BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: 28 December 2025

A. Sajassi
L. Krattiger
K. Ananthamurthy
Cisco
J. Rabadan
Nokia
W. Lin
Juniper Networks, Inc.

26 June 2025

EVPN First Hop Security
draft-sajassi-bess-evpn-first-hop-security-04

Abstract

The Dynamic Host Configuration Protocol (DHCP) snoop database stores valid IPv4-to-MAC and IPv6-to-MAC bindings by snooping on DHCP messages. These bindings are used by security functions like Dynamic Address Resolution Protocol Inspection (DAI), Neighbor Discovery Inspection (NDI), IPv4 Source Guard, and IPv6 Source Guard to safeguard against traffic received with a spoofed address. These functions are collectively referred to as First Hop Security (FHS). This document proposes BGP extensions and new procedures for Ethernet VPN (EVPN) will distribute and synchronize the DHCP snoop database to support FHS. Such synchronization is needed to support EVPN host mobility and multi-homing.

Requirements Language

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

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

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

Copyright Notice

Table of Contents

1. Introduction

DHCP snoop database stores valid IPv4-to-MAC and IPv6-to-MAC bindings by snooping on Dynamic Host Configuration Protocol (DHCP) messages. These bindings are used by security functions like Dynamic ARP Inspection (DAI), Neighbor Discovery Inspection (NDI), IPv4 Source Guard, and IPv6 Source Guard to safeguard against traffic received with a spoofed address. These functions are collectively referred to as First Hop Security (FHS).

FHS may be leveraged by Ethernet VPN (EVPN) [RFC7432] PEs operating in bridge mode or in IRB mode (with distributed anycast default gateway functionality [RFC9135]) in Data Center (DC), Enterprise, and/or Service Provider (SP) networks to enhance the security of such networks. This document proposes BGP extensions and new procedures for EVPN to support FHS in the presence of EVPN multi-homing and host mobility by distributing DHCP snoop bindings among EVPN PEs participating in that EVPN Broadcast Domain. These bindings not only need to be distributed among multi-homing PEs to ensure the synchronization of these PEs are for DHCP messages but also need to be distributed among the PEs participating in that EVPN Broadcast Domain to provide a host mobility procedures can operate adequately. I.e., when a host moves from the current EVPN peer to a new EVPN peer, then the new EVPN peer shall have the bindings so that it can continue to do FHS without any interruption.

DAI and NDI use the DHCP snoop database to validate received ARP messages and ND messages, respectively. Likewise, IPv4 Source Guard and IPv6 Source Guard uses this database to validate source IPv4 and IPv6 addresses, respectively, before forwarding traffic. While FHS running on top of DHCP snoop database are widely deployed on access switches (without standard-based multi-homing or host mobility), there is a need to extend the application of FHS on EVPN PEs supporting Network Virtualization Overlay (NVO) and running multi-homing (All-Active or Single-Active) with host mobility.

Unfortunately, the lack of DHCP snoop binding on EVPN PEs would lead to failure of FHS (i.e., IP Source Guard, DAI, and NDI) when a host is multi-homed to multiple PEs (e.g., All-Active or Single-Active) and/or when a host moves from one PE to another PE. This is because when the host is All-Active multi-homed among multiple PEs, DHCP

messages can arrive on different multi-homing PEs without a single PE (in the multi-homing/redundancy group) seeing DHCP exchanges needed to build DHCP snoop database as described in Section 5. Since there is a possibility of none of the PEs in the redundancy group see the complete DHCP message exchanges needed to build DHCP snoop database, then none of the PEs in the group can establish the DHCP snoop binding, which in turn, causes failure of FHS. Furthermore, when a host moves from an old PE to a new PE, the new PE does not have the DHCP binding for that host. Since the new PE would not have the DHCP snoop binding, both IP Source Guard and DAI/NDI would start dropping packets originating from that host, resulting in FHS failure, which in turn results in service failure.

[RFC7513] proposes procedures that enable adding source address validation on a device based on DHCP exchanges. Their approach differs from that of ours in two ways. First, when the host moves from one PE to another PE, [RFC7513] Section 7.1 offers a probabilistic solution. Our approach provides a deterministic solution by proactively sending DHCP snoop updates from one PE to another so that the new PE would have the information it needs before the host moves to it. Second, [RFC7513] Section 5 identifies the need to distribute the DHCP snoop bindings but does not provide a procedure for distribution. Our approach offers an extension to EVPN protocol to distribute the DHCP snoop bindings.

2. Requirements Language

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Terminology

All-Active Redundancy Mode: When all PEs attached to an Ethernet segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

ASBR: Autonomous System Boundary Router.

Backup-DF (BDF): Backup-Designated Forwarder.

BD: Broadcast Domain.

DC: Data Center.

DF: Designated Forwarder. A DF is a PE device that is selected from among a group of PE devices that participate in EVPN multi-homing. It is the role of DF PE to forward Broadcast, Unicast, and Multicast (BUM) Layer 2 messages to the host that is multi-homed to all the PEs. DF PE is selected on a per-EVI basis.

DHCP: Dynamic Host Configuration Protocol.

DHCP Client: A DHCP client is a host that gets an address assignment from a DHCP server.

DHCP Server: A server that assigns network addresses to its clients.

DHCP Snoop Anchor: A PE device that originates a DHCP Snoop Route. It is this device that uses the DHCP Snoop bindings to do source address validation for hosts that sit behind it.

DHCP Snoop Route (DSR): EVPN Route to sync DHCP Snoop binding.

DORA : Discover, Offer, Request, Acknowledge.

EPOCH: The epoch is 1st January 1970 at 00:00 UTC.

Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an Ethernet Segment Identifier'.

Ethernet Tag: Used to represent a BD that is configured on a given ES for the purposes of DF election and <EVI, BD> identification for frames received from the CE. Note that any of the following may be used to represent a BD: VIDs (including Q-in-Q tags), configured IDs, VNIs (Virtual Extensible Local Area Network (VXLAN) Network Identifiers), normalized VIDs, I-SIDs (Service Instance Identifiers), etc., as long as the representation of the BDs is configured consistently across the multihomed PEs attached to that ES.

Ethernet Tag ID: Normalized network wide ID that is used to identify a BD within an EVI and carried in EVPN routes.

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN. An EVI may be comprised of one BD (VLAN-based, VLAN Bundle, or Port-based services) or multiple BDs (VLAN-aware Bundle or Port-based VLAN-Aware services).

IRB: Integrated Routing and Bridging interface, with EVPN procedures described in [RFC9135]

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.

Non-DF (NDF): Non-Designated Forwarder.

NVO: Network Virtualization Overlay as described in [RFC8365]

PE: Provider Edge device.

Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

SP: Service Provider.

UTC: Coordinated Universal Time.

VID: VLAN Identifier.

4. DHCP Snoop Primer

DHCP basic operation understanding is paramount to understnd the DHCP snooping operation on a non-distributed switch where no synchronization is needed. DHCP snooping is based on snooping of DHCP handshake between the host and the DHCP server. The handshake sequence has four steps, sometimes known as the DORA exchange (Figure 1) which is described in [RFC2131].

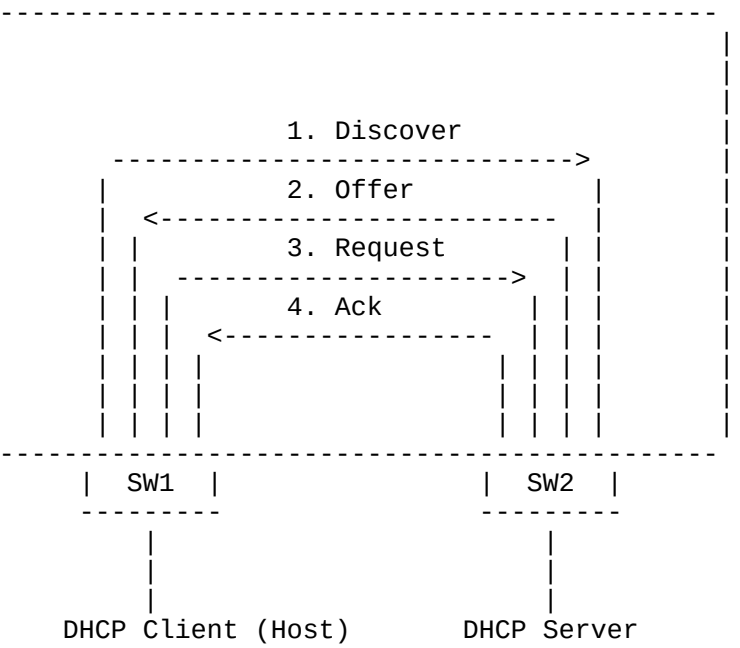

Figure 1: Typical DHCP DORA Exchange

1. Discover (DHCPDISCOVER): Initial DHCP message sent by the host (or the DHCP client) to discover DHCP server(s) in the network.

2. Offer (DHCPOFFER): Once a DHCP server receives the Discover message, it responds with an offer of an IP address that can be assigned to the host. There can be multiple DHCP servers in the network and hence multiple servers can respond to the Discover message by sending their own Offer message.

3. Request (DHCPREQUEST): Once the host receives one or more of the above offers, it sends a request to one of the DHCP servers confirming that it has accepted its offer.

4. Acknowledge (DHCPACK): The DHCP server sends the last DHCP message for which the Request message was sent. The message is sent to indicate the completion of the IP assignment mechanism.

4.1. DHCP Snoop binding entry

DHCP snoop binding is created using DHCPREQUEST and DHCPACK messages. Section 2 of [RFC2131] defines the DHCP message fields and the following are some of the key fields to understand the exchange of DHCPREQUEST and DHCPACK messages.

* 'ciaddr': Client IP address; only filled in if client is in BOUND, RENEW or REBINDING state and can respond to ARP requests.

* 'giaddr': Relay agent IP address, used in booting via a relay agent.

* 'yiaddr': 'your' (client) IP address.

DHCP client-server interaction is defined in section 3 of [RFC2131], which are

1. Client-server interaction - allocating a network address.

2. Client-server interaction - reusing a previously allocated network address

When a host is connected to a single switch (e.g., SW1), both DHCPREQUEST and DHCPACK messages pass through the same switch. Thus, the switch (SW1 in this case) can build and validate its state for DHCP snoop for that host. If SW1 relies on just a single DHCP message (such as DHCPACK that contains all the needed info) instead of both DHCPREQUEST and DHCPACK to build its DHCP snoop state, then it exposes itself to security risks and hijacking MAC/IP binding when a rouge DHCPACK is received.

5. Synchronizing DHCP Snoop Database

Considering the distributed nature of EVPN application in providing distributed bridge and distributed host gateway functions over a DC, Enterprise, and/or SP network, the synchronization challenges of providing FHS over such a distributed system needs to be addressed. The two main challenges are the synchronization of the DHCP snoop database (used in FHS) for both EVPN multi-homing and EVPN host mobility.

The synchronization procedure needed in EVPN to address these two challenges are dependent on the type of EVPN service being provided - i.e., bridge service vs. Integrated Routing and Bridging (IRB) service. Therefore, we organize the synchronization procedures needed based on the EVPN services in the following subsections.

EVPN single-homing is analogous to the scenario described in Section 4, where a host is connected to a single switch. If it wasn't for EVPN host mobility, then the existing DHCP snoop procedures could be leveraged as is. However, additional extensions are needed for EVPN host mobility and EVPN multi-homing will be described in the following subsections.

The solution described here addresses both the multi-homing and the host mobility issues of FHS by distributing DHCP snoop bindings among the EVPN peers. A new EVPN route is proposed DHCP Snoop Route (DSR) to carry the DHCP snoop binding information and detailed in Section 9.

5.1. DHCP Snoop Anchor PE

The PE where the host is attached sees completion of DHCPREQUEST and DHCPACK exchange between a DHCP Client (host) and a DHCP server, we refer to this PE as the DHCP Snoop Anchor PE.

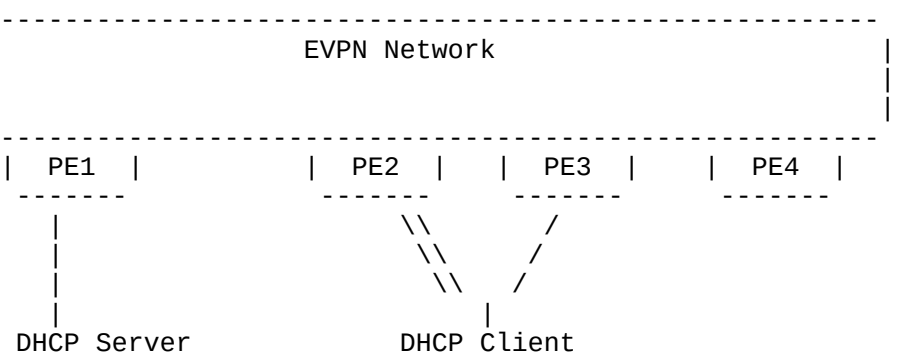

Figure 2: Single-Homed and Multi-Homed hosts.

DHCP Snoop Anchor PE (e.g., PE2) originates the DSR. When a remote BGP peer receives the DSR (e.g., PE4), it imports locally and updates its DHCP Snoop Database. With this information, if the host moved to a new PE (e.g., PE4), the new PE would already have the DSR update from the old PE. As a result, the DHCP Snoop procedure running on the new PE would successfully validate the host and immediately start accepting that host's messages.

* For initial IP address assignment, both DHCPREQUEST and DHCPACK needs to be received by the same multi-homing PE in order for that PE to become DHCP Snoop Anchor PE and to originate DSR

* For subsequent IP address renewal, ONLY DHCPACK needs to be received by one of the multi-homing PEs in order for that PE to become DHCP Snoop Anchor PE and to originate DSR

5.2. DHCP Message Synchronization

The synchronization procedure for DHCP snoop bindings avoid synchronization of DHCPREQUEST and DHCPACK message among the PEs and instead for the most part relies on a single PE to receive DHCPREQUEST and DHCPACK message exchanges for initial IP address assignment and ONLY DHCPACK for subsequent IP address renewal. After the completion of such exchange, it will distribute the DHCP snoop binding to the PEs participating in that EVPN Broadcast Domain.

The following sections describe the DHCP snoop procedures and associated synchronization needed for EVPN All-Active multi-homing and host mobility for DHCP initial IP address allocation/lease and IP address renewal when EVPN PEs participate in a bridged and IRB service.

5.3. Bridged Service

When EVPN bridged service is used with DHCP snooping, it is assumed that both DHCP clients and servers reside in the same subnet (same bridge domain and EVI). If DHCP servers reside in a subnet different then one of the DHCP clients, then EVPN IRB service along with DHCP relay function needs to be deployed, which will be described in Section 5.4.

Just as in the use-case of FHS application in traditional switches, we assume that the PE interfaces on which DHCP information is exchanged with the DHCP server is secure and the DHCP server itself is not compromised.

5.3.1. DHCP IP Address Allocation and Lease for Bridged Service

In this section, we describe how an anchor PE for DHCP snoop is selected among PEs participating in an EVPN multi-homing for a given BD. Furthermore, we explain why we don’t need synchronization for individual DHCPREQUEST and DHCPACK messages among these multi-homing PEs for anchor PE selection, but rather we need to synchronize the final DHCP snoop state among the PEs participating in that EVI after verification of DHCPREQUEST and DHCPACK exchange and the anchor PE selection. The synchronization of the final DHCP snoop state is achieved when the anchor PE distributes this information is via DSR.

When a DHCP client is multi-homed to two or more PEs on the same Ethernet Segment operating in All-Active mode, DORA messages can arrive at different PEs. However, only one PE in the multi-homing redundancy group receives both DHCPREQUEST and DHCPACK messages and thus designates itself as DHCP Snoop Anchor PE. The behavior in the case of Single-Active multi-homing applies to other multi-homing modes, such as port-active [I-D.ietf-bess-evpn-mh-pa] or single-flow active [I-D.ietf-bess-evpn-l2gw-proto] multi-homing, DHCPREQUEST and DHCPACK messages can only arrive at a single PE in the redundancy group, which is the active PE for that ESI/EVI, hence and thus the anchor PE for DHCP snoop.

1. A DHCP client initiates a DORA exchange by sending a DHCPDISCOVER broadcast message. Because of All-Active multi-homing, this broadcast message arrives at only one PE in the redundancy group (e.g., PE2), which forwards it to all the other participating PEs for that BD, including PE1, PE3 and PE4.

2. Each DHCP server for that subnet replies with a DHCPOFFER, while DHCPOFFER may be broadcast or unicast in the following cases.

   * Broadcast: If DHCPDISCOVER has ’ciaddr’ and ’giaddr’ set to ZERO with Broadcast bit option.

   * Unicast: If DHCPDISCOVER has ’ciaddr’ and ’giaddr’ set to ZERO without the Broadcast bit option, then the client’s hardware address and ’yiaddr’ address are used.

   Since client MAC is not learned in the EVPN network before the client obtains the IP address, even if DHCPOFFER is unicast, it will be sent as an unknown unicast (from PE1’s perspective). Effectively, PE (e.g., PE1) attached to the DHCP server sends this broadcast/unknown unicast message to all other PEs in that BD/EVI and thus all the multi-homing PEs for that DHCP client (e.g., PE2 and PE3) receive the DHCPOFFER broadcast/unknown unicast message and the DF PE (e.g., PE3) forwards the message to the DHCP client.

3. The DHCP client responds with a DHCPREQUEST message of type broadcast and gets hashed to PE2 again. PE2 will create incomplete DHCP snoop binding entry and forwards this broadcast message to all other PEs in that BD, including PE1. PE1 delivers this broadcast message to the DHCP server.

4. DHCP server responds with DHCPACK. Since ’ciaddr’ and ’giaddr’ are ZERO during initial setup, the client does not yet have the ’yiaddr’ address. DHCPACK will be sent as broadcast PE (e.g., PE1) attached to the DHCP server sends this broadcast message to

all other PEs in that BD/EVI and thus all the multi-homing PEs for that DHCP client (e.g., PE2 and PE3) receive the DHCPACK broadcast message and the DF PE (e.g., PE3) forwards the message to the DHCP client. PE2 received the DHCPREQUEST earlier on its local attachment circuit, and with DHCPACK, it creates the complete DHCP snoop binding, claims the Anchor, and originates the DSR.

As the above example illustrates, only one PE in the redundancy group (e.g., PE2) receives DHCPREQUEST on its local attachment circuit and DHCPACK messages. After verification of this exchange, it creates a DHCP snoop state and designates itself as the DHCP anchor for that client. Next, the anchor PE sends an EVPN DSR with the snooped MAC/IP binding, lease time, and other pertinent information to all PEs in that BD, including multi-homing PEs in the same redundancy group.

When multi-homing PEs in the same redundancy group receive this DSR message from the anchor PE, they register the DHCP snoop state for that host sitting behind that ESI. Therefore, from this time forward, when ARP/ND message (or data traffic) is received from that host, the host MAC address is learned and advertised in EVPN MAC/IP RT-2 is in the EVPN network and the traffic is forwarded accordingly.

5.3.2. DHCP IP Address Renewal for Bridged Service

A DHCP client will send DHCPREQUEST to renew the lease, which can be unicast or broadcast. Client will set ’yiaddr’ address as it already knows the address. If DHCPREQUEST is a broadcast message then the procedure defined in Section 5.3.1 will apply. If DHCPREQUEST is a unicast, because of All-Active multi-homing, DHCPREQUEST unicast message arrives at only one of the PEs in the redundancy group (e.g., PE2), which forwards it to DHCP server.

DHCP server responds with DHCPACK. Since Client had set ’yiaddr’ address in DHCPREQUEST, DHCPACK will be a unicast and either PE2 or PE3 will receive the DHCPACK.

* If PE2 receives the DHCPACK which is the anchor PE then, lease time will be updated and DSR update will be sent with the new lease time. All other PEs including the multihomed PEs will receive and update the lease time in the snoop entry that they have created with the previous DSR update.

* If PE3 receives the DHCPACK which is not the anchor PE and determines that it has received the snoop entry from the multihomed PE (e.g., PE2), which is the anchor then it claims itself as an anchor and advertises DSR updates with a MAC Mobility extended community attribute with a sequence number one greater

than the sequence number in the MAC Mobility extended community attribute of the received DSR. Suppose the snoop entry does not have the MAC Mobility extended community attribute; the value of the sequence number in the received DSR is assumed to be 0 for the purpose of this processing.

* PE2, which is the previous anchor, receives DSR with a higher sequence number from its ESI peer PE3, determines that ESI peer has claimed the anchor and withdraws the previously advertised DSR. Note that when MAC/IP routes are received from the same ESI, no mobility event is triggered irrespective of the sequence number. But for MAC/IP routes, the ES peer will not withdraw its own MAC/IP route, so the case for the DSR is different indeed.

5.4. IRB Service

When EVPN IRB service is used with DHCP snooping, if both DHCP clients and servers reside in the same subnet (same bridge domain and EVI), then the procedure defined in Section 5.3 will apply. If DHCP servers reside in a subnet different than one of the DHCP clients, then EVPN IRB service and DHCP relay function MUST be deployed. The solution described here addresses the multi-homing and host mobility issues by distributing DHCP snoop bindings among the EVPN peers.

5.4.1. DHCP IP Address Allocation and Lease for IRB Service

A DHCP client initiates a DORA exchange by sending a DHCPDISCOVER broadcast message. Because of All-Active multi-homing, this broadcast message arrives at only one PE in the redundancy group (e.g., PE2), which forwards it to the DHCP server defined in the relay config. The source IP address used in the relay message will be a unique IP configured on multihomed PEs such that the DHCP server response comes to the PE, which initiates the DHCP relay message.

There could be multiple DHCP relays configured with different servers. Each DHCP server can reply with a DHCPOFFER broadcast message and will be unicasted to the PE, which originated the DHCPDISCOVER relay message, which broadcasts on its local interfaces.

The DHCP client responds with a DHCPREQUEST message of type broadcast and gets hashed to PE2 again. PE2 will create an incomplete DHCP snoop binding entry and forward this broadcast message via the DHCP relay.

DHCP server responds with a DHCPACK message, which will be unicasted to the PE (e.g., PE2), which originated the DHCPREQUEST relay message. PE will broadcast this message on its local interfaces.

As the above example illustrates, only one PE in the redundancy group (e.g., PE2) receives DHCPREQUEST on its local attachment circuit and DHCPACK messages. After verification of this exchange, it creates a DHCP snoop state and designates itself as the DHCP anchor for that client. Next, the anchor PE sends an EVPN DSR with the snooped MAC/IP binding, lease time, and other pertinent information to all PEs in that BD, including multi-homing PEs in the same redundancy group.

When multi-homing PEs in the same redundancy group receive this DSR message from the anchor PE, they register the DHCP snoop state for that host sitting behind that ESI. Therefore, from this time forward, when an ARP/ND message (or data traffic) is received from that host, the host MAC address is learned and advertised in EVPN MAC/IP RT-2 in the EVPN network, and the traffic is forwarded accordingly.

5.4.2. DHCP IP Address Renewal for IRB Service

A DHCP client will send a DHCPREQUEST to renew the lease. Because of All-Active multi-homing, the DHCPREQUEST unicast message arrives at only one of the PEs in the redundancy group, which forwards it to the DHCP server defined in the relay config.

Suppose DHCPREQUEST arrives on PE2, which forwards it to the DHCP server defined in the relay config. If PE2 is the anchor PE, then after receiving the DHCPACK, the DHCP snoop entry lease time will be updated, and a DSR update will be sent with the new lease time. All other PEs, including the multihomed PEs, will receive and update the lease time in the snoop entry created with the previous DSR update.

Suppose DHCPREQUEST arrives on PE3, which forwards it to the DHCP server defined in the relay config. If PE3 is not the anchor PE, then after receiving the DHCPACK, it determines that it has received the snoop entry from the multihomed PE, which is the anchor (e.g., PE2), then it claims itself as an anchor.

* DHCP snoop entry lease time will be updated.

* DSR update will be sent with the new lease time with a MAC Mobility extended community attribute with a sequence number one greater than the sequence number in the MAC Mobility extended community attribute of the received DSR. Suppose the snoop entry does not have the MAC Mobility extended community attribute; the value of the sequence number in the received DSR is assumed to be 0 for the purpose of this processing.

* PE2, the previous anchor, receives DSR with a higher sequence number from its multihomed PE3, determines that multihomed PE3 has claimed the anchor, and withdraws the previously advertised DSR. Note that no mobility event is triggered when MAC/IP routes are received from the same ESI, irrespective of the sequence number. But for MAC/IP routes, the ES peer will not withdraw its own MAC/IP route, so the case for the DSR route is different indeed.

* All other PEs will receive and update the lease time in the snoop entry that they have created with the previous DSR update.

5.5. DSR handling on non-ESI PEs

When other PEs (e.g., PE4) in the same BD receive this DSR message advertised by the anchor PE, they also register and synchronize the DHCP snoop state for that host with that of the anchor PE.

Contrary to EVPN MAC/IP Advertisement Routes (RT-2), EVPN DSR (RT-x) does not need to use EVPN Ethernet AD per ES Route (RT-1) for route resolution as described in section 9.2.2 of [I-D.ietf-bess-rfc7432bis] because DSR is only used for DHCP snoop state and not traffic forwarding. It is better to maintain the last state of DHCP snoop for a given MAC/IP binding than to have no state at all. Furthermore, there is no impact on traffic forwarding in the case of DSR, whereas if route resolution based on RT-1 is not performed for RT-2, traffic destined to that MAC can be blackholed till it is learned again at the remote PEs.

6. DHCP Snoop Anchor Mobility

The host move will be detected via the data plane or GARP/RARP when the host moves from Anchor PE to remote PE. Since DHCP snoop entry was synced via the DSR from Anchor on remote PE, the EVPN mobility procedure will be initiated as defined in [RFC7432]. After completion of the mobility procedure, the anchor will be moved to the remote PE, where the host is moved. A duplicate-wait-timer with a default value of 30 sec will be started to identify the duplicate case. After the duplicate-wait-timer expires, the anchor will be moved if MAC/IP in the DSR is learned locally. If not, then Anchor will not be moved. Subsequent Host mobility will again start the duplicate-wait-timer.

If Anchor is moved from a remote location to a local one, the MAC Mobility extended community attribute defined [RFC7432] will be used for the DSR. Every Anchor mobility event for a given DSR will contain a sequence number that is set using the following rules:

1. A PE advertising given DSR for the first time advertises it with no MAC Mobility extended community attribute.

2. A PE detecting a locally attached DSR for which it had previously received a DSR with a different Ethernet segment identifier advertises the DSR tagged with a MAC Mobility extended community attribute with a sequence number one greater than the sequence number in the MAC Mobility extended community attribute of the received DSR. In the case of the first mobility event for a given DSR, where the received DSR does not carry a MAC Mobility extended community attribute, the value of the sequence number in the received DSR is assumed to be 0 for the purpose of this processing.

3. A PE detecting a locally attached DSR for which it had previously received a DSR with the same non-zero Ethernet segment identifier advertises it with the following:

   * No MAC Mobility extended community attribute if the received DSR did not carry said attribute.

   * a MAC Mobility extended community attribute with the sequence number equal to the highest of the sequence number(s) in the received DHCP Snoop Route (s) if the received route(s) is (are) tagged with a MAC Mobility extended community attribute.

4. A PE detecting a locally attached DSR for which it had previously received a DSR with the same zero Ethernet segment identifier (single-homed scenarios) advertises it with a MAC Mobility extended community attribute with the sequence number appropriately set. In the case of single-homed scenarios, there is no need for an ESI comparison. ESI comparison is made for multi-homing to prevent false detection of DSR moves among the PEs attached to the same multihomed site.

A PE receiving a DSR for a MAC/IP address with a different Ethernet segment identifier and a higher sequence number than that which it had previously advertised withdraws its DSR. If two (or more) PEs advertise the same DSR with the same sequence number but different Ethernet segment identifiers, a PE that receives these routes selects the route advertised by the PE with the lowest IP address, which is the best route. If the PE is the originator of the DSR and it receives the same DSR with the same sequence number that it generated, it will compare its IP address with the IP address of the remote PE and will select the lowest IP. If its route is not the best one, it will withdraw the route.

Previous Anchor PE receiving DSR from remote check whether the MAC/IP is learned remotely; if so, it will withdraw the local DSR and use the remote DSR. If MAC/IP is learned locally, then it will increment the sequence number by ONE, then the received sequence number.

7. Host Mobility and Age-Out

When using the DSR, the baseline host mobility procedures in EVPN are not affected. When the host moves from one PE to another and both PEs have the same BD, the new PE would already have the remote DHCP Snoop Entry. As a result, it would accept the incoming ARP/ND messages. Once it learns the new host, the new PE can send a new MAC/IP update.

When the host ages out, the PE would withdraw the EVPN MAC/IP advertisement route without bothering about the DSR. If the DHCP Lease expiration timer is running on the PE, then the PE does not send a withdrawal of the DSR. Once the Lease expires, the PE can withdraw the DSR as well.

8. Race Conditions

8.1. Inter-ES Mobility

A race-condition can happen when the host moves from one PE device (say PE1) to another PE device (say PE2). Let us say that as soon as DHCPREQUEST is validated on PE1 and PE1 advertises the DSR to other PE devices. The host moves from PE1 to PE2. Upon moving, the host generates a GARP (Gratuitous ARP) message. The GARP message MAY arrive sooner on PE2 than the DSR. In other words, PE2 receives the GARP before it has populated its DHCP binding and thus discards GARP.

We can address the above race-condition by storing an ARP entry associated with the GARP message and a flag indicating that we should keep the entry for T seconds. If DSR arrives within T, then the flag is removed and ARP entry is made permanent. Otherwise, we delete the ARP entry after the expiration of T seconds. In other words, the ARP entry is created, but it stays inactive until the DSR arrives and activates the ARP entry.

8.2. Intra-ES Synchronization

A similar race-condition can occur when multiple PEs are connected to the same Ethernet-Segment. Let us say, that upon successfully getting the DHCP handshake done, the host generates an ARP message. The ARP message MAY reach PE2, which is different from PE1, which has the Snoop DB binding. However, they are in the same Ethernet Segment. In other words, PE2 receives the GARP before it has

populated its DHCP binding and thus discards the ARP.

Once again, we can address the above race-condition by storing an ARP entry associated with the ARP message and a flag indicating that it will be kept for T seconds. If DSR arrives within T, the flag is removed and ARP entry is made permanent. Otherwise, the ARP entry is deleted after the expiration of T seconds.

9. BGP EVPN DSR Route

The BGP EVPN NLRI as defined in [RFC7432] is shown below:

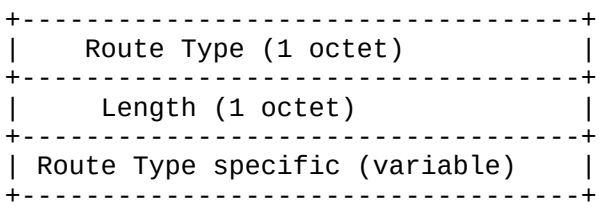

Figure 3: BGP EVPN NLRI

We propose a new EVPN route type called DHCP Snoop Route with the following format:

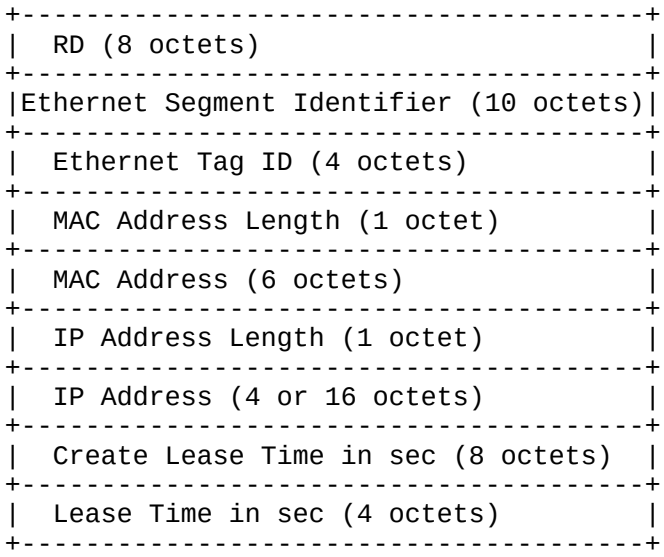

Figure 4: EVPN DSR Route

* The Route Distinguisher (RD) and Ethernet Tag ID MUST be used as defined in [RFC7432] and [RFC8365]. In particular, the RD is unique per MAC-VRF.

* Ethernet Segment Identifier (ESI) is a unique non-zero identifier that identifies an Ethernet segment. The ESI format is described in [RFC7432].

* The MAC Address and the IP Address fields are the MAC address and IP address of the host respectively. The MAC Address length (in bits) field specifies the host’s MAC address length. The IP address Length (in bits) field specifies the host’s IP address length.

* Create-Time is the value when DHCP entry is created, it also gets updated when DHCP Lease renewal happens. The value is calculated from EPOCH time -1st January 1970 UTC I,e how many seconds elapsed from EPOCH.

* Lease-Time is the value of lease time remaining for the DHCP snoop entry in seconds.

* For the purpose of BGP route key processing, only the Ethernet Tag ID, MAC Address Length, MAC Address, IP Address Length, and IP Address fields are considered to be part of the prefix in the NLRI.

* The BGP advertisement for the DSR MUST also carry the Route Target (RT) associated with the BD.

9.1. Create and Lease Time Handling

Anchor PE originates the DSR when the DORA exchange is complete. DHCP Snoop DB entry will maintain the create time and lease time. When DHCP lease renewal is complete, the create time and lease time are updated. The Create time will be in seconds. For example, the Create time on January 1, 2022 12:00:01 A.M will be represented in seconds a 1640995201.

All EVPN peers will be expected to synchronize the timestamp using NTP such that Create time will be interpreted correctly.

The PE router will calculate the lease time as follows.

Lease Time = Received Lease time - (Current time - Create time)

There are no lease time calculations in transit BGP EVPN peers like route reflectors of ASBRs.

10. Security Considerations

Security considerations discussed in [RFC7432] and [RFC8365] apply to this document as well.

11. IANA Considerations

This document defines a new EVPN route type called DHCP Snoop Route and request the following registration in the EVPN Route Type registry:

Value : 12
Description: DHCP Snoop Route

12. References

12.1. Normative References

[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>.

[RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, DOI 10.17487/RFC2131, March 1997, <https://www.rfc-editor.org/info/rfc2131>.

[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>.

[RFC7513] Bi, J., Wu, J., Yao, G., and F. Baker, "Source Address Validation Improvement (SAVI) Solution for DHCP", RFC 7513, DOI 10.17487/RFC7513, May 2015, <https://www.rfc-editor.org/info/rfc7513>.

[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>.

12.2. Informative References

[I-D.ietf-bess-evpn-l2gw-proto]
Burdet, L. A., Brissette, P., Sajassi, A., Maheshwari, P., and I. Bhatt, "EVPN Multi-Homing Mechanism for Layer-2 Gateway Protocols", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-l2gw-proto-05, 3 March 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-l2gw-proto-05>.

[I-D.ietf-bess-evpn-mh-pa]
Brissette, P., Burdet, L. A., Wen, B., Leyton, E., and J. Rabadan, "EVPN Port-Active Redundancy Mode", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-mh-pa-13, 5 December 2024, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-mh-pa-13>.

[I-D.ietf-bess-rfc7432bis]
Sajassi, A., Burdet, L. A., Drake, J., and J. Rabadan, "BGP MPLS-Based Ethernet VPN", Work in Progress, Internet-Draft, draft-ietf-bess-rfc7432bis-13, 24 June 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-rfc7432bis-13>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <https://www.rfc-editor.org/info/rfc9135>.

Contributors

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

Samir Thoria

Authors' Addresses

Ali Sajassi
Cisco
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: sajassi@cisco.com

Lukas Krattiger
Cisco
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: lkrattig@cisco.com

Krishnaswamy Ananthamurthy
Cisco
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: kriswamy@cisco.com

Jorge Rabadan
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: jorge.rabadan@nokia.com

Wen Lin
Juniper Networks, Inc.
10 Technology Park Drive
Westford, Massachusetts 01886
United States of America
Email: wlin@juniper.net