Authors: N. Malhotra, Ed.   A. Sajassi   A. Pattekar
         Cisco Systems   Cisco Systems   Cisco Systems
         J. Rabadan   A. Lingala   J. Drake
         Nokia   ATT   Juniper Networks

Extended Mobility Procedures for EVPN-IRB

Abstract

   Procedure to handle host mobility in a layer 2 Network with EVPN
   control plane is defined as part of RFC 7432. EVPN has since evolved
   to find wider applicability across various IRB use cases that
   include distributing both MAC and IP reachability via a common EVPN
   control plane. MAC Mobility procedures defined in RFC 7432 are
   extensible to IRB use cases if a fixed 1:1 mapping between VM IP and
   MAC is assumed across VM moves. Generic mobility support for IP and
   MAC that allows these bindings to change across moves is required to
   support a broader set of EVPN IRB use cases, and requires further
   consideration. EVPN all-active multi-homing further introduces
   scenarios that require additional consideration from mobility
   perspective. This document enumerates a set of design considerations
   applicable to mobility across these EVPN IRB use cases and defines
   generic sequence number assignment procedures to address these IRB
   use cases.

Status of This Memo

Table of Contents

1. Requirements Language and Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this
document are to be interpreted as described in [[RFC2119]()].

 *EVPN-IRB: A BGP-EVPN distributed control plane based integrated
  routing and bridging fabric overlay discussed in [EVPN-IRB]

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

 *Overlay: VPN or service layer network consisting of EVPN PEs OR
  VPN provider-edge (PE) switch-router devices that runs on top of
  an underlay routed core.

 *EVPN PE: A PE switch-router in a data-center fabric that runs
  overlay BGP-EVPN control plane and connects to overlay CE host
  devices. An EVPN PE may also be the first-hop layer-3 gateway for
  CE/host devices. This document refers to EVPN PE as a logical
  function in a data-center fabric. This EVPN PE function may be
  physically hosted on a top-of-rack switching device (ToR) OR at
  layer(s) above the ToR in the Clos fabric. An EVPN PE is
  typically also an IP or MPLS tunnel end-point for overlay VPN
  flow

 *Symmetric EVPN-IRB: An overlay fabric first-hop routing
  architecture as defined in [EVPN-IRB], wherein, overlay host-to-
  host routed inter-subnet flows are routed at both ingress and
  egress EVPN PEs.

 *Asymmetric EVPN-IRB: An overlay fabric first-hop routing
  architecture as defined in [EVPN-IRB], wherein, overlay host-to-
  host routed inter-subnet flows are routed and bridged at ingress
  PE and bridged at egress PEs.

 *ARP: Address Resolution Protocol [RFC 826]. ARP references in
  this document are equally applicable to ND as well.

*ND: IPv6 Neighbor Discovery Protocol [RFC 4861].

*Ethernet-Segment: physical Ethernet or LAG port that connects an
 access device to an EVPN PE, as defined in [RFC 7432].

*ESI: Ethernet Segment Identifier as defined in [RFC 7432].

*LAG: Layer-2 link-aggregation, also known as layer-2 bundle port-
 channel, or bond interface.

*EVPN all-active multi-homing: PE-CE all-active multi-homing
 achieved via a multi-homed layer-2 LAG interface on a CE with
 member links to multiple PEs and related EVPN procedures on the
 PEs.

*RT-2: EVPN route type 2 carrying both MAC and IP reachability.

*RT-5: EVPN route type 5 carrying IP prefix reachability.

*MAC-IP: IP association for a MAC, referred to in this document
 may be IPv4, IPv6 or both.

*SYNC MAC route: In the context of EVPN multi-homing, this refers
 to a local MAC route SYNCed from another PE sharing the same ESI.

*SYNC MAC-IP route: In the context of EVPN multi-homing, this
 refers to a local MAC-IP route SYNCed from another PE sharing the
 same ESI.

*SYNC MAC sequence number: In the context of EVPN multi-homing,
 this refers to sequencee number received with a SYNC MAC route.

*SYNC MAC-IP sequence number: In the context of EVPN multi-homing,
 this refers to sequencee number received with a SYNC MAC-IP
 route.

2. Introduction

EVPN-IRB enables capability to advertise both MAC and IP routes via
a single MAC+IP RT-2 advertisement. MAC is imported into local
bridge MAC table and enables L2 bridged traffic across the network
overlay. IP is imported into the local ARP table in an asymmetric
IRB design OR imported into the IP routing table in a symmetric IRB
design, and enables routed traffic across the layer 2 network
overlay. Please refer to [EVPN-IRB] for more background on EVPN IRB
forwarding modes.

To support EVPN mobility procedure, a single sequence number
mobility attribute is advertised with the combined MAC+IP route. A
single sequence number advertised with the combined MAC+IP route to

resolve both MAC and IP reachability implicitly assumes a 1:1 fixed mapping between IP and MAC. While a fixed 1:1 mapping between IP and MAC is a common use case that could be addressed via existing MAC mobility procedure, additional IRB scenarios need to be considered, that don't necessarily adhere to this assumption. Following IRB mobility scenarios are considered:

  *VM move results in VM IP and MAC moving together

  *VM move results in VM IP moving to a new MAC association

  *VM move results in VM MAC moving to a new IP association

While existing MAC mobility procedure can be leveraged for MAC+IP move in the first scenario, subsequent scenarios result in a new MAC- IP association. As a result, a single sequence number assigned independently per-[MAC, IP] is not sufficient to determine most recent reachability for both MAC and IP, unless the sequence number assignment algorithm is designed to allow for changing MAC-IP bindings across moves.

Purpose of this draft is to define additional sequence number assignment and handling procedures to adequately address generic mobility support across EVPN-IRB overlay use cases that allow MAC-IP bindings to change across VM moves and can support mobility for both MAC and IP components carried in an EVPN RT-2 for these use cases.

In addition, for hosts on an ESI multi-homed to multiple GW devices, additional procedure is proposed to ensure synchronized sequence number assignments across the multi-homing devices.

Content presented in this draft is independent of data plane encapsulation used in the overlay being MPLS or NVO Tunnels. It is also largely independent of the EVPN IRB solution being based on symmetric OR asymmetric IRB design as defined in [EVPN-INTER-SUBNET].

In addition to symmetric and asymmetric IRB, mobility solution for a routed overlay, where traffic to an end host in the overlay is always IP routed using EVPN RT-5 is also presented in this document.

To summarize, this draft covers mobility mobility for the following independent of the overlay encapsulation being MPLS or an NVO Tunnel:

  *Symmetric EVPN IRB overlay

  *Asymmetric EVPN IRB overlay

  *Routed EVPN overlay

2.1. Document Structure

Following sections of the document should be condidered informnative:

   *section 4 and 5 provide the necessary background and problem
    statement being addressed in this document.

   *section 6 lists the resulting design considerations for the
    document.

Following sections of the document should be condidered normative:

   *section 8 describes the mobility and sequence number assigment
    procedures in an EVPN-IRB overlay required to address the
    scenarios described in section 4.

   *section 9 describes the mobility procedures for a routed overlay
    nextwork as opposed to an IRB overlay.

   *section 10 describes corresponding duplicate detection procedures
    for EVPN-IRB and routed overlays.

3. 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 is
redundant for host MACs that are advertised via MAC+IP RT-2. As a
result, a MAC only RT-2 is an optional route that may not be
advertised from or received at an EVPN PE. This is an important
consideration for mobility scenarios discussed in subsequent
sections.

MAC only RT-2 may still be advertised for non-IP host MACs that are
not advertised via MAC+IP RT-2.

4. Mobility Use Cases

This section describes the IRB mobility use cases considered in this
document. Procedures to address them are covered later in section 6
and section 7.

   *Host move results in Host IP and MAC moving together

   *Host move results in Host IP moving to a new MAC association

   *Host move results in Host MAC moving to a new IP association

4.1. Host MAC+IP Move

This is the baseline case, wherein a host move results in both host MAC and IP moving together with no change in MAC-IP binding across a move. Existing MAC mobility defined in RFC 7432 may be leveraged to apply to corresponding MAC+IP route to support this mobility scenario.

4.2. Host IP Move to new MAC

This is the case, where a host move results in VM IP moving to a new MAC binding.

4.2.1. VM Reload

A host reload or an orchestrated host move that results in host being re-spawned at a new location may result in host getting a new MAC assignment, while maintaining existing IP address. This results in a host IP move to a new MAC binding:

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

4.2.2. MAC Sharing

This takes into account scenarios, where multiple hosts, each with a unique IP, may share a common MAC binding, and a host move results in a new MAC binding for the host IP.

As an example, hosts running on a single physical server, each with a unique IP, may share the same physical server MAC. In yet another scenario, an L2 access network may be behind a firewall, such that all hosts IPs on the access network are learnt with a common firewall MAC. In all such "shared MAC" use cases, multiple local MAC-IP ARP entries may be learnt with the same MAC. A host IP move, in such scenarios (for e.g., to a new physical server), could result in new MAC association for the host IP.

4.2.3. Problem

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

```
                  +------------------------+
                  | Underlay Network Fabric|
                  +------------------------+

    +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
    | PE1 |   | PE2 |      | PE3 |   | PE4 |      | PE5 |   | PE6 |
    +-----+   +-----+      +-----+   +-----+      +-----+   +-----+
      \          /           \         /           \          /
       \ ESI-1  /             \ ESI-2 /             \ ESI-3  /
        \      /               \     /               \      /
        +\---/+                +\---/+                +\---/+
        | \ / |                | \ / |                | \ / |
        +--+--+                +--+--+                +--+--+
           |                      |                      |
      Server-MAC1            Server-MAC2            Server-MAC3
           |                      |                      |
    [VM-IP1, VM-IP2]       [VM-IP3, VM-IP4]       [VM-IP5, VM-IP6]
```
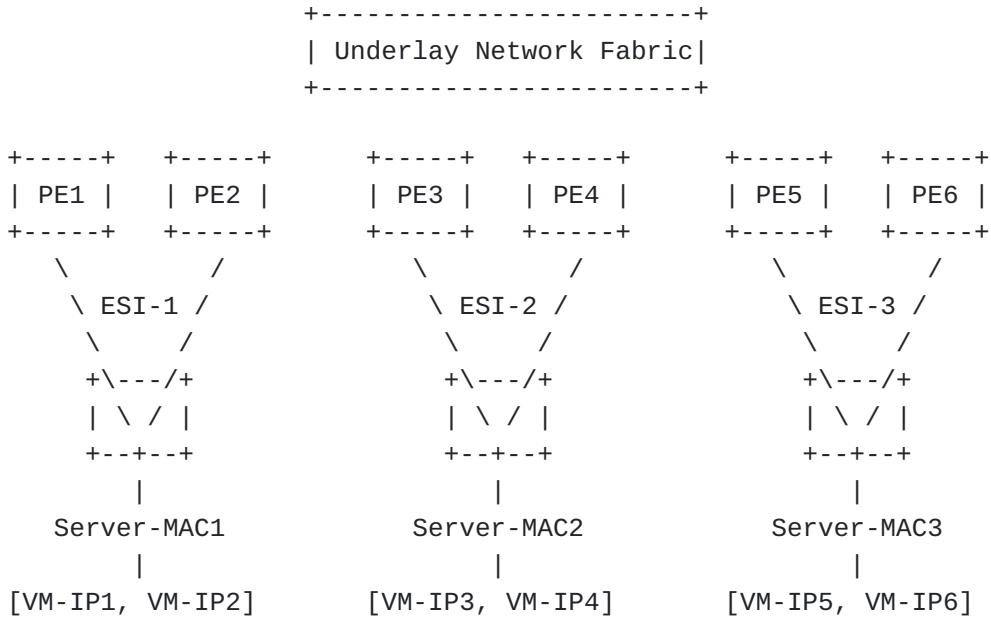
Figure 1

As an example, consider a topology shown in Figure 1, with host VMs
sharing the physical server MAC. In steady state, [IP1, MAC1] route
is learnt at [PE1, PE2] and advertised to remote PEs with a sequence
number N. Now, VM-IP1 is moved to Server-MAC2. ARP or ND based local
learning at [PE3, PE4] would now result in a new [IP1, MAC2] route
being learnt. If route [IP1, MAC2] is learnt as a new MAC+IP route
and assigned a new sequence number of say 0, mobility procedure for
VM-IP1 will not trigger across the overlay network.

A sequence number assignment procedure needs to be defined to
unambiguously determine the most recent IP reachability, IP to MAC
binding, and MAC reachability for such a MAC sharing scenario.

4.3. Host MAC move to new IP

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

4.3.1. Problem

Complication with this scenario is that MAC reachability could be
carried via a combined MAC+IP route while a MAC only route may not
be advertised at all. A single sequence number association with the
MAC+IP route again implicitly assumes a fixed mapping between MAC
and IP. A MAC move resulting in a new IP association for the host
MAC breaks this assumption and results in a new MAC+IP route. If
this new MAC+IP route independently assumes a new sequence number,

this mobility attribute can no longer be used to determine most
recent host MAC reachability.

```
                  +------------------------+
                  | Underlay Network Fabric|
                  +------------------------+
+-----+   +-----+     +-----+   +-----+     +-----+   +-----+
| PE1 |   | PE2 |     | PE3 |   | PE4 |     | PE5 |   | PE6 |
+-----+   +-----+     +-----+   +-----+     +-----+   +-----+
   \         /           \         /           \         /
    \ ESI-1 /             \ ESI-2 /             \ ESI-3 /
     \     /               \     /               \     /
     +\---/+               +\---/+               +\---/+
     | \ / |               | \ / |               | \ / |
     +--+--+               +--+--+               +--+--+
        |                     |                     |
      Server1               Server2               Server3
        |                     |                     |
[VM-IP1-M1, VM-IP2-M2]  [VM-IP3-M3, VM-IP4-M4] [VM-IP5-M5, VM-IP6-M6]
```
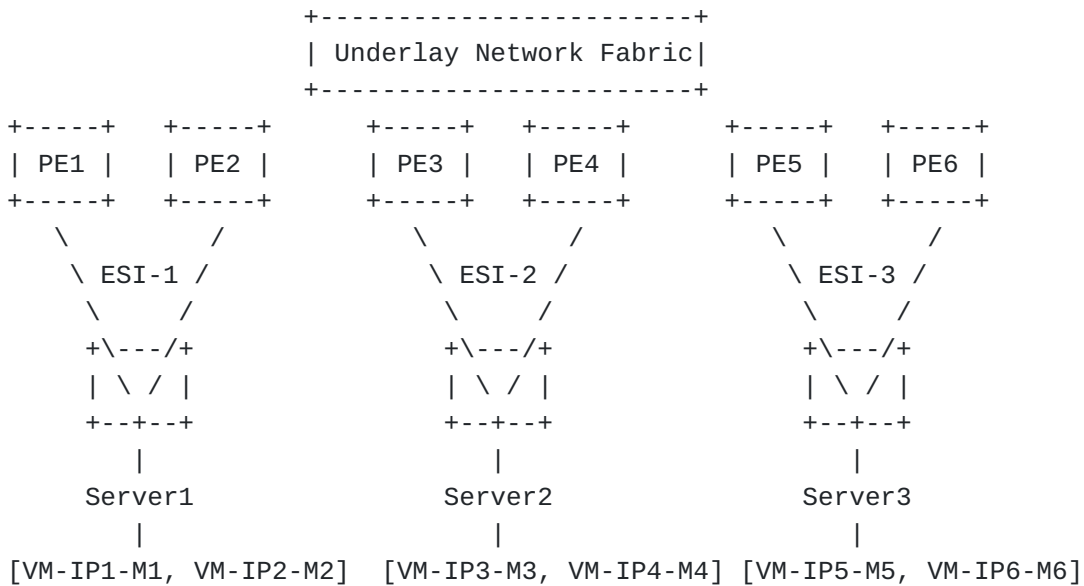
Figure 2

As an example, consider a host VM IP1-M1 that is learnt locally at
[PE1, PE2] and advertised to remote hosts with a sequence number N.
Consider a scenario where this VM with MAC M1 is re-provisioned at
server 2, however, as part of this re-provisioning, assigned a
different IP address say IP7. [IP7, M1] is learnt as a new route at
[PE3, PE4] and advertised to remote PEs with a sequence number of 0.
As a result, L3 reachability to IP7 would be established across the
overlay, however, MAC mobility procedure for MAC1 will not trigger
as a result of this MAC-IP route advertisement. If an optional MAC
only route is also advertised, sequence number associated with the
MAC only route would trigger MAC mobility as per [RFC7432]. However,
in the absence of an additional MAC only route advertisement, a
single sequence number advertised with a combined MAC+IP route may
not be sufficient to update MAC reachability across the overlay.

A MAC-IP sequence number assignment procedure needs to be defined to
unambiguously determine the most recent MAC reachability in such a
scenario without a MAC only route being advertised.

Further, PE1/PE2, on learning new reachability for [IP7, M1] via
PE3/PE4 MUST probe and delete any local IPs associated with MAC M1,
such as [IP1, M1] in the above example.

Arguably, MAC mobility sequence number defined in [RFC7432], could
be interpreted to apply only to the MAC part of MAC-IP route, and
would hence cover this scenario. It could hence be interpreted as a

clarification to [RFC7432] and one of the considerations for a
common sequence number assignment procedure across all MAC-IP
mobility scenarios detailed in this document.

5. **EVPN All Active multi-homed ES**

```
                    +------------------------+
                    | Underlay Network Fabric|
                    +------------------------+

                  +-----+   +-----+
                  | PE1 |   | PE2 |
                  +-----+   +-----+
                    \\          //
                     \\ ESI-1 //
                      \\      /X
                     +\\---//+
                     | \\ // |
                     +---+---+
                         |
                        CE1
```
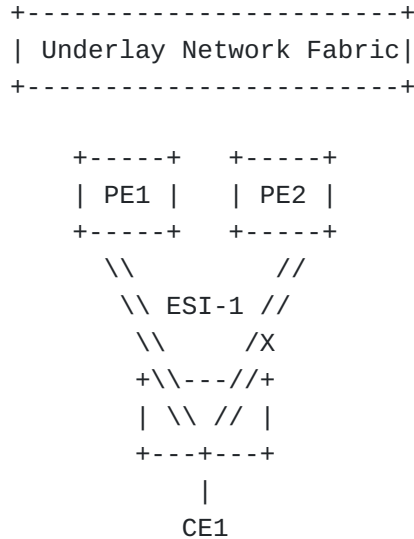
                        Figure 3

Consider an EVPN-IRB overlay network shown in Figure 2, with hosts
multi-homed to two or more PE devices via an all-active multi-homed
ES. MAC and ARP entries learnt on a local ES may also be
synchronized across the multi-homing PE devices sharing this ES.
This MAC and ARP SYNC enables local switching of intra and inter
subnet ECMP traffic flows from remote hosts. In other words, local
MAC and ARP entries on a given ES may be learnt via local learning
and / or via sync 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, local learning of host MAC and MAC-IP at each
PE device is an independent asynchronous event, that is dependent on
traffic flow and or ARP / ND response from the host hashing to a
directly connected PE on the MC-LAG interface. As a result, sequence
number mobility attribute value assigned to a locally learnt 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.

As an example, consider a host VM that is deleted from ESI-2 and
moved to ESI-1. It is possible for host to be learnt on say, PE1
following deletion of the remote route from [PE3, PE4], while being
learnt on PE2 prior to deletion of remote route from [PE3, PE4]. If
so, 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, N being the existing sequence number assigned at
[PE3, PE4].

Inconsistent sequence numbers advertised from multi-homing devices
introduces:

  *Ambiguity with respect to how the remote PEs should handle paths
   with same ESI and different sequence numbers. A remote PE may not
   program ECMP paths if it receives routes with different sequence
   numbers from a set of multi-homing PEs sharing the same ESI.

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

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

There is need for a mechanism to ensure consistency of sequence
numbers advertised from a set of multi-homing devices for EVPN
mobility to work reliably.

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

6. **Design Considerations**

   To summarize, sequence number assignment scheme and implementation
   must take following considerations into account:

     *MAC+IP may be learnt on an ES multi-homed to multiple PE devices,
      hence requires sequence numbers to be synchronized across multi-
      homing PE devices.

     *MAC only RT-2 is optional in an IRB scenario and may not
      necessarily be advertised in addition to MAC+IP RT-2.

     *Single MAC may be associated with multiple IPs, i.e., multiple
      host IPs may share a common MAC.

     *Host IP move could result in host moving to a new MAC, resulting
      in a new IP to MAC association and a new MAC+IP route.

*Host MAC move to a new location could result in host MAC being
 associated with a different IP address, resulting in a new MAC to
 IP association and a new MAC+IP route.

*LOCAL MAC-IP learn via ARP would always accompanied by a LOCAL
 MAC learn event resulting from the ARP packet. MAC and MAC-IP
 learning, however, could happen in any order.

*Use cases discussed earlier that do not maintain a constant 1:1
 MAC-IP mapping across moves could potentially be addressed by
 using separate sequence numbers associated with MAC and IP
 components of MAC+IP route. Maintaining two separate sequence
 numbers however adds significant overhead with respect to
 complexity, debugability, and backward compatibility. Hence, this
 document addresses these requirements via a single sequence
 number attribute.

7. Solution Components

This section goes over main components of the EVPN IRB mobility
solution proposed in this draft. Later sections will go over exact
sequence number assignment procedures resulting from concepts
described in this section.

7.1. Sequence Number Inheritance

Main idea presented here is to view a LOCAL MAC-IP route as a child
of the corresponding LOCAL MAC only route that inherits the sequence
number attribute from the parent LOCAL MAC only route:

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

As a result, both parent MAC and child MAC-IP routes share one
common sequence number associated with the parent MAC route. Doing
so ensures that a single sequence number attribute carried in a
combined MAC+IP route represents sequence number for both a MAC only
route as well as a MAC+IP route, and hence makes the MAC only route
truly optional. As a result, optional MAC only route with its own
sequence number is not required to establish most recent
reachability for a MAC in the overlay network. Specifically, this
enables a MAC to assume a different IP address on a move, and still
be able to establish most recent reachability to the MAC across the
overlay network via mobility attribute associated with the MAC+IP
route advertisement. As an example, when Mx moves to a new location,
it would result in LOCAL Mx being assigned a higher sequence number
at its new location as per RFC 7432. If this move results in Mx
assuming a different IP address, IPz, LOCAL Mx+IPz route would
inherit the new sequence number from Mx.

LOCAL MAC and LOCAL MAC-IP routes would typically be sourced from
data plane learning and ARP learning respectively, and could get
learnt in control plane in any order. Implementation could either
replicate inherited sequence number in each MAC-IP entry OR maintain
a single attribute in the parent MAC by creating a forward reference
LOCAL MAC object for cases where a LOCAL MAC-IP is learnt before the
LOCAL MAC.

Arguably, this inheritance may be assumed from RFC 7432, in which
case, the above may be interpreted as a clarification with respect
to interpretation of a MAC sequence number in a MAC-IP route.

7.2. MAC Sharing

Further, for the shared MAC scenario, this would result in multiple
LOCAL MAC-IP siblings inheriting sequence number attribute from a
common parent MAC route:

```
Mx-IP1 -----
 |          |
Mx-IP2 -----
   .        |
   .        +---> Mx (seq# = N)
   .        |
Mx-IPw -----
 |          |
Mx-IPx -----
```
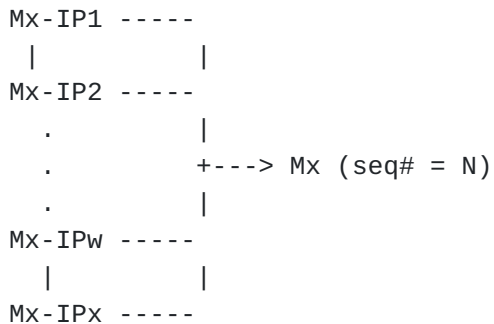
                           Figure 4

In such a case, a host-IP move to a different physical server would
result in IP moving to a new MAC binding. A new MAC-IP route
resulting from this move must now be advertised with a sequence
number that is higher than the previous MAC-IP route for this IP,
advertised from the prior location. As an example, consider a route
Mx-IPx that is currently advertised with sequence number N from PE1.
IPx moving to a new physical server behind PE2 results in IPx being
associated with MAC Mz. A new local Mz-IPx route resulting from this
move at PE2 must now be advertised with a sequence number higher
than N. This is so that PE devices, including PE1, PE2, and other
remote PE devices that are part of the overlay can clearly determine
and program the most recent MAC binding and reachability for the IP.
PE1, on receiving this new Mz-IPx route with sequence number say,
N+1, for symmetric IRB case, would update IPx reachability via PE2
in forwarding, for asymmetric IRB case, would update IPx's ARP
binding to Mz. In addition, PE1 would clear and withdraw the stale
Mx-IPx route with the lower sequence number.

This also implies that sequence number associated with local MAC Mz and all local MAC-IP children of Mz at PE2 must now be incremented to N+1, and re-advertised across the overlay. While this re-advertisement of all local MAC-IP children routes affected by the parent MAC route is an overhead, it avoids the need for two separate sequence number attributes to be maintained and advertised for IP and MAC components of MAC+IP RT-2. Implementation would need to be able to lookup MAC-IP routes for a given IP and update sequence number for it's parent MAC and its MAC-IP children.

7.3. Multi-homing Mobility Synchronization

In order to support mobility for multi-homed hosts, local MAC and MAC-IP routes learnt on a shared ES MUST be advertised with the same sequence number by all PE devices that the ES is multi-homed to. This also applies to local MAC only routes. LOCAL MAC and MAC-IP may be learnt natively via data plane and ARP/ND respectively as well as via SYNC from another multi-homing PE to achieve local switching. Local and SYNC route learning can happen 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 learnt. This implies:

  *On local or SYNC MAC-IP route learning, sequence number for the
   local MAC-IP route MUST be compared and updated to the higher
   value.

  *On local or SYNC MAC route learning, sequence number for the
   local MAC route MUST be compared and updated to the higher value.

If an update to local MAC-IP sequence number is required as a result of above comparison with SYNC MAC-IP route, it would essentially amount to a sequence number update on the parent local MAC, resulting in inherited sequence number update on the MAC-IP route.

8. Requirements for Sequence Number Assignment

Following sections summarize sequence number assignment procedure needed on local and SYNC MAC and MAC-IP route learning events in order to accomplish the above.

8.1. LOCAL MAC-IP learning

A local Mx-IPx learning via ARP or ND should result in computation OR re-computation of parent MAC Mx's sequence number, following which the MAC-IP route Mx-IPx would simply inherit parent MAC's

sequence number. Parent MAC Mx Sequence number should be computed as
follows:

  *MUST be higher than any existing remote MAC route for Mx, as per
   RFC 7432.

  *MUST be at least equal to corresponding SYNC MAC sequence number
   if one is present.

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

Once new sequence number for MAC route Mx is computed as per above,
all LOCAL MAC-IPs associated with MAC Mx MUST inherit the updated
sequence number.

8.2. LOCAL MAC learning

Local MAC Mx Sequence number should be computed as follows:

  *MUST be higher than any existing remote MAC route for Mx, as per
   RFC 7432.

  *MUST be at least equal to corresponding SYNC MAC sequence number
   if one is present.

  *Once new sequence number for MAC route Mx is computed as per
   above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the
   updated sequence number.

Note that the local MAC sequence number might already be present if
there was a local MAC-IP learnt prior to the local MAC, in which
case the above may not result in any change in local MAC's sequence
number.

8.3. Remote MAC OR MAC-IP Update

On receiving a remote MAC OR MAC-IP route update associated with a
MAC 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 because of lower
   VTEP IP as per [RFC 7432],

following handling is required on the receiving PE:

   *PE MUST trigger probe and deletion procedure for all LOCAL IPs
    associated with MAC Mx.

   *PE MUST trigger deletion procedure for LOCAL MAC route for Mx.

8.4. REMOTE (SYNC) MAC update

Corresponding local MAC Mx (if present) sequence number should be
re- computed as follows:

   *If the current sequence number is less than the received SYNC MAC
    sequence number, it MUST be increased to be equal to received
    SYNC MAC sequence number.

   *If a LOCAL MAC sequence number is updated as a result of the
    above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the
    updated sequence number.

8.5. REMOTE (SYNC) MAC-IP update

If this is a SYNCed MAC-IP on a local ES, it would also result in a
derived SYNC MAC Mx route entry, as MAC only RT-2 advertisement is
optional. Corresponding local MAC Mx (if present) sequence number
should be re-computed as follows:

   *If the current sequence number is less than the received SYNC MAC
    sequence number, it MUST be increased to be equal to received
    SYNC MAC sequence number.

   *If a LOCAL MAC sequence number is updated as a result of the
    above, all LOCAL MAC-IPs associated with MAC Mx MUST inherit the
    updated sequence number.

8.6. Inter-op

In general, if all PE nodes in the overlay network follow the above
sequence number assignment procedure, and the PE is advertising both
MAC+IP and MAC routes, sequence number advertised with the MAC and
MAC+IP routes with the same MAC would always be the same. However,
an inter-op scenario with a different implementation could arise,
where a PE implementation non-compliant with this document or with
RFC 7432 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 and corresponding remote MAC

routes from a remote PE, sequence number associated with the remote
MAC route should be computed as:

  *Highest of the all received sequence numbers with remote MAC+IP
   and MAC routes with the same MAC.

  *MAC sequence number would be re-computed on a MAC or MAC+IP route
   withdraw as per above.

A MAC and / or IP move to the local PE would now result in the MAC
(and hence all MAC-IP) sequence numbers incremented from the above
computed remote MAC 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,
sequence number associated with the derived remote MAC route should
still be computed as the highest of the all received MAC+IP sequence
numbers with the same MAC.

8.7. MAC Sharing Race Condition

In a MAC sharing use case described in section 6.2, a race condition
is possible with simultaneous host moves between a pair of PEs. As
an example, consider PE1 with local host IPs I1 and I2 sharing MAC
M1, and PE2 with local host IPs I3 and I4 sharing MAC M2. A
simultaneous move of I1 from PE1 to PE2 and of I3 from PE2 to PE1,
such that I3 is learnt on PE1 before I1's local entry has been
probed out on PE1 and/or I1 is learnt on PE2 before I3's local entry
has been probed out on PE2 may trigger a race condition. This race
condition together with MAC sequence number assignment rules defined
in section 7.1 can cause new mac-ip routes [I1, M2] and [I3, M1] to
bounce a couple of times with an incremented sequence number 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, ND, and local MAC entries, following
a move, on learning remote routes as defined in section 7.3 (and as
per [EVPN-IRB]) to minimize exposure to this race condition.

8.8. Mobility Convergence

This sections is to be treated as optional and details ARP and ND
probing procedures that MAY be implemented to achieve faster host
re- learning and convergence on mobility events.

  *Following a host move from PE1 to PE2, the host's MAC is
   discovered at PE2 as a local MAC via a data frames received from
   the host. If PE2 has a prior REMOTE MAC-IP host route for this
   MAC from PE1, an ARP/ND probe MAY be triggered at PE2 to learn
   the MAC-IP as a local adjacency and trigger EVPN RT-2
   advertisement for this MAC-IP across the overlay with new

reachability via PE2. This results in a reliable "event based" host IP learning triggered by a "MAC learning event" across the overlay, and hence 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/ND probe MAY be triggered at PE1 to clear the stale local MAC-IP neighbor adjacency OR re-learn the local MAC-IP in case the host has moved back or is duplicate.

*Following a local MAC age-out, if there is a local IP adjacency with this MAC, an ARP/ND probe MAY be triggered for this IP to either re-learn the local MAC and maintain local l3 and l2 reachability to this host OR to clear the ARP/ND entry in case the host is indeed no longer local. Note that this accomplishes clearing of stale ARP entries, triggered by a MAC age-out event even when the ARP refresh timer was longer than the MAC age-out timer. Clearing of stale IP neighbor entries in-turn facilitates traffic convergence in the event that the host was silent and not discovered at its new location. Once stale neighbor entry for the host is cleared, routed traffic flow destined for the host can re-trigger ARP/ND discovery for this host at the new location.

8.8.1. Generalized Probing Logic

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 being inconsistent is defined as being not present OR conflicting in terms of the route source being local OR remote. MAC-IP to MAC parent relationship described earlier in this document in section 6.1 MAY be used to achieve this logic.

9. Routed Overlay

An additional use case is possible, such that traffic to an end host in the overlay is always IP routed. In a purely routed overlay such as this:

*A host MAC is never advertised in 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 [RFC 7814] for more details.

Host mobility within the stretched subnet would still need to be
supported for this use. In the absence of any host MAC routes,
sequence number mobility EXT-COMM specified in [RFC7432], section
7.7 may be associated with a /32 OR /128 host IP prefix advertised
via EVPN route type 5. MAC mobility procedures defined in RFC 7432
can now be applied as is to host IP prefixes:

  *On LOCAL learning of a host IP, on a new ESI, host IP MUST be
   advertised with a sequence number attribute that is 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/ND probe and deletion procedure on
   any LOCAL route for that IP with a lower sequence number. A PE
   would essentially move the forwarding entry to point to the
   remote route with a higher sequence number and send an ARP/ND
   PROBE for the local IP route. If the IP has indeed moved, PROBE
   would timeout and the local IP host route would be deleted.

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

Note that this mobility procedure would not apply to "anycast IPv6"
hosts advertised via NA messages with 0-bit=0. Please refer to
[EVPN-PROXY-ARP].

10. **Duplicate Host Detection**

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

     *Scenario A: where two hosts have the same MAC (host IPs may or
      may not be duplicate).

     *Scenario B: where two hosts have the same IP but different MACs.

     *Scenario C: where two hosts have the same IP and host MAC is not
      advertised at all.

   Duplicate detection procedures for scenario B and C would not apply
   to "anycast IPv6" hosts advertised via NA messages with 0-bit=0.
   Please refer to [EVPN-PROXY-ARP].

10.1. Scenario A

For all use cases where duplicate hosts have the same MAC, MAC is
detected as duplicate via duplicate MAC detection procedure
described in RFC 7432. Corresponding MAC-IP routes with the same MAC
do not require duplicate detection and MUST simply inherit the
DUPLICATE property from the corresponding MAC route. In other words,
if a MAC route is in DUPLICATE state, all corresponding MAC-IP
routes MUST also be treated as DUPLICATE. Duplicate detection
procedure need only be applied to MAC routes.

10.2. Scenario B

Due to misconfiguration, a situation may arise where hosts with
different MACs are configured with the same IP. This scenario would
not be detected by existing duplicate MAC detection procedure and
would result in incorrect forwarding of routed traffic destined to
this IP.

Such a situation, on LOCAL MAC-IP learning, would be detected as a
move scenario via the following local MAC sequence number
computation procedure described earlier in section 6.1:

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

Such a move that results in sequence number increment on local MAC
because of a remote MAC-IP route associated with a different MAC
MUST be counted as an "IP move" against the "IP" independent of MAC.
Duplicate detection procedure described in RFC 7432 can now be
applied to an "IP" entity independent of MAC. Once an IP is detected
as DUPLICATE, corresponding MAC-IP route should be treated as
DUPLICATE. Associated MAC routes and any other MAC-IP routes
associated with this MAC should not be affected.

10.2.1. Duplicate IP Detection Procedure for Scenario B

Duplicate IP detection procedure for such a scenario is specified in
[EVPN-PROXY-ARP]. What counts as an "IP move" in this scenario is
further clarified as follows:

  *On learning a LOCAL MAC-IP route Mx-IPx, check if there is an
   existing REMOTE OR LOCAL route for IPx with a different MAC
   association, say, Mz-IPx. If so, count this as an "IP move" count
   for IPx, independent of the MAC.

  *On learning a REMOTE MAC-IP route Mz-IPx, check if there is an
   existing LOCAL route for IPx with a different MAC association,
   say, Mx-IPx. If so, count this as an "IP move" count for IPx,
   independent of the MAC.

A MAC-IP route SHOULD be treated as DUPLICATE if either of the
following two conditions are met:

  *Corresponding MAC route is marked as DUPLICATE via existing
   duplicate detection procedure.

  *Corresponding IP is marked as DUPLICATE via extended procedure
   described above.

10.3. Scenario C

For a purely routed overlay scenario described in section 8, where
only a host IP is advertised via EVPN RT-5, together with a sequence
number mobility attribute, duplicate MAC detection procedures
specified in RFC 7432 can be intuitively applied to IP only host
routes for the purpose of duplicate IP detection.

  *On learning a LOCAL host IP route IPx, check if there is an
   existing REMOTE OR LOCAL route for IPx with a different ESI
   association. If so, count this as an "IP move" count for IPx.

  *On learning a REMOTE host IP route IPx, check if there is an
   existing LOCAL route for IPx with a different ESI association. If
   so, count this as an "IP move" count for IPx.

  *With configurable parameters "N" and "M", If "N" IP moves are
   detected within "M" seconds for IPx, treat IPx as DUPLICATE.

10.4. Duplicate Host Recovery

Once a MAC or IP is marked as DUPLICATE and FROZEN, corrective
action must be taken to un-provision one of the duplicate MAC or IP.
Un- provisioning a duplicate MAC or IP in this context refers to a
corrective action taken on the host side. Once one of the duplicate
MAC or IP is un-provisioned, normal operation would not resume until
the duplicate MAC or IP ages out, following this correction, unless
additional action is taken to speed up recovery.

This section lists possible additional corrective actions that could
be taken to achieve faster recovery to normal operation.

10.4.1. Route Un-freezing Configuration

Unfreezing the DUPLICATE OR FROZEN MAC or IP via a CLI can be
leveraged to recover from DUPLICATE and FROZEN state following
corrective un-provisioning of the duplicate MAC or IP.

Unfreezing the frozen MAC or IP via a CLI at a PE should result in
that MAC OR IP being advertised with a sequence number that is

higher than the sequence number advertised from the other location
of that MAC or IP.

Two possible corrective un-provisioning scenarios exist:

  *Scenario A: A duplicate MAC or IP may have been un-provisioned at
   the location where it was NOT marked as DUPLICATE and FROZEN.

  *Scenario B: A duplicate MAC or IP may have been un-provisioned at
   the location where it was marked as DUPLICATE and FROZEN.

Unfreezing the DUPLICATE and FROZEN MAC or IP, following the above
corrective un-provisioning scenarios would result in recovery to
steady state as follows:

  *Scenario A: If the duplicate MAC or IP was un-provisioned at the
   location where it was NOT marked as DUPLICATE, unfreezing the
   route at the FROZEN location will result in the route being
   advertised with a higher sequence number. This would in-turn
   result in automatic clearing of local route at the PE location,
   where the host was un-provisioned via ARP/ND PROBE and DELETE
   procedure specified earlier in section 8 and in [RFC 7432].

  *Scenario B: If the duplicate host is un-provisioned at the
   location where it was marked as DUPLICATE, unfreezing the route
   will trigger an advertisement with a higher sequence number to
   the other location. This would in-turn trigger re-learning of
   local route at the remote location, resulting in another
   advertisement with a higher sequence number from the remote
   location. Route at the local location would now be cleared on
   receiving this remote route advertisement, following the ARP/ND
   PROBE.

Note that the probes referred to in the above scenarios are event
driven probes resulting from receiving a route with a higher
sequence number. Periodic probes resulting from refresh timers may
also occur in addition as completely independent probes.

10.4.2. Route Clearing Configuration

In addition to the above, route clearing CLIs may also be leveraged
to clear the local MAC or IP route, to be executed AFTER the
duplicate host is un-provisioned:

  *clear mac CLI: A clear MAC CLI can be leveraged to clear a
   DUPLICATE MAC route, to recover from a duplicate MAC scenario.

  *clear ARP/ND: A clear ARP/ND CLI may be leveraged to clear a
   DUPLICATE IP route to recover from a duplicate IP scenario.

Note that the route unfreeze CLI may still need to be run if the route was un-provisioned and cleared from the NON-DUPLICATE / NON-FROZEN location. Given that unfreezing of the route via the un-freeze CLI would any ways result in auto-clearing of the route from the "un- provisioned" location, as explained in the prior section, need for a route clearing CLI for recovery from DUPLICATE / FROZEN state is truly optional.

11. Security Considerations

This document raises no new security issues for EVPN.

12. IANA Considerations

None.

13. Acknowledgements

Authors would like to thank Vibov Bhan and Patrice Brisset for feedback the process of design and implementation of procedures defined in this document. Authors would like to thank Wen Lin for a detailed review and valuable comments related to MAC sharing race conditions. Authors would also like to thank Saumya Dikshit for a detailed review and valuable comments across the document.

14. Normative References

[EVPN-IRB] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J.
            Rabadan, "Integrated Routing and Bridging in EVPN", Work
            in Progress, Internet-Draft, draft-ietf-bess-evpn-inter-
            subnet-forwarding-13, 10 February 2021, <http://
            www.ietf.org/internet-drafts/draft-ietf-bess-evpn-inter-
            subnet-forwarding-13.txt>.

[EVPN-PROXY-ARP] Rabadan, J., Sathappan, S., Nagaraj, K., Hankins,
            G., and T. King, "Operational Aspects of Proxy-ARP/ND in
            EVPN Networks", Work in Progress, Internet-Draft, draft-
            ietf-bess-evpn-proxy-arp-nd-11, 7 January 2021, <https://
            tools.ietf.org/html/draft-ietf-bess-evpn-proxy-arp-
            nd-11.txt>.

[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>.

[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://tools.ietf.org/html/rfc7814>.

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
ATT
200 S. Laurel Avenue
Middletown, CA 07748
United States of America

Email: ar977m@att.com

John Drake
Juniper Networks

Email: jdrake@juniper.net