                   Proxy IP->MAC Advertisement in EVPNs
                  draft-rbickhart-evpn-ip-mac-proxy-adv-00

   Abstract

      This document specifies procedures for EVPN PEs connected to a common
      multihomed site to generate proxy EVPN type 2 IP->MAC advertisements
      on behalf of other PEs to facilitate preservation of ARP/ND state
      across link or node failures.

   Status of This Memo

      This Internet-Draft is submitted in full conformance with the
      provisions of BCP 78 and BCP 79.

      Internet-Drafts are working documents of the Internet Engineering
      Task Force (IETF).  Note that other groups may also distribute
      working documents as Internet-Drafts.  The list of current Internet-
      Drafts is at https://datatracker.ietf.org/drafts/current/.

      Internet-Drafts are draft documents valid for a maximum of six months
      and may be updated, replaced, or obsoleted by other documents at any
      time.  It is inappropriate to use Internet-Drafts as reference
      material or to cite them other than as "work in progress."

      This Internet-Draft will expire on September 6, 2019.

Table of Contents

1.  Conventions and Terminology

   This document assumes familiarity with the terminology used in EVPN
   [RFC7432].  A few key terms used in this document are defined below:

   CE: Customer edge device.

   PE: Provider edge device.

   IP->MAC: An IP address associated with a MAC address.

   ARP: Address Resolution Protocol.

   ND: Neighbor Discovery Protocol.

   DF: Designated Forwarder.

   R-bit: Router Flag in NA messages, as per ND for IPv6 [RFC4861].

   O-bit: Override Flag in NA messages, as per ND for IPv6 [RFC4861].

2.  Introduction

   EVPN [RFC7432] allows for the distribution of IP->MAC bindings of
   connected hosts learned through IPv4 ARP and IPv6 neighbor discovery
   (ND) using type 2 MAC/IP advertisement routes.  When end hosts are
   connected to a multihomed site of an EVPN running in all-active mode,
   depending on the learning mechanisms of the multihoming PEs and the
   load balancing mechanisms implemented by the CE devices multihomed to
   the EVPN PEs, local learning of an IP->MAC may be limited to a subset

of the total number of multihoming PEs, possibly only a single PE
device.  In the steady state, the IP->MAC originally learned locally
on one or more of the set of multihoming PEs is synchronized to all
remaining PEs attached to the same multihoming site via the EVPN
control plane.  Once synchronized, the IP->MAC is available on each
multihoming PE as well as other remote PEs not connected to the
multihomed site for use in forwarding traffic or suppressing ARP or
ND messaging in the EVPN.

In the event of the complete failure of a multihoming PE or the
failure of a multihoming PE's link to the multihomed site, any
IP->MAC locally learned on that PE or locally attached link will be
invalidated and will be withdrawn from the EVPN control plane.  If
the source of the failure was the only origin of any particular
IP->MAC, that IP->MAC will completely disappear from the EVPN until
such time that one of the remaining multihoming PEs is able to
relearn the IP->MAC that was lost.  Traffic forwarding or other EVPN
features like ARP/ND suppression may fail during the intermediate
period between the loss of the IP->MAC from the original local
learning PE and later learning and distribution of the IP->MAC from a
new local learning PE.

This document specifies procedures to preserve IP->MAC state across
the remaining multihoming PEs in the EVPN to bridge the gap between
the initial failure and the subsequent relearning of the IP->MAC on
one of the remaining multihoming PEs.

3.  IP->MAC Proxy Advertisements

Preserving IP->MAC state in the EVPN until relearning and
distribution of the new IP->MAC state to all PEs is completed can be
accomplished by using IP->MAC proxy advertisements.  When an IP->MAC
for a host connected to a multihomed site is locally learned by a PE,
the PE will advertise the IP->MAC via an EVPN MAC/IP route as usual.
When other PEs learn that IP->MAC from the control plane upon
reception of the MAC/IP route, they will install the ARP/ND state
derived from the received IP->MAC for local use as usual.
Additionally, if the receiving PE is locally connected to the same
multihomed ethernet segment where the received IP->MAC originated and
the IP->MAC was not previously locally learned and advertised, the
receiving PE will inject its own EVPN MAC/IP route carrying the same
IP->MAC (and with the same ESI) into the control plane and mark that
injected route with a special proxy IP->MAC indication.  Assuming
that all PEs attached to the multihomed site support this proxy
advertisement functionality, the result is that each PE attached to
the site will originate the given IP->MAC using an EVPN MAC/IP route,
some of the route advertisements possibly carrying the proxy

indication and at least one route advertisement not marked with the proxy indication.

A subsequent PE failure, link failure, or other event triggering the loss of all non-proxy IP->MAC state on a multihoming PE will cause that PE to start an aging timer for the proxy IP->MAC the PE had previously advertised.  The aging timer should be initialized to the same age-time as the system default for ARP/ND aging, but an implementation may allow the initial age-time used for proxy a IP->MAC to be set administratively.  While the aging timer is running, the multihoming PE will take no other actions and will continue using the proxy IP->MAC state for local forwarding and ARP/ND purposes and will continue to advertise the IP->MAC in the control plane with the proxy indication set.  Remote PEs not connected to the multihomed site will ignore the proxy indication completely, and will be unaware of the difference between proxy and non-proxy IP->MAC advertisements.  In this state, the EVPN will continue working as before the failure, with the exception of the failed link or PE being removed from the forwarding path.

In the event that one of the remaining multihoming PEs now learns the IP->MAC locally, it will restart the aging timer for the IP->MAC with the default ARP/ND age-time and remove the proxy indication from the EVPN MAC/IP route for the IP->MAC previously advertised in the control plane.  When any other multihoming PE observes the removal of the proxy indication from at least one of the sources advertising the IP->MAC, that PE will stop the aging timer for the locally advertised proxy IP->MAC and continue advertising the IP->MAC with the proxy indication set as before.

In the event that a multihoming PE fails to learn the IP->MAC locally before the aging timer for the proxy IP->MAC expires, that PE will withdraw the EVPN MAC/IP route for proxy IP->MAC that it had advertised previously.  In this way, if all multihoming PEs fail to learn the IP->MAC locally within the age-time, the proxy IP->MAC advertisements will expire on every PE and will be withdrawn, completely removing the IP->MAC from the EVPN.

In the case that a non-proxy IP->MAC is withdrawn from the EVPN because the original dynamically learned ARP/ND entry ages out due to end host inactivity or shutdown rather than a PE node or link failure, PEs which advertised a proxy IP->MAC will still follow the same procedures as above and retain their proxy IP->MAC advertisements until the age-time has passed.  Implementations may allow the proxy IP->MAC age-time to be administratively specified separately from the regular system ARP/ND age-time to tune how fast stale proxy IP->MAC advertisements are cleared from the EVPN. Additionally, a PE may optionally use a mechanism like send-refresh

[I-D.ietf-bess-evpn-proxy-arp-nd] to probe the liveness of the
IP->MAC and withdraw the proxy IP->MAC from the control plane before
the age-time if the PE determines that the IP->MAC is no longer
active.

## 3.1.  Interoperation with Legacy PEs

A heterogenous mix of new PEs supporting proxy IP->MAC advertisement
and legacy PEs not supporting proxy IP->MAC advertisement is
supported in the event of incremental configuration of the feature or
incremental upgrades of PEs attached to the same ethernet segment.
Although legacy PE devices will continue to operate with the
traditional mechanisms and advertise only locally learned IP->MAC
entries, they can make use of any remotely learned proxy IP->MAC
advertised by other PEs supporting proxy advertisement.

## 3.2.  Single-Active Multihoming Considerations

Proxy IP->MAC advertisement is not applicable to ethernet segments
configured for single-active multihoming because MAC advertisements
are the indication of which multihoming PE is the DF for remote PEs
not directly connected ethernet segment.  Advertisement of a proxy
IP->MAC by a non-DF multihoming PE will prevent remote PEs not
directly attached to the ethernet segment from determining the
correct DF.
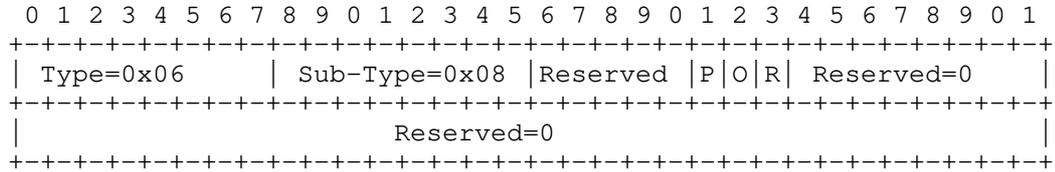
## 3.3.  MAC Route Attribute Considerations

When a PE advertises a proxy IP->MAC that was originally learned from
the control plane with a MAC mobility extended community attached
with a nonzero sequence number, the PE should advertise the new proxy
IP->MAC with the same sequence number as originally received.  The
presence or lack of a proxy indication for a received IP->MAC
advertisement should not be used in active MAC determination for MAC
mobility purposes.

When a PE advertises a proxy IP->MAC for an IPv6 address learned from
the control plane that has the 'R' or 'O' bits set in the EVPN ND
extended community, the new proxy IP->MAC should carry an EVPN ND
extended community with the same 'R' and 'O' bits as originally
received.

## 4.  EVPN ARP/ND Extended Community

EVPN already provides an extended community to signal additional
state relevant to IPv6 ND (the override and router bits).  Because
the proxy indication described in this document is equally applicable
to both ARP and ND, we propose renaming the EVPN Neighbor Discovery

(ND) Extended Community (type 0x08) to EVPN ARP/ND Extended Community
and allocating an additional bit from the flags field of the
community to signal the proxy advertisement state.

```
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Type=0x06      | Sub-Type=0x08 |Reserved |P|O|R| Reserved=0   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                        Reserved=0                             |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The following bits in the flags field in the third octet of the
extended community are defined.  The remaining bits must be set to
zero when sending and must be ignored when receiving this community.

```
+--------+------------------------------------------------------------+
| Bit    | Meaning                                                    |
| Name   |                                                            |
+--------+------------------------------------------------------------+
| O,R    | Defined in IPv6 Neighbor Advertisement Flags in EVPN       |
|        | [I-D.ietf-bess-evpn-na-flags]                              |
| P      | Proxy IP->MAC advertisement defined in this draft          |
+--------+------------------------------------------------------------+
```

                    EVPN ARP/ND Extended Community Flags

5.  IANA Considerations

   This document requests the value of 0x04 of the EVPN ARP/ND extended
   community flags field to be assigned to the proxy IP->MAC
   advertisement (P) bit.

6.  Operational Considerations

   Depending on the number of multihoming PEs and MAC/IP scaling of an
   EVPN, proxy advertisement of IP->MAC entries by other PEs in addition
   to the devices initially learning IP->MAC entries locally in the data
   plane could cause scalability concerns for operators.  Proxy
   advertisements would increase the total number of EVPN routes
   maintained in the route tables of PEs, as well as increase the time
   required for PEs to download all remotely learned EVPN routes.
   Protocol implementations should provide administrative controls for
   operators to limit proxy advertisement functionality to situations
   where the benefits are required and the scale overhead is acceptable.

7.  Security Considerations

   This draft does not introduce any new security considerations to
   EVPN.

8.  Normative References

   [I-D.ietf-bess-evpn-na-flags]
             Rabadan, J., Sathappan, S., and K. Nagaraj, "Propagation
             of IPv6 Neighbor Advertisement Flags in EVPN", draft-ietf-
             bess-evpn-na-flags-02 (work in progress), October 2018.

   [I-D.ietf-bess-evpn-proxy-arp-nd]
             Rabadan, J., Sathappan, S., Nagaraj, K., Henderickx, W.,
             Hankins, G., King, T., Melzer, D., and E. Nordmark,
             "Operational Aspects of Proxy-ARP/ND in EVPN Networks",
             draft-ietf-bess-evpn-proxy-arp-nd-05 (work in progress),
             November 2018.

   [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>.

Authors' Addresses

   Ryan Bickhart
   Juniper Networks

   Email: rbickhart@juniper.net


   Wen Lin
   Juniper Networks

   Email: wlin@juniper.net


   John Drake
   Juniper Networks

   Email: jdrake@juniper.net

   Jorge Rabadan
   Nokia

   Email: jorge.rabadan@nokia.com