BESS Workgroup
Internet Draft

Intended status: Informational

J. Rabadan, Ed.
S. Sathappan
K. Nagaraj
G. Hankins
Nokia

T. King
DE-CIX

Expires: January 9, 2020

July 8, 2019

Operational Aspects of Proxy-ARP/ND in EVPN Networks
draft-ietf-bess-evpn-proxy-arp-nd-07

Abstract

The EVPN MAC/IP Advertisement route can optionally carry IPv4 and IPv6 addresses associated with a MAC address. Remote PEs can use this information to reply locally (act as proxy) to IPv4 ARP requests and IPv6 Neighbor Solicitation messages (or 'unicast-forward' them to the owner of the MAC) and reduce/suppress the flooding produced by the Address Resolution procedure. This EVPN capability is extremely useful in Internet Exchange Points (IXPs) and Data Centers (DCs) with large broadcast domains, where the amount of ARP/ND flooded traffic causes issues on routers and CEs. This document describes how the EVPN Proxy-ARP/ND function may be implemented to help IXPs and other operators deal with the issues derived from Address Resolution in large broadcast domains.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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." The list of current Internet-Drafts can be accessed at

http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on January 9, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Terminology 3
2. Introduction 4
2.1. The DC Use-Case 5
2.2. The IXP Use-Case 5
3. Solution Requirements 6
4. Solution Description 7
4.1. Learning Sub-Function 9
4.1.1. Proxy-ND and the NA Flags 10
4.2. Reply Sub-Function 11
4.3. Unicast-forward Sub-Function 12
4.4. Maintenance Sub-Function 13
4.5. Flooding (to Remote PEs) Reduction/Suppression 14
4.6. Duplicate IP Detection 15
5. Solution Benefits 17
6. Deployment Scenarios 17
6.1. All Dynamic Learning 17
6.2. Dynamic Learning with Proxy-ARP/ND 18
6.3. Hybrid Dynamic Learning and Static Provisioning with Proxy-ARP/ND 18
6.4 All Static Provisioning with Proxy-ARP/ND 18
6.5 Deployment Scenarios in IXPs 18
6.6 Deployment Scenarios in DCs 20
7. Security Considerations 20

8. IANA Considerations 21
9. References 21
9.1. Normative References 21
9.2. Informative References 22
10. Acknowledgments 22
11. Contributors 22
Authors' Addresses 22

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.

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

ARP: Address Resolution Protocol.

GARP: Gratuitous ARP message.

ND: Neighbor Discovery Protocol.

NS: Neighbor Solicitation message.

NA: Neighbor Advertisement.

IXP: Internet eXchange Point.

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

DC: Data Center.

IP->MAC: it refers to an IP address associated to a MAC address. The entries may be of three different types: dynamic, static or EVPN-learned.

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

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

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

SLLA: Source Link Layer Address, as per [RFC4861].

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

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

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

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

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

MAC or IP DA: MAC or IP Destination Address.

MAC or IP SA: MAC or IP Source Address.

AS-MAC: Anti-spoofing MAC.

LAG: Link Aggregation Group.

BD: Broadcast Domain.

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

2. Introduction

As specified in [RFC7432] the IP Address field in the 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 routes. The remote PEs 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 Data Centers (DCs) or Internet Exchange Points (IXPs) with large broadcast domains, where the amount of ARP/ND flooded traffic causes issues on routers and CEs. [RFC6820] describes the Address Resolution problems in Large Data Center networks.

This document describes how the [RFC7432] Proxy-ARP/ND function 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 DC 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 with the use of EVPN and its Proxy-ARP/ND function.

2.2. The IXP Use-Case

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

A typical IXP provides access to a large layer-2 peering network, where (hundreds of) Internet routers are connected. Because of the requirement to connect all routers to a single layer-2 network the peering networks use IPv4 layer-3 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 Customer Edge (CE) devices and therefore floods any ARP request or NS messages to all the CEs in the network. Unsolicited GARP 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 may be better in IPv6 routers, 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 go worse as new CEs are added.

In order to deal with this issue, IXPs have developed certain solutions over the past years. One example is the ARP-Sponge daemon [ARP-Sponge], which can reduce significantly the amount of ARP messages sent to an absent router. 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 Requirements

The distributed EVPN Proxy-ARP/ND function described in this document meets the following requirements:

- The solution supports the learning of the CE IP->MAC entries on the EVPN PEs via the management, control or data planes. An implementation should allow to intentionally enable or disable those possible learning mechanisms.

- 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.

- 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.

- 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.

- The solution provides a mechanism to detect duplicate IP addresses.

In case of duplication, the detecting PE should not reply to requests for the duplicate IP. Instead, the PE should alert the operator and may optionally prevent any other CE from sending traffic to the duplicate IP.

o The solution should not change any existing behavior in the CEs connected to the EVPN PEs.

4. Solution Description

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

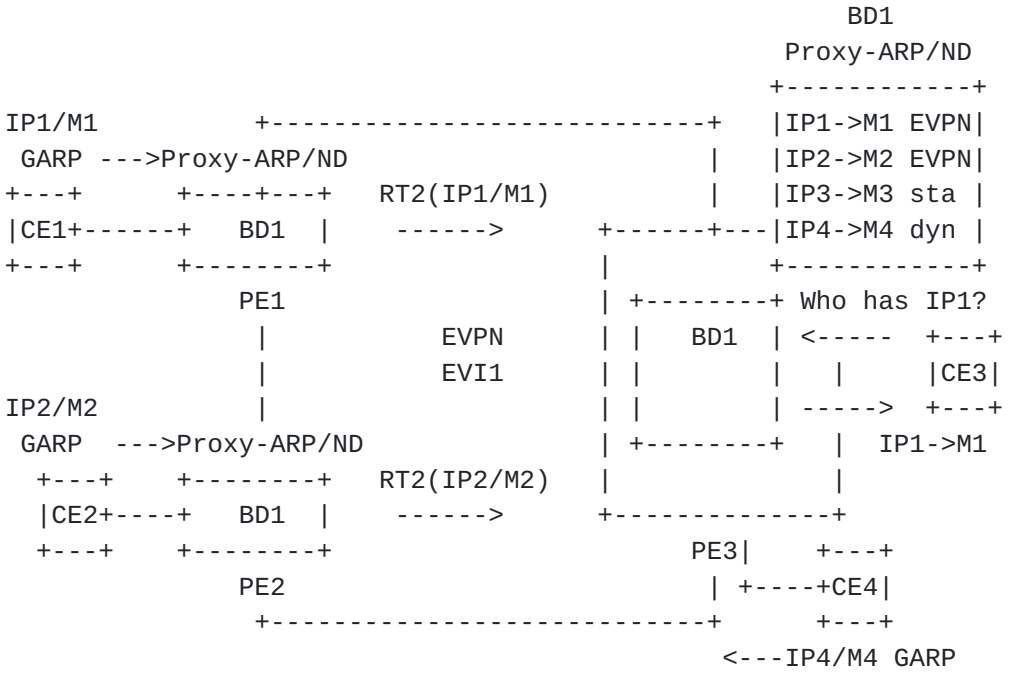


Figure 1 Proxy-ARP/ND network example

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.

As PE3 learns more and more host entries in the Proxy-ARP/ND table, the flooding of ARP Request messages 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 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 is in fact totally suppressed.

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. Flooding reduction/suppression sub-function
6. Duplicate IP detection sub-function

A Proxy-ARP/ND implementation MAY support all those sub-functions or only a subset of them. The following sections describe each individual sub-function.

4.1. Learning Sub-Function

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

Static entries are provisioned from the management plane. The provisioned static IP->MAC entry SHOULD be advertised in EVPN with an ARP/ND extended community where the Immutable ARP/ND Binding Flag flag (I) is set to 1, as per [EVPN-ARP-ND-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 MAC/IP 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). When there is more than one MAC in the list of allowed MACs, the PE will not advertise any IP->MAC in EVPN until a local ARP/NA message or any other frame is received from the CE. Upon receiving traffic from the CE, the PE will check that the source MAC is included in the list of allowed MACs. Only in that case, the PE will activate the IP->MAC and advertise it in EVPN.

EVPN-learned entries MUST be learned from 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:

They SHOULD be learned by snooping any ARP packet (Ethertype 0x0806) received from the CEs attached to the BD. The Learning function will add the Sender MAC and Sender IP of the snooped ARP packet to the Proxy-ARP table. Note that MAC and IPs with value 0 SHOULD NOT be learned.

b) Proxy-ND dynamic entries:

They SHOULD be learned out of the Target Address and TLLA information in NA messages (Ethertype 0x86DD, ICMPv6 type 136) received from the CEs attached to the BD. A Proxy-ND implementation SHOULD NOT learn IP->MAC entries from NS messages, since they don't contain the R-bit Flag required by the Proxy-ND reply function. See section 4.1.1 for more information about the R-bit flag.

Note that if the O-bit is zero in the received NA message, the IP->MAC SHOULD only be learned in case IPv6 'anycast' is enabled in the EVI.

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 an unsolicited GARP or NA message to the access CEs. The PE SHOULD send an unsolicited GARP/NA message for dynamic entries only if the ARP/NA message creating the entry was NOT flooded before. This unsolicited GARP/NA message makes sure the CE ARP/ND caches are updated even if the ARP/NS/NA messages from remote CEs 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.

4.1.1. Proxy-ND and the NA Flags

[RFC4861] describes the use of the R-bit flag in IPv6 Address Resolution:

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

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

The use of the R-bit flag in NA messages has an impact on how hosts select their default gateways when sending packets off-link:

- Hosts build a Default Router List based on the received RAs and NAs with R-bit=1. Each cache entry has an IsRouter flag, which must be set based on the R-bit 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-bit and O-bit will be learned in the following ways:

o Static entries SHOULD have the R-bit information added by the management interface. The O-bit information MAY also be added by the management interface.

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

o EVPN-learned entries SHOULD learn the R-bit and MAY learn the O-bit from the ND Extended Community received from EVPN along with the RT2 used to learn the IP->MAC itself. Please refer to [EVPN-ARP-ND-FLAGS]. If no ND extended community is received, the PE will add the default R-bit/O-bit to the entry. The default R-bit SHOULD be an administrative choice. The default O-bit SHOULD be 1.

Note that the O-bit SHOULD only be learned if 'anycast' is enabled in the EVI. If so, 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 'anycast' is disabled, NA messages with O-bit = 0 will not create a Proxy-ND entry, hence no EVPN advertisement with ND extended community will be generated.

4.2. 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:

a) When replying to ARP Request or NS messages, the PE SHOULD use the Proxy-ARP/ND entry MAC address 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.

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 switch/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) A PE SHOULD include the R-bit learned for the IP->MAC entry in the NA messages (see section 4.1.1). The S-bit will be set/unset as per [RFC4861]. The O-bit will be included if IPv6 'anycast' is enabled in the EVI and it is learned for the IP->MAC entry. If 'anycast' is enabled and there are more than one MAC for a given IP, the PE will reply to NS messages with as many NA responses as 'anycast' entries are in the Proxy-ND table.

e) A PE SHOULD 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) A PE SHOULD only reply to ARP-Request and NS messages with the format specified in [RFC0826] and [RFC4861] respectively. Received ARP-Requests and NS messages with unknown options SHOULD be either forwarded (as unicast packets) to the owner of the requested IP (assuming the MAC is known in the Proxy-ARP/ND table and BD) or discarded. An administrative option to control this behavior ('unicast-forward' or 'discard') SHOULD be supported. The 'unicast-forward' option is described in section 4.3.

4.3. Unicast-forward Sub-Function

As discussed in section 4.2. in some cases the operator may want to 'unicast-forward' certain ARP-Request and NS messages as opposed to reply to them. The operator SHOULD be able to activate this option 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 EVI 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 4.2. As an example, this would allow to enable Proxy-ND and Secure ND [RFC3971] in the same EVI. The 'unicast-forward unknown-options' configuration allows the support of new applications using ARP/ND in the EVI while still reducing the flooding at the same time.

4.4. 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 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 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 MAC and MAC-IP 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.

4.5. Flooding (to Remote PEs) Reduction/Suppression

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 may be all provisioned 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 to 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.

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 in the CEs.

4.6. Duplicate IP Detection

The Proxy-ARP/ND function SHOULD support duplicate IP detection so that ARP/ND-spoofing attacks or duplicate IPs due to human errors can be detected.

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 function supported by [RFC7432] for MAC addresses.

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

- 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.

- In order to detect the duplicate IP faster, the PE MAY 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 timer, the new entry will be confirmed and activated. 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 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 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.

o Upon detecting a duplicate IP situation:

a) 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.

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

c) Optionally the PE MAY associate an "anti-spoofing-mac" (AS-MAC) to the duplicate IP. 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 force all the CEs in the EVI to use the AS-MAC as MAC DA for IP1, and prevent the spoofer from attracting any traffic for IP1. Since the AS-MAC is a managed MAC address known by all the PEs in the EVI, 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" that can be used directly in the BD to drop frames is for further study.

o 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, Duplicate IP Detection SHOULD only monitor IP moves for IP->MACs learned from NA messages with O-bit=1. NA messages with O-bit=0 would not override the ND cache entries for an existing IP. Duplicate IP Detection for IPv6 SHOULD be disabled when IPv6 'anycast' is activated in a given EVI.

5. Solution Benefits

The solution described in this document provides the following benefits:

a) The solution may suppress completely the flooding of the ARP/ND and unknown-unicast messages in the EVPN network, in cases where all the CE IP->MAC addresses local to the PEs are known and provisioned on the PEs from a management system.

b) The solution reduces significantly the flooding of the ARP/ND messages in the EVPN network, in cases where 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 reduces the control plane overhead and unnecessary BGP MAC/IP Advertisements and Withdrawals in a network with active CEs that do not send packets frequently.

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

6. 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.

6.1. All Dynamic Learning

In this scenario for minimum security and mitigation, EVPN is deployed in the peering network with the Proxy-ARP/ND function shutdown. PEs do not intercept ARP/ND requests and flood all requests, as in a conventional layer-2 network. While no ARP/ND mitigation is used in this scenario, the IXP can still take advantage of EVPN features such as control plane learning and all-active multihoming in the peering network. Existing mitigation solutions, such as the ARP-Sponge daemon [ARP-Sponge] MAY also be used in this scenario.

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 6.2, 6.3 and 6.4 are only

possible in EVPN networks in combination with their Proxy-ARP/ND capabilities.

6.2. 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 intercept and respond to CE requests.

The solution MAY further reduce the flooding of the ARP/ND messages in the EVPN network by snooping ARP/ND messages issued by the CEs.

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

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

Some IXPs want to protect particular hosts on the peering network while allowing dynamic learning of peering router addresses. For example, an IXP may want to configure static MAC->IP entries for management and infrastructure hosts that provide critical services. In this scenario, static entries are provisioned from the management plane for protected MAC->IP addresses, and dynamic learning with Proxy-ARP/ND is enabled as described in section 6.2 on the peering network.

6.4 All Static Provisioning with Proxy-ARP/ND

For a solution that maximizes security and eliminates flooding and unknown unicast in the peering network, all MAC-IP 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 traffic to unknown entries is discarded at the ingress PE. This scenario provides an IXP the most control over MAC->IP entries and allows an IXP to manage all entries from a management system.

6.5 Deployment Scenarios in IXPs

Nowadays, almost all IXPs installed some security rules in order to protect the IXP-LAN. These rules are often called port security. Port security summarizes different operational steps that limit the access to the IXP-LAN, to the customer router and controls the kind of traffic that the routers are allowed to be exchange (e.g., Ethernet, IPv4, IPv6). Due to this, the deployment scenario as described in 6.4 "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 white-listed on the IXP's switch port. All other MAC addresses are blocked. Pre-defined IPv4 and IPv6 addresses of the IXP's 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.

6.6 Deployment Scenarios in DCs

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 not a requirement in IXP networks. Therefore if the DC needs IPv6 'anycast' it 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, Duplicate IP Detection 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.

7. Security Considerations

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 4.5. can provide an additional level of security against ARP-Requests/NS messages issued to non-existing hosts.

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 4.6. will definitely protect the BD against ARP/ND spoofing.

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 and SHOULD follow 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.

8. IANA Considerations

No IANA considerations.

9. References

9.1. 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., "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>.

[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>.

[RFC7342] Dunbar, L., Kumari, W., and I. Gashinsky, "Practices for Scaling ARP and Neighbor Discovery (ND) in Large Data Centers", RFC 7342, DOI 10.17487/RFC7342, August 2014, <https://www.rfc-editor.org/info/rfc7342>.

[RFC3971] Arkko, J., Ed., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, DOI 10.17487/RFC3971, March 2005, <https://www.rfc-editor.org/info/rfc3971>.

[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, <http://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <http://www.rfc-editor.org/info/rfc8174>.

9.2. Informative References

[ARP-Sponge] Wessel M. and Sijm N., Universiteit van Amsterdam, "Effects of IPv4 and IPv6 address resolution on AMS-IX and the ARP Sponge", July 2009.

[EVPN-ARP-ND-FLAGS] Sathappan S., Nagaraj K. and Rabadan J., "Propagation of ARP/ND Flags in EVPN", draft-ietf-bess-evpn-na-flags-04, Work in Progress, July 2019.

[Euro-IX BCP] https://www.euro-ix.net/pages/28/1/bcp_ixp.html

10. 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.

11. 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

Authors' Addresses

Jorge Rabadan (Editor)
Nokia
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@nokia.com

Senthil Sathappan
Nokia
Email: senthil.sathappan@nokia.com

Kiran Nagaraj
Nokia
Email: kiran.nagaraj@nokia.com

Greg Hankins
Nokia
Email: greg.hankins@nokia.com

Thomas King
DE-CIX Management GmbH
Lichtstrasse 43i, Cologne 50825, Germany
Email: thomas.king@de-cix.net