BESS Workgroup
Internet Draft

Intended status: Informational

J. Rabadan, Ed.
S. Sathappan
K. Nagaraj
W. Henderickx
Alcatel-Lucent

T. King
D. Melzer
DE-CIX

Expires: September 10, 2015
March 9, 2015

Proxy-ARP/ND function in EVPN networks
draft-snr-bess-evpn-proxy-arp-nd-00

Abstract

The MAC/IP Advertisement route specified in [RFC7432] 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 and reduce/suppress the flooding produced by the Address Resolution procedure. This EVPN capability is extremely useful in Internet Exchange Points (IXPs) with large broadcast domains, where the amount of ARP/ND flooded traffic causes issues on routers and CEs, as explained in [RFC6820]. This document describes how the [RFC7432] EVPN proxy-ARP/ND function should be implemented to help IXPs and other operators deal with the issues derived from Address Resolution in large broadcast domains.

Status of this Memo

Copyright Notice

Table of Contents

1. Terminology

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.

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

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.

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, 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 should be implemented to help IXPs and other operators deal with the issues derived from Address Resolution in large broadcast domains.

2.1. 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. 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 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]. 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 SHOULD meet the following requirements:

- The solution SHOULD support 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 traffic can also be suppressed, since all the expected unicast traffic will be destined to known MAC addresses in the PE MAC-VRFs.

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

- The solution MAY provide a way to refresh periodically the CE IP->MAC entries learnt 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 SHOULD provide 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.

4. Solution description

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

```
                                                         MAC-VRF1
                                                       Proxy-ARP/ND
                                                      +------------+
     IP1/M1          +----------------------------+   |IP1->M1 EVPN|
      GARP --->Proxy-ARP/ND                       |   |IP2->M2 EVPN|
     +---+      +----+---+   RT2(IP1/M1)          |   |IP3->M3 sta |
     |CE1+------+MAC-VRF1|    ------>      +------+---|IP4->M4 dyn |
     +---+      +--------+                 |          +------------+
                    PE1                    | +--------+ Who has IP1?
                     |           EVPN      | |MAC-VRF1| <-----  +---+
                     |           EVI1      | |        |   |     |CE3|
     IP2/M2          |                     | |        | ----->  +---+
      GARP  --->Proxy-ARP/ND               | +--------+   |  IP1->M1
       +---+    +--------+   RT2(IP2/M2)   |              |
       |CE2+----+MAC-VRF1|    ------>      +--------------+
       +---+    +--------+                       PE3|    +---+
                    PE2                            | +----+CE4|
                     +----------------------------+      +---+
                                                   <---IP4/M4 GARP
```

Figure 1 Proxy-ARP/ND network example

When the Proxy-ARP/ND function is enabled in the MAC-VRFs of the EVPN PEs, each PE creates a Proxy table specific to that MAC-VRF that can contain three types of Proxy-ARP/ND entries:

a) Dynamic entries: learnt 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-learnt entries: learnt 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 MAC-VRF1 of PE1, PE2 and PE3.

2. The PEs start adding dynamic, static and EVPN-learnt 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 learnt as dynamic entries in PE1 and PE2 respectively, and advertised in

BGP EVPN.

b. PE3 adds IP4->M4 as dynamic. This entry is learnt 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 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 five sub-functions or procedures:

1. Learning sub-function
2. Reply sub-function
3. Maintenance sub-function
4. Flooding reduction/suppression sub-function
5. 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-learnt entries.

Static entries are provisioned from the management plane. The provisioned static IP->MAC entry SHOULD be advertised in EVPN with a MAC Mobility extended community where the static flag is set to 1, as per [RFC7432].

EVPN-learnt entries MUST be learnt from received valid EVPN MAC/IP Advertisement routes containing a MAC and IP address.

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

a) Proxy-ARP dynamic entries:

They SHOULD be learnt by snooping any ARP packet (Ethertype 0x0806) received from the CEs attached to the MAC-VRF. The Learning function will add the Sender MAC and Sender IP of the snooped ARP packet to the Proxy-ARP table.

b) Proxy-ND dynamic entries:

They SHOULD be learnt out of the Target Address and TLLA information in NA messages (Ethertype 0x86DD, ICMPv6 type 136) received from the CEs attached to the MAC-VRF. 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 NOT be learnt.

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

o When a new Proxy-ARP/ND EVPN or static active entry is learnt (or provisioned) the PE SHOULD send an unsolicited GARP or NA message to the access CEs. This 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-learnt 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:

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

o 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:

- o 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.

- o 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 will be learnt in the following ways:

- o Static entries SHOULD have the R-bit information added by the management interface.

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

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

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 learnt in the MAC FIB of potential Layer-2 switches seating 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 learnt. 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 learnt for the IP->MAC entry in the NA messages (see section 4.1.1). The S-bit and O-bit will be set/unset as per [RFC4861].

4.3. 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 SHOULD 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.

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. 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-learnt 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 MAC-VRF 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 MAC-VRF 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.

4.4. 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-learnt 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, it is not recommended to suppress the flooding of unknown ARP-Requests/NS or GARPs/unsolicited-NAs.

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:

o Unknown ARP-Request and NS messages (unknown means that the lookups for the requested IPs do not succeed). o GARP and unsolicited-NA messages.

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

4.5. 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:

o When a new active IP1->MAC1 entry is learnt, 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.

o In order to detect the duplicate IP faster, the PE MAY send a CONFIRM message to the former owner of the IP. If the PE does not receive an answer within a given timer, the new entry will be confirmed and activated. 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 and the spoofer to reply so that the PE can detect the duplicate IP within the M timer.

o Upon detecting a duplicate IP situation:

a) The entry in duplicate detected state cannot be updated with new dynamic or EVPN-learnt 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 as well) 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/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 MAC-VRF to drop frames is for further study.

- 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 learnt from NA messages with O-bit=1. NA messages with O-bit=0 would not override the ND cache entries for an existing IP.

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 learnt 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. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

7. Security Considerations

They will be added in future revisions.

8. IANA Considerations

9. References

9.1. Normative References

[RFC7432]Sajassi A., Aggarwal R. et al, "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015, <http://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, September 2007, <http://www.rfc-editor.org/info/rfc4861>.

[RFC6820]Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", RFC 6820, January 2013, <http://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, August 2014, <http://www.rfc-editor.org/info/rfc7342>.

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-ND-FLAGS] Sathappan S., Nagaraj K. and Rabadan J., "Propagation of IPv6 Neighbor Advertisement Flags in EVPN", draft-snr-bess-evpn-na-flags-00, Work in Progress, March 2015.

10. Acknowledgments

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

Authors' Addresses

Jorge Rabadan (Editor)
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@alcatel-lucent.com

Senthil Sathappan
Alcatel-Lucent
Email: senthil.sathappan@alcatel-lucent.com

Kiran Nagaraj
Alcatel-Lucent
Email: kiran.nagaraj@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.com

Thomas King
DE-CIX
Email: thomas.king@de-cix.net

Daniel Melzer
DE-CIX
Email: daniel.melzer@de-cix.net