BESS Workgroup
Internet-Draft
Updates: 7432 (if approved)
Intended status: Standards Track
Expires: December 10, 2021

J. Rabadan, Ed.
S. Sathappan
K. Nagaraj
G. Hankins
Nokia
T. King
DE-CIX

June 8, 2021

Operational Aspects of Proxy-ARP/ND in Ethernet Virtual Private Networks

draft-ietf-bess-evpn-proxy-arp-nd-14

Abstract

This document describes the Ethernet Virtual Private Networks (EVPN) Proxy-ARP/ND function, augmented by the capability of the ARP/ND Extended Community. From that perspective this document updates the EVPN specification to provide more comprehensive documentation of the operation of the Proxy-ARP/ND function. The EVPN Proxy-ARP/ND function and the ARP/ND Extended Community help operators of Internet Exchange Points, Data Centers, and other networks deal with IPv4 and IPv6 address resolution issues associated with large Broadcast Domains by reducing and even suppressing the flooding produced by address resolution in the EVPN network.

Status of This Memo

Copyright Notice

Table of Contents

1. Terminology 2
2. Introduction 4
  2.1. The Data Center Use-Case 5
  2.2. The Internet Exchange Point Use-Case 5
3. Solution Description 6
  3.1. Proxy-ARP/ND Sub-Functions 8
  3.2. Learning Sub-Function 9
    3.2.1. Proxy-ND and the NA Flags 11
  3.3. Reply Sub-Function 12
  3.4. Unicast-forward Sub-Function 13
  3.5. Maintenance Sub-Function 14
  3.6. Flood (to Remote PEs) Handling 15
  3.7. Duplicate IP Detection 16
4. Solution Benefits 18
5. Deployment Scenarios 19
  5.1. All Dynamic Learning 19
  5.2. Dynamic Learning with Proxy-ARP/ND 20
  5.3. Hybrid Dynamic Learning and Static Provisioning with Proxy-ARP/ND 20
  5.4. All Static Provisioning with Proxy-ARP/ND 20
  5.5. Example of Deployment in Internet Exchange Points 21
  5.6. Example of Deployment in Data Centers 22
6. Security Considerations 22
7. IANA Considerations 23
8. Acknowledgments 23
9. Contributors 24
10. References 24
  10.1. Normative References 24
  10.2. Informative References 25
Authors' Addresses 25

1. 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

BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

ARP: Address Resolution Protocol.

AS-MAC: Anti-spoofing MAC. It is a especial MAC configured on all the PEs attached to the same BD and used for the Duplicate IP Detection procedures.

BD: Broadcast Domain.

BUM: Broadcast, Unknown unicast and Multicast layer-2 traffic.

CE: Customer Edge router.

DAD: Duplicate Address Detection, as per [RFC4861].

DC: Data Center.

EVI: EVPN Instance.

EVPN: Ethernet Virtual Private Networks, as per [RFC7432].

GARP: Gratuitous ARP message.

IP->MAC: an IP address associated to a MAC address. IP->MAC entries are programmed in Proxy-ARP/ND tables and may be of three different types: dynamic, static or EVPN-learned.

IXP: Internet eXchange Point.

IXP-LAN: the IXP's large Broadcast Domain to where Internet routers are connected.

LAG: Link Aggregation Group.

MAC or IP DA: MAC or IP Destination Address.

MAC or IP SA: MAC or IP Source Address.

ND: Neighbor Discovery Protocol.

NS: Neighbor Solicitation message.

NA: Neighbor Advertisement.

NUD: Neighbor Unreachability Detection, as per [RFC4861].

O Flag: Override Flag in NA messages, as per [RFC4861].

PE: Provider Edge router.

R Flag: Router Flag in NA messages, as per [RFC4861].

RT2: EVPN Route type 2 or EVPN MAC/IP Advertisement route, as per [RFC7432].

S Flag: Solicited Flag in NA messages, as per [RFC4861].

SN-multicast address: Solicited-Node IPv6 multicast address used by NS messages.

TLLA: Target Link Layer Address, as per [RFC4861].

VPLS: Virtual Private LAN Service.

This document assumes familiarity with the terminology used in [RFC7432].

2. Introduction

As specified in [RFC7432] the IP Address field in the Ethernet Virtual Private Networks (EVPN) MAC/IP Advertisement route may optionally carry one of the IP addresses associated with the MAC address. A PE may learn local IP->MAC pairs and advertise them in EVPN MAC/IP Advertisement routes. Remote PEs importing those routes in the same Broadcast Domain (BD) may add those IP->MAC pairs to their Proxy-ARP/ND tables and reply to local ARP requests or Neighbor Solicitations (or 'unicast-forward' those packets to the owner MAC), reducing and even suppressing in some cases the flooding in the EVPN network.

EVPN and its associated Proxy-ARP/ND function are extremely useful in DCs or Internet Exchange Points (IXPs) with large broadcast domains, where the amount of ARP/ND flooded traffic causes issues on connected routers and CEs. [RFC6820] describes the address resolution problems in large DC networks.

This document describes the Proxy-ARP/ND function in [RFC7432] networks, augmented by the capability of the ARP/ND Extended Community [I-D.ietf-bess-evpn-na-flags]. From that perspective this document updates [RFC7432].

Proxy-ARP/ND may be implemented to help IXPs, DCs and other operators deal with the issues derived from address resolution in large broadcast domains.

2.1. The Data Center Use-Case

As described in [RFC6820] the IPv4 and IPv6 address resolution can create a lot of issues in large DCs. In particular, the issues created by the IPv4 Address Resolution Protocol procedures may be significant.

On one hand, ARP Requests use broadcast MAC addresses, therefore any Tenant System in a large Broadcast Domain will see a large amount of ARP traffic, which is not addressed to most of the receivers.

On the other hand, the flooding issue becomes even worse if some Tenant Systems disappear from the broadcast domain, since some implementations will persistently retry sending ARP Requests. As [RFC6820] states, there are no clear requirements for retransmitting ARP Requests in the absence of replies, hence an implementation may choose to keep retrying endlessly even if there are no replies.

The amount of flooding that address resolution creates can be mitigated by the use of EVPN and its Proxy-ARP/ND function.

2.2. The Internet Exchange Point Use-Case

The implementation described in this document is especially useful in IXP networks.

A typical IXP provides access to a large layer-2 Broadcast Domain for peering purposes (referred to as 'the peering network'), where (hundreds of) Internet routers are connected. We refer to these Internet routers as Customer Edge (CE) devices in this section. Because of the requirement to connect all routers to a single layer-2 network the peering networks use IPv4 addresses in length ranges from /21 to /24 (and even bigger for IPv6), which can create very large broadcast domains. This peering network is transparent to the CEs, and therefore, floods any ARP request or NS messages to all the CEs in the network. Gratuitous ARP and NA messages are flooded to all the CEs too.

In these IXP networks, most of the CEs are typically peering routers and roughly all the BUM traffic is originated by the ARP and ND address resolution procedures. This ARP/ND BUM traffic causes significant data volumes that reach every single router in the peering network. Since the ARP/ND messages are processed in "slow path" software processors and they take high priority in the routers, heavy loads of ARP/ND traffic can cause some routers to run out of resources. CEs disappearing from the network may cause address resolution explosions that can make a router with limited processing power fail to keep BGP sessions running.

The issue might be better in IPv6 routers if MLD-snooping was enabled, since ND uses SN-multicast address in NS messages; however, ARP uses broadcast and has to be processed by all the routers in the network. Some routers may also be configured to broadcast periodic GARPs [RFC5227]. The amount of ARP/ND flooded traffic grows exponentially with the number of IXP participants, therefore the issue can only grow worse as new CEs are added.

In order to deal with this issue, IXPs have developed certain solutions over the past years. While these solutions may mitigate the issues of address resolution in large broadcasts domains, EVPN provides new more efficient possibilities to IXPs. EVPN and its Proxy-ARP/ND function may help solve the issue in a distributed and scalable way, fully integrated with the PE network.

3. Solution Description

Figure 1 illustrates an example EVPN network where the Proxy-ARP/ND function is enabled.

```
                                                         BD1
                                                     Proxy-ARP/ND
                                                    +------------+
   IP1/M1          +----------------------------+   |IP1->M1 EVPN|
    GARP --->Proxy-ARP/ND                       |   |IP2->M2 EVPN|
   +---+      +---------+   RT2(IP1/M1)         |   |IP3->M3 sta |
   |CE1+------|   BD1   |    ------>      +------+---|IP4->M4 dyn |
   +---+      +---------+                 |         +------------+
                  PE1                     | +--------+ Who has IP1?
                   |           EVPN       | |   BD1  | <-----  +---+
                   |           EVI1       | |        | ----->  |CE3|
   IP2/M2          |                      | |        | IP1->M1 +---+
    GARP  --->Proxy-ARP/ND                | +--------+   |    IP3/M3
     +---+    +---------+   RT2(IP2/M2)   |              |
     |CE2+----|   BD1   |    ------>      +--------------+
     +---+    +---------+                     PE3|    +---+
                  PE2                         | +----+CE4|
                   +----------------------------+      +---+
                                               <---IP4/M4 GARP
```

Figure 1: Proxy-ARP/ND network example

When the Proxy-ARP/ND function is enabled in a BD (Broadcast Domain) of the EVPN PEs, each PE creates a Proxy table specific to that BD that can contain three types of Proxy-ARP/ND entries:

a. Dynamic entries: learned by snooping CE's ARP and ND messages. For instance, IP4->M4 in Figure 1.

b. Static entries: provisioned on the PE by the management system. For instance, IP3->M3 in Figure 1.

c. EVPN-learned entries: learned from the IP/MAC information encoded in the received RT2's coming from remote PEs. For instance, IP1->M1 and IP2->M2 in Figure 1.

As a high level example, the operation of the EVPN Proxy-ARP/ND function in the network of Figure 1 is described below. In this example we assume IP1, IP2 and IP3 are IPv4 addresses:

1. Proxy-ARP/ND is enabled in BD1 of PE1, PE2 and PE3.

2. The PEs start adding dynamic, static and EVPN-learned entries to their Proxy tables:

   A. PE3 adds IP1->M1 and IP2->M2 based on the EVPN routes received from PE1 and PE2. Those entries were previously learned as dynamic entries in PE1 and PE2 respectively, and advertised in BGP EVPN.
   B. PE3 adds IP4->M4 as dynamic. This entry is learned by snooping the corresponding ARP messages sent by CE4.
   C. An operator also provisions the static entry IP3->M3.

3. When CE3 sends an ARP Request asking for the MAC address of IP1, PE3 will:

   A. Intercept the ARP Request and perform a Proxy-ARP lookup for IP1.
   B. If the lookup is successful (as in Figure 1), PE3 will send an ARP Reply with IP1->M1. The ARP Request will not be flooded to the EVPN network or any other local CEs.
   C. If the lookup is not successful, PE3 will flood the ARP Request in the EVPN network and the other local CEs.

In the same example, if we assume IP1, IP2, IP3 and IP4 are now IPv6 addresses and Proxy-ARP/ND is enabled in BD1:

1. PEs will start adding entries in a similar way as for IPv4, however there are some differences:

   A. IP1->M1 and IP2->M2 are learned as dynamic entries in PE1 and PE2 respectively, by snooping NA messages and not by snooping NS messages. In the IPv4 case, any ARP frame can be snooped to learn the dynamic Proxy-ARP entry. When learning the dynamic entries, the R and O Flags contained in the snooped NA messages will be added to the Proxy-ND entries too.

B. PE1 and PE2 will advertise those entries in EVPN MAC/IP Advertisement routes, including the corresponding learned R and O Flags in the ARP/ND Extended Community.

C. PE3 also adds IP4->M4 as dynamic, after snooping an NA message sent by CE4.

2. When CE3 sends an NS message asking for the MAC address of IP1, PE3 behaves as in the IPv4 example, by intercepting the NS, doing a lookup on the IP and replying with an NA if the lookup is successful. If it is successful the NS is not flooded to the EVPN PEs or any other local CEs.

3. If the lookup is not successful, PE3 will flood the NS to remote EVPN PEs attached to the same BD and the other local CEs as in the IPv4 case.

As PE3 learns more and more host entries in the Proxy-ARP/ND table, the flooding of ARP Request messages among PEs is reduced and in some cases it can even be suppressed. In a network where most of the participant CEs are not moving between PEs and they advertise their presence with GARPs or unsolicited-NA messages, the ARP/ND flooding among PEs, as well as the unknown unicast flooding, can practically be suppressed. In an EVPN-based IXP network, where all the entries are Static, the ARP/ND flooding among PEs is in fact totally suppressed.

In a network where CEs move between PEs, the Proxy-ARP/ND function relies on the CE signaling its new location via GARP or unsolicited-NA messages so that tables are immediately updated. If a CE moves "silently", that is, without issuing any GARP or NA message upon getting attached to the destination PE, the mechanisms described in Section 3.5 make sure that the Proxy-ARP/ND tables are eventually updated.

3.1. Proxy-ARP/ND Sub-Functions

The Proxy-ARP/ND function can be structured in six sub-functions or procedures:

1. Learning sub-function

2. Reply sub-function

3. Unicast-forward sub-function

4. Maintenance sub-function

5. Flood handling sub-function

6. Duplicate IP detection sub-function

A Proxy-ARP/ND implementation MUST at least support the Learning, Reply, Maintenance and Duplicate IP detection sub-functions. The following sections describe each individual sub-function.

3.2. Learning Sub-Function

A Proxy-ARP/ND implementation in an EVPN BD MUST support dynamic and EVPN-learned entries, and SHOULD support static entries.

Static entries are provisioned from the management plane. A static entry is configured on the PE attached to the host using the IP address in that entry. The provisioned static IP->MAC entry MUST be advertised in EVPN with an ARP/ND Extended Community where the Immutable ARP/ND Binding Flag (I) is set to 1, as per [I-D.ietf-bess-evpn-na-flags]. When the I flag in the ARP/ND Extended Community is 1, the advertising PE indicates that the IP address must not be associated to a MAC, other than the one included in the EVPN MAC/IP Advertisement route. The advertisement of I=1 in the ARP/ND Extended Community is compatible with any value of the Sticky bit (S) or Sequence Number in the [RFC7432] MAC Mobility Extended Community. Note that the I bit in the ARP/ND Extended Community refers to the immutable configured association between the IP and the MAC address in the IP->MAC binding, whereas the S bit in the MAC Mobility Extended Community refers to the fact that the advertised MAC address is not subject to the [RFC7432] mobility procedures.

An entry may associate a configured static IP to a list of potential MACs, i.e. IP1->(MAC1,MAC2..MACN). Until a frame (including local ARP/NA message) is received from the CE, the PE will not advertise any IP1->MAC in EVPN. Upon receiving traffic from the CE, the PE will check that the source MAC, E.g., MAC1, is included in the list of allowed MACs. Only in that case, the PE will activate the IP1->MAC1 and advertise only that IP1 and MAC1 in an EVPN MAC/IP Advertisement route.

The PE MUST create EVPN-learned entries from the received valid EVPN MAC/IP Advertisement routes containing a MAC and IP address.

Dynamic entries are learned in different ways depending on whether the entry contains an IPv4 or IPv6 address:

a. Proxy-ARP dynamic entries:

The PE MUST snoop all ARP packets (that is, all frames with Ethertype 0x0806) received from the CEs attached to the BD in order to learn dynamic entries. ARP packets received from remote EVPN PEs attached to the same BD are not snooped. The Learning function will add the Sender MAC and Sender IP of the snooped ARP packet to the Proxy-ARP table. Note that a MAC or an IP address with value 0 SHOULD NOT be learned.

b. Proxy-ND dynamic entries:

The PE MUST snoop the NA messages (Ethertype 0x86dd, ICMPv6 type 136) received from the CEs attached to the BD and learn dynamic entries from the Target Address and TLLA information. NA messages received from remote EVPN PEs are not snooped. A PE implementing Proxy-ND as in this document MUST NOT create dynamic IP->MAC entries from NS messages, since they don't contain the R Flag required by the Proxy-ND reply function. See Section 3.2.1 for more information about the R Flag.

This document specifies an "anycast" capability that can be configured for the proxy-ND function of the PE, and affects how dynamic Proxy-ND entries are learned based on the O Flag of the snooped NA messages. If the O Flag is zero in the received NA message, the IP->MAC SHOULD only be learned in case the IPv6 "anycast" capability is enabled in the BD. Irrespective, an NA message with O Flag = 0 will be normally forwarded by the PE based on a MAC DA lookup.

The following procedure associated to the Learning sub-function is RECOMMENDED:

- When a new Proxy-ARP/ND EVPN or static active entry is learned (or provisioned), the PE SHOULD send a GARP or unsolicited-NA message to all the connected access CEs. The PE SHOULD send a GARP or unsolicited-NA message for dynamic entries only if the ARP/NA message that previously created the entry on the PE was NOT flooded to all the local connected CEs before. This GARP/ unsolicited-NA message makes sure the CE ARP/ND caches are updated even if the ARP/NS/NA messages from CEs connected to remote PEs are not flooded in the EVPN network.

Note that if a Static entry is provisioned with the same IP as an existing EVPN-learned or Dynamic entry, the Static entry takes precedence.

In case of a PE reboot, the static and EVPN entries will be re-added as soon as the PE is back online and receives all the EVPN routes for the BD. However, the dynamic entries will be gone. Due to that

reason, new NS and ARP Requests will be flooded by the PE to remote PEs and dynamic entries gradually re-learned again.

3.2.1. Proxy-ND and the NA Flags

[RFC4861] describes the use of the R Flag in IPv6 address resolution:

- Nodes capable of routing IPv6 packets must reply to NS messages with NA messages where the R Flag is set (R Flag=1).

- Hosts that are not able to route IPv6 packets must indicate that inability by replying with NA messages that contain R Flag=0.

The use of the R Flag in NA messages has an impact on how hosts select their default gateways when sending packets off-link, as per [RFC4861]:

- Hosts build a Default Router List based on the received RAs and NAs with R Flag=1. Each cache entry has an IsRouter flag, which must be set for received RAs and is set based on the R flag in the received NAs. A host can choose one or more Default Routers when sending packets off-link.

- In those cases where the IsRouter flag changes from TRUE to FALSE as a result of a NA update, the node must remove that router from the Default Router List and update the Destination Cache entries for all destinations using that neighbor as a router, as specified in [RFC4861] section 7.3.3. This is needed to detect when a node that is used as a router stops forwarding packets due to being configured as a host.

The R Flag and O Flag for a Proxy-ARP/ND entry will be learned in the following ways:

- The R Flag information SHOULD be added to the Static entries by the management interface. The O Flag information MAY also be added by the management interface. If the R and O Flags are not configured, the default value is 1.

- Dynamic entries SHOULD learn the R Flag and MAY learn the O Flag from the snooped NA messages used to learn the IP->MAC itself.

- EVPN-learned entries SHOULD learn the R Flag and MAY learn the O Flag from the ARP/ND Extended Community [I-D.ietf-bess-evpn-na-flags] received from EVPN along with the RT2 used to learn the IP->MAC itself. If no ARP/ND Extended Community is received, the PE will add a configured R Flag/O Flag to the entry. These configured R and O Flags MAY be an

administrative choice with a default value of 1. The configuration of this administrative choice provides a backwards compatible option with EVPN PEs that follow [RFC7432] but do not support this specification.

Note that, typically, IP->MAC entries with O=0 will not be learned, and therefore the Proxy-ND function will reply to NS messages with NA messages that contain O=1. However, this document allows the configuration of the "anycast" capability in the BD where the Proxy-ND function is enabled. If "anycast" is enabled in the BD and an NA message with O=0 is received, the associated IP->MAC entry will be learned with O=0. If this "anycast" capability is enabled in the BD, Duplicate IP Detection must be disabled so that the PE is able to learn the same IP mapped to different MACs in the same Proxy-ND table. If the "anycast" capability is disabled, NA messages with O Flag = 0 will not create a Proxy-ND entry (although they will be forwarded normally), hence no EVPN advertisement with ARP/ND Extended Community will be generated.

3.3. Reply Sub-Function

This sub-function will reply to address resolution requests/ solicitations upon successful lookup in the Proxy-ARP/ND table for a given IP address. The following considerations should be taken into account, assuming that the ARP Request/NS lookup hits a Proxy-ARP/ND entry IP1->MAC1:

a. When replying to ARP Request or NS messages:

- the PE SHOULD use the Proxy-ARP/ND entry MAC address MAC1 as MAC SA. This is RECOMMENDED so that the resolved MAC can be learned in the MAC FIB of potential layer-2 switches sitting between the PE and the CE requesting the address resolution.

- for an ARP reply, the PE MUST use the Proxy-ARP entry IP1 and MAC1 addresses in the Sender Protocol Address and Hardware Address fields, respectively.

- for an NA message in response to an address resolution NS or DAD NS, the PE MUST use IP1 as the IP SA and Target Address. M1 MUST be used as the Target Link Local Address (TLLA).

b. A PE SHOULD NOT reply to a request/solicitation received on the same attachment circuit over which the IP->MAC is learned. In this case the requester and the requested IP are assumed to be connected to the same layer-2 CE/access network linked to the PE's attachment circuit, and therefore the requested IP owner will receive the request directly.

c. A PE SHOULD reply to broadcast/multicast address resolution messages, that is, ARP-Request, NS messages as well as DAD NS messages. A PE SHOULD NOT reply to unicast address resolution requests (for instance, NUD NS messages).

d. When replying to an NS, a PE SHOULD set the Flags in the NA messages as follows:

   - The R-bit is set as it was learned for the IP->MAC entry in the NA messages that created the entry (see Section 3.2.1).

   - The S Flag will be set/unset as per [RFC4861].

   - The O Flag will be set in all the NA messages issued by the PE, except in the case the BD is configured with the "anycast" capability and the entry was previously learned with O=0. If "anycast" is enabled and there are more than one MAC for a given IP in the Proxy-ND table, the PE will reply to NS messages with as many NA responses as 'anycast' entries there are in the Proxy-ND table.

e. A PE MUST NOT reply to ARP probes received from the CEs. An ARP probe is an ARP request constructed with an all-zero sender IP address that may be used by hosts for IPv4 Address Conflict Detection [RFC5227].

f. For Proxy-ARP, a PE MUST only reply to ARP-Request with the format specified in [RFC0826]. For Proxy-ND, a PE MUST reply to NS messages with the format and options specified in [RFC4861], and MAY reply to NS messages containing other options. Received NS messages with unknown options MAY be forwarded (as unicast packets) to the owner of the requested IP (assuming the MAC is known in the Proxy-ARP/ND table and BD). An administrative choice to control the behavior for received NS messages with unknown options ('unicast-forward', 'discard' or 'forward') MAY be supported. The 'forward' option implies flooding the NS message based on the MAC DA. The 'unicast-forward' option is described in Section 3.4. If 'discard' is available, the operator should assess if flooding NS unknown options may be a security risk for the EVPN BD (and is so, enable 'discard'), or if, on the contrary, not forwarding NS unknown options may disrupt connectivity.

3.4. Unicast-forward Sub-Function

As discussed in Section 3.3, in some cases the operator may want to 'unicast-forward' certain ARP-Request and NS messages as opposed to reply to them. The implementation of a 'unicast-forward' function is

RECOMMENDED. This option can be enabled with one of the following parameters:

a. unicast-forward always

b. unicast-forward unknown-options

If 'unicast-forward always' is enabled, the PE will perform a Proxy-ARP/ND table lookup and in case of a hit, the PE will forward the packet to the owner of the MAC found in the Proxy-ARP/ND table. This is irrespective of the options carried in the ARP/ND packet. This option provides total transparency in the BD and yet reduces the amount of flooding significantly.

If 'unicast-forward unknown-options' is enabled, upon a successful Proxy-ARP/ND lookup, the PE will perform a 'unicast-forward' action only if the ARP-Request or NS messages carry unknown options, as explained in Section 3.3. The 'unicast-forward unknown-options' configuration allows the support of new applications using ARP/ND in the BD while still reducing the flooding.

Irrespective of the enabled option, if there is no successful Proxy-ARP/ND lookup, the unknown ARP-Request/NS will be flooded in the context of the BD, as per Section 3.6.

3.5. Maintenance Sub-Function

The Proxy-ARP/ND tables SHOULD follow a number of maintenance procedures so that the dynamic IP->MAC entries are kept if the owner is active and flushed (and the associated RT2 withdrawn) if the owner is no longer in the network. The following procedures are RECOMMENDED:

a. Age-time

A dynamic Proxy-ARP/ND entry MUST be flushed out of the table if the IP->MAC has not been refreshed within a given age-time. The entry is refreshed if an ARP or NA message is received for the same IP->MAC entry. The age-time is an administrative option and its value should be carefully chosen depending on the specific use case: in IXP networks (where the CE routers are fairly static) the age-time may normally be longer than in DC networks (where mobility is required).

b. Send-refresh option

The PE MAY send periodic refresh messages (ARP/ND "probes") to the owners of the dynamic Proxy-ARP/ND entries, so that the

entries can be refreshed before they age out. The owner of the IP->MAC entry would reply to the ARP/ND probe and the corresponding entry age-time reset. The periodic send-refresh timer is an administrative option and is RECOMMENDED to be a third of the age-time or a half of the age-time in scaled networks.

An ARP refresh issued by the PE will be an ARP-Request message with the Sender's IP = 0 sent from the PE's MAC SA. If the PE has an IP address in the subnet, for instance on an Integrated Routing and Bridging (IRB) interface, then it MAY use it as a source for the ARP request (instead of Sender's IP = 0). An ND refresh will be a NS message issued from the PE's MAC SA and a Link Local Address associated to the PE's MAC.

The refresh request messages SHOULD be sent only for dynamic entries and not for static or EVPN-learned entries. Even though the refresh request messages are broadcast or multicast, the PE SHOULD only send the message to the attachment circuit associated to the MAC in the IP->MAC entry.

The age-time and send-refresh options are used in EVPN networks to avoid unnecessary EVPN RT2 withdrawals: if refresh messages are sent before the corresponding BD FIB and Proxy-ARP/ND age-time for a given entry expires, inactive but existing hosts will reply, refreshing the entry and therefore avoiding unnecessary EVPN MAC/IP Advertisement withdrawals in EVPN. Both entries (MAC in the BD and IP->MAC in Proxy-ARP/ND) are reset when the owner replies to the ARP/ND probe. If there is no response to the ARP/ND probe, the MAC and IP->MAC entries will be legitimately flushed and the RT2s withdrawn.

3.6. Flood (to Remote PEs) Handling

The Proxy-ARP/ND function implicitly helps reducing the flooding of ARP Request and NS messages to remote PEs in an EVPN network. However, in certain use cases, the flooding of ARP/NS/NA messages (and even the unknown unicast flooding) to remote PEs can be suppressed completely in an EVPN network.

For instance, in an IXP network, since all the participant CEs are well known and will not move to a different PE, the IP->MAC entries for the local CEs may be all provisioned on the PEs by a management system. Assuming the entries for the CEs are all provisioned on the local PE, a given Proxy-ARP/ND table will only contain static and EVPN-learned entries. In this case, the operator may choose to suppress the flooding of ARP/NS/NA from the local PE to the remote PEs completely.

The flooding may also be suppressed completely in IXP networks with dynamic Proxy-ARP/ND entries assuming that all the CEs are directly connected to the PEs and they all advertise their presence with a GARP/unsolicited-NA when they connect to the network. If any of those two assumptions is not true and any of the PEs may not learn all the local Proxy-ARP/ND entries, flooding of the ARP/NS/NA messages from the local PE to the remote PEs SHOULD NOT be suppressed, or the address resolution process for some CEs will not be completed.

In networks where fast mobility is expected (DC use case), it is NOT RECOMMENDED to suppress the flooding of unknown ARP-Requests/NS or GARPs/unsolicited-NAs. Unknown ARP-Requests/NS refer to those ARP-Request/NS messages for which the Proxy-ARP/ND lookups for the requested IPs do not succeed.

In order to give the operator the choice to suppress/allow the flooding to remote PEs, a PE MAY support administrative options to individually suppress/allow the flooding of:

- Unknown ARP-Request and NS messages.

- GARP and unsolicited-NA messages.

The operator will use these options based on the expected behavior on the CEs.

3.7. Duplicate IP Detection

The Proxy-ARP/ND function MUST support duplicate IP detection as per this section so that ARP/ND-spoofing attacks or duplicate IPs due to human errors can be detected. For IPv6 addresses, CEs will continue to carry out the DAD procedures as per [RFC4862]. The solution described in this section is an additional security mechanism carried out by the PEs that guarantees IPv6 address moves between PEs are legit and not the result of an attack. [RFC6957] describes a solution for IPv6 Duplicate Address Detection Proxy, however, it is defined for point-to-multipoint topologies with a split-horizon forwarding, where the 'CEs' have no direct communication within the same L2 link and therefore it is not suitable for EVPN Broadcast Domains. In addition, the solution described in this section includes the use of the AS-MAC for additional security.

ARP/ND spoofing is a technique whereby an attacker sends "fake" ARP/ND messages onto a broadcast domain. Generally the aim is to associate the attacker's MAC address with the IP address of another host causing any traffic meant for that IP address to be sent to the attacker instead.

The distributed nature of EVPN and Proxy-ARP/ND allows the easy detection of duplicated IPs in the network, in a similar way to the MAC duplication detection function supported by [RFC7432] for MAC addresses.

Duplicate IP detection monitors "IP-moves" in the Proxy-ARP/ND table in the following way:

a. When an existing active IP1->MAC1 entry is modified, a PE starts an M-second timer (default value of M=180), and if it detects N IP moves before the timer expires (default value of N=5), it concludes that a duplicate IP situation has occurred. An IP move is considered when, for instance, IP1->MAC1 is replaced by IP1->MAC2 in the Proxy-ARP/ND table. Static IP->MAC entries, that is, locally provisioned or EVPN-learned entries with I=1 in the ARP/ND Extended Community, are not subject to this procedure. Static entries MUST NOT be overridden by dynamic Proxy-ARP/ND entries.

b. In order to detect the duplicate IP faster, the PE SHOULD send a Confirm message to the former owner of the IP. A Confirm message is a unicast ARP-Request/NS message sent by the PE to the MAC addresses that previously owned the IP, when the MAC changes in the Proxy-ARP/ND table. The Confirm message uses a sender's IP 0.0.0.0 in case of ARP (if the PE has an IP address in the subnet then it MAY use it) and an IPv6 Link Local Address in case of NS. If the PE does not receive an answer within a given time, the new entry will be confirmed and activated. The default RECOMMENDED time to receive the confirmation is 30 seconds. In case of spoofing, for instance, if IP1->MAC1 moves to IP1->MAC2, the PE may send a unicast ARP-Request/NS message for IP1 with MAC DA= MAC1 and MAC SA= PE's MAC. This will force the legitimate owner to respond if the move to MAC2 was spoofed, and make the PE issue another Confirm message, this time to MAC DA= MAC2. If both, legitimate owner and spoofer keep replying to the Confirm message, the PE will detect the duplicate IP within the M-second timer:

   - If the IP1->MAC1 pair was previously owned by the spoofer and the new IP1->MAC2 was from a valid CE, then the issued Confirm message would trigger a response from the spoofer.

   - If it were the other way around, that is, IP1->MAC1 was previously owned by a valid CE, the Confirm message would trigger a response from the CE.

   Either way, if this process continues, then duplicate detection will kick in.

c. Upon detecting a duplicate IP situation:

1. The entry in duplicate detected state cannot be updated with new dynamic or EVPN-learned entries for the same IP. The operator MAY override the entry, though, with a static IP->MAC.

2. The PE SHOULD alert the operator and stop responding to ARP/NS for the duplicate IP until a corrective action is taken.

3. Optionally the PE MAY associate an "anti-spoofing-mac" (AS-MAC) to the duplicate IP in the Proxy-ARP/ND table. The PE will send a GARP/unsolicited-NA message with IP1->AS-MAC to the local CEs as well as an RT2 (with IP1->AS-MAC) to the remote PEs. This will update the ARP/ND caches on all the CEs in the BD, and hence all the CEs in the BD will use the AS-MAC as MAC DA when sending traffic to IP1. This procedure prevents the spoofer from attracting any traffic for IP1. Since the AS-MAC is a managed MAC address known by all the PEs in the BD, all the PEs MAY apply filters to drop and/or log any frame with MAC DA= AS-MAC. The advertisement of the AS-MAC as a "black-hole MAC" (by using an indication in the RT2) that can be used directly in the BD to drop frames is for further study.

d. The duplicate IP situation will be cleared when a corrective action is taken by the operator, or alternatively after a HOLD-DOWN timer (default value of 540 seconds).

The values of M, N and HOLD-DOWN timer SHOULD be a configurable administrative option to allow for the required flexibility in different scenarios.

For Proxy-ND, the Duplicate IP Detection described in this section SHOULD only monitor IP moves for IP->MACs learned from NA messages with O Flag=1. NA messages with O Flag=0 would not override the ND cache entries for an existing IP, and therefore the procedure in this section would not detect duplicate IPs. This Duplicate IP Detection for IPv6 SHOULD be disabled when the IPv6 "anycast" capability is activated in a given BD.

4. Solution Benefits

The solution described in this document provides the following benefits:

a. The solution may suppress completely the flooding of the ARP/ND messages in the EVPN network, assuming that all the CE IP->MAC

addresses local to the PEs are known or provisioned on the PEs from a management system. Note that in this case, the unknown unicast flooded traffic can also be suppressed, since all the expected unicast traffic will be destined to known MAC addresses in the PE BDs.

b. The solution reduces significantly the flooding of the ARP/ND messages in the EVPN network, assuming that some or all the CE IP->MAC addresses are learned on the data plane by snooping ARP/ND messages issued by the CEs.

c. The solution provides a way to refresh periodically the CE IP->MAC entries learned through the data plane, so that the IP->MAC entries are not withdrawn by EVPN when they age out unless the CE is not active anymore. This option helps reducing the EVPN control plane overhead in a network with active CEs that do not send packets frequently.

d. Provides a mechanism to detect duplicate IP addresses and avoid ARP/ND-spoof attacks or the effects of duplicate addresses due to human errors.

5. Deployment Scenarios

Four deployment scenarios with different levels of ARP/ND control are available to operators using this solution, depending on their requirements to manage ARP/ND: all dynamic learning, all dynamic learning with Proxy-ARP/ND, hybrid dynamic learning and static provisioning with Proxy-ARP/ND, and all static provisioning with Proxy-ARP/ND.

5.1. All Dynamic Learning

In this scenario for minimum security and mitigation, EVPN is deployed in the BD with the Proxy-ARP/ND function shutdown. PEs do not intercept ARP/ND requests and flood all requests issued by the CEs, as a conventional layer-2 network among those CEs would do. While no ARP/ND mitigation is used in this scenario, the operator can still take advantage of EVPN features such as control plane learning and all-active multihoming in the peering network.

Although this option does not require any of the procedures described in this document, it is added as baseline/default option for completeness. This option is equivalent to VPLS as far as ARP/ND is concerned. The options described in Section 5.2, Section 5.3 and Section 5.4 are only possible in EVPN networks in combination with their Proxy-ARP/ND capabilities.

5.2. Dynamic Learning with Proxy-ARP/ND

This scenario minimizes flooding while enabling dynamic learning of IP->MAC entries. The Proxy-ARP/ND function is enabled in the BDs of the EVPN PEs, so that the PEs snoop ARP/ND messages issued by the CEs and respond to CE ARP-requests/NS messages.

PEs will flood requests if the entry is not in their Proxy table. Any unknown source IP->MAC entries will be learned and advertised in EVPN, and traffic to unknown entries is discarded at the ingress PE.

This scenario makes use of the Learning, Reply and Maintenance sub-functions, with an optional use of the Unicast-forward and Duplicate IP detection sub-functions. The Flood handling sub-function uses default flooding for unknown ARP-Request/NS messages.

5.3. Hybrid Dynamic Learning and Static Provisioning with Proxy-ARP/ND

Some IXPs and other operators want to protect particular hosts on the BD while allowing dynamic learning of CE addresses. For example, an operator may want to configure static IP->MAC entries for management and infrastructure hosts that provide critical services. In this scenario, static entries are provisioned from the management plane for protected IP->MAC addresses, and dynamic learning with Proxy-ARP/ND is enabled as described in Section 5.2 on the BD.

This scenario makes use of the same sub-functions as in Section 5.2, but adding static entries added by the Learning sub-function.

5.4. All Static Provisioning with Proxy-ARP/ND

For a solution that maximizes security and eliminates flooding and unknown unicast in the peering network, all IP->MAC entries are provisioned from the management plane. The Proxy-ARP/ND function is enabled in the BDs of the EVPN PEs, so that the PEs intercept and respond to CE requests. Dynamic learning and ARP/ND snooping is disabled so that ARP-Requests and NS to unknown IPs are discarded at the ingress PE. This scenario provides an operator the most control over IP->MAC entries and allows an operator to manage all entries from a management system.

In this scenario, the Learning sub-function is limited to static entries, the Maintenance sub-function will not require any procedures due to the static entries, and the Flood handling sub-function will completely suppress Unknown ARP-Requests/NS messages as well as GARP and unsolicited-NA messages.

5.5. Example of Deployment in Internet Exchange Points

Nowadays, almost all IXPs install some security rules in order to protect the peering network (BD). These rules are often called port security. Port security summarizes different operational steps that limit the access to the IXP-LAN and the customer router, and controls the kind of traffic that the routers are allowed to exchange (e.g., Ethernet, IPv4, IPv6). Due to this, the deployment scenario as described in Section 5.4 "All Static Provisioning with Proxy-ARP/ND" is the predominant scenario for IXPs.

In addition to the "All Static Provisioning" behavior, in IXP networks it is recommended to configure the Reply Sub-Function to 'discard' ARP-Requests/NS messages with unrecognized options.

At IXPs, customers usually follow a certain operational life-cycle. For each step of the operational life-cycle specific operational procedures are executed.

The following describes the operational procedures that are needed to guarantee port security throughout the life-cycle of a customer with focus on EVPN features:

1. A new customer is connected the first time to the IXP:

   Before the connection between the customer router and the IXP-LAN is activated, the MAC of the router is allow-listed on the IXP's switch port. All other MAC addresses are blocked. Pre-defined IPv4 and IPv6 addresses of the IXP peering network space are configured at the customer router. The IP->MAC static entries (IPv4 and IPv6) are configured in the management system of the IXP for the customer's port in order to support Proxy-ARP/ND.

   In case a customer uses multiple ports aggregated to a single logical port (LAG) some vendors randomly select the MAC address of the LAG from the different MAC addresses assigned to the ports. In this case the static entry will be used associated to a list of allowed MACs.

2. Replacement of customer router:

   If a customer router is about to be replaced, the new MAC address(es) must be installed in the management system besides the MAC address(es) of the currently connected router. This allows the customer to replace the router without any active involvement of the IXP operator. For this, static entries are also used. After the replacement takes place, the MAC address(es) of the replaced router can be removed.

3. Decommissioning a customer router

   If a customer router is decommissioned, the router is disconnected from the IXP PE. Right after that, the MAC address(es) of the router and IP->MAC bindings can be removed from the management system.

5.6. Example of Deployment in Data Centers

DCs normally have different requirements than IXPs in terms of Proxy-ARP/ND. Some differences are listed below:

a. The required mobility in virtualized DCs makes the "Dynamic Learning" or "Hybrid Dynamic and Static Provisioning" models more appropriate than the "All Static Provisioning" model.

b. IPv6 'anycast' may be required in DCs, while it is typically not a requirement in IXP networks. Therefore if the DC needs IPv6 anycast addresses, the "anycast" capability will be explicitly enabled in the Proxy-ND function, hence the Proxy-ND sub-functions modified accordingly. For instance, if IPv6 'anycast' is enabled in the Proxy-ND function, the Duplicate IP Detection procedure in Section 3.7 must be disabled.

c. DCs may require special options on ARP/ND as opposed to the address resolution function, which is the only one typically required in IXPs. Based on that, the Reply Sub-function may be modified to forward or discard unknown options.

6. Security Considerations

The security considerations of [RFC7432] and [I-D.ietf-bess-evpn-na-flags] apply to this document too. Note that EVPN does not inherently provide cryptographic protection (including confidentiality protection).

The procedures in this document reduce the amount of ARP/ND message flooding, which in itself provides a protection to "slow path" software processors of routers and Tenant Systems in large BDs. The ARP/ND requests that are replied by the Proxy-ARP/ND function (hence not flooded) are normally targeted to existing hosts in the BD. ARP/ND requests targeted to absent hosts are still normally flooded; however, the suppression of Unknown ARP-Requests and NS messages described in Section 3.6 can provide an additional level of security against ARP-Requests/NS messages issued to non-existing hosts.

While the unicast-forward and/or flood suppression sub-functions provide an added security mechanism for the BD, they can also

increase the risk of blocking the service for a CE if the EVPN PEs cannot provide the ARP/ND resolution that the CE needs.

The solution also provides protection against Denial Of Service attacks that use ARP/ND-spoofing as a first step. The Duplicate IP Detection and the use of an AS-MAC as explained in Section 3.7 protects the BD against ARP/ND spoofing.

The Proxy-ARP/ND function specified in this document does not allow the learning of an IP address mapped to multiple MAC addresses in the same table, unless the "anycast" capability is enabled (and only in case of Proxy-ND). When "anycast" is enabled in the Proxy-ND function, the number of allowed entries for the same IP address MUST be limited by the operator to prevent DoS attacks that attempt to fill the Proxy-ND table with a significant number of entries for the same IP.

The document provides some examples and guidelines that can be used by IXPs in their EVPN BDs. When EVPN and its associated Proxy-ARP/ND function are used in IXP networks, they provide ARP/ND security and mitigation. IXPs must still employ additional security mechanisms that protect the peering network as per the established BCPs such as the ones described in [Euro-IX-BCP]. For example, IXPs should disable all unneeded control protocols, and block unwanted protocols from CEs so that only IPv4, ARP and IPv6 Ethertypes are permitted on the peering network. In addition, port security features and ACLs can provide an additional level of security.

Finally, it is worth noting that the Proxy-ARP/ND solution in this document will not work if there is a mechanism securing ARP/ND exchanges among CEs, because the PE is not able to secure the "proxied" ND messages.

7. IANA Considerations

No IANA considerations.

8. Acknowledgments

The authors want to thank Ranganathan Boovaraghavan, Sriram Venkateswaran, Manish Krishnan, Seshagiri Venugopal, Tony Przygienda, Robert Raszuk and Iftekhar Hussain for their review and contributions. Thank you to Oliver Knapp as well, for his detailed review.

9. Contributors

In addition to the authors listed on the front page, the following co-authors have also contributed to this document:

Wim Henderickx
Nokia

Daniel Melzer
DE-CIX Management GmbH

Erik Nordmark
Zededa

10. References

10.1. Normative References

[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>.

[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>.

[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>.

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

[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>.

[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>.

[I-D.ietf-bess-evpn-na-flags]
Rabadan, J., Sathappan, S., Nagaraj, K., and W. Lin, "Propagation of ARP/ND Flags in EVPN", draft-ietf-bess-evpn-na-flags-09 (work in progress), December 2020.

10.2. Informative References

[Euro-IX-BCP]
Euro-IX, "European Internet Exchange Association Best Practises - https://www.euro-ix.net/en/forixps/set-ixp/ixp-bcops/".

[RFC6820] Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", RFC 6820, DOI 10.17487/RFC6820, January 2013, <https://www.rfc-editor.org/info/rfc6820>.

[RFC6957] Costa, F., Combes, J-M., Ed., Pougnard, X., and H. Li, "Duplicate Address Detection Proxy", RFC 6957, DOI 10.17487/RFC6957, June 2013, <https://www.rfc-editor.org/info/rfc6957>.

[RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, DOI 10.17487/RFC4862, September 2007, <https://www.rfc-editor.org/info/rfc4862>.

Authors' Addresses

Jorge Rabadan (editor)
Nokia
777 Middlefield Road
Mountain View, CA 94043
USA

Email: jorge.rabadan@nokia.com

Senthil Sathappan
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA

Email: senthil.sathappan@nokia.com

Kiran Nagaraj
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA

Email: kiran.nagaraj@nokia.com

Greg Hankins
Nokia

Email: greg.hankins@nokia.com

Thomas King
DE-CIX Management GmbH

Email: thomas.king@de-cix.net