Methods described in this document further extend the consumption of sequence numbers for IRB deployments. Hence, they are subject to the same considerations if the control plane or data plane was to be compromised. As an example, if the host-facing data plane is compromised, spoofing attempts could result in a legitimate host being perceived as moved, eventually resulting in the host being marked as duplicate. The considerations for protecting control and data planes described in [RFC7432] are equally applicable to such mobility spoofing use cases.

10. IANA Considerations

This document has no IANA actions.

11. References

11.1. Normative References

[RFC0826] Plummer, D., "An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware", STD 37,

RFC 826, DOI 10.17487/RFC0826, November 1982,
<https://www.rfc-editor.org/info/rfc826>.

[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>.

[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>.

[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>.

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

[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>.

[RFC9161] Rabadan, J., Ed., Sathappan, S., Nagaraj, K., Hankins, G., and T. King, "Operational Aspects of Proxy ARP/ND in Ethernet Virtual Private Networks", RFC 9161, DOI 10.17487/RFC9161, January 2022, <https://www.rfc-editor.org/info/rfc9161>.

11.2. Informative References

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[RFC7814] Xu, X., Jacquenet, C., Raszuk, R., Boyes, T., and B. Fee, "Virtual Subnet: A BGP/MPLS IP VPN-Based Subnet Extension Solution", RFC 7814, DOI 10.17487/RFC7814, March 2016, <https://www.rfc-editor.org/info/rfc7814>.

[RFC8926] Gross, J., Ed., Ganga, I., Ed., and T. Sridhar, Ed., "Geneve: Generic Network Virtualization Encapsulation", RFC 8926, DOI 10.17487/RFC8926, November 2020, <https://www.rfc-editor.org/info/rfc8926>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

[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>.

Acknowledgements

The authors would like to thank Gunter Van de Velde for the significant contributions to improve the readability of this document. The authors would also like to thank Sonal Agarwal and Larry Kreeger for multiple contributions through the implementation process. The authors would like to thank Vibov Bhan and Patrice Brissette for early feedback during the implementation and testing of several procedures defined in this document. The authors would like to thank Wen Lin for a detailed review and valuable comments related to MAC sharing race conditions. The authors would also like to thank Saumya Dikshit for a detailed review and valuable comments across the document.

Contributors

Gunter Van de Velde
Nokia
Email: van_de_velde@nokia.com

Wen Lin
Juniper
Email: wlin@juniper.net

Sonal Agarwal
Arrcus
Email: sonal@arrcus.com

Authors' Addresses

Neeraj Malhotra (editor)
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: nmalhotr@cisco.com

Ali Sajassi
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: sajassi@cisco.com

Aparna Pattekar
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: apjoshi@cisco.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043
United States of America
Email: jorge.rabadan@nokia.com

Avinash Lingala
AT&T
3400 W Plano Pkwy
Plano, TX 75075
United States of America
Email: ar977m@att.com

John Drake
Independent
Email: je_drake@yahoo.com