INTERNET-DRAFT N. Malhotra, Ed.
K. Patel
Arrcus
Intended Status: Proposed Standard
J. Rabadan
Nokia

Expires: Sept 12, 2019 Mar 11, 2019

LSoE-based PE-CE Control Plane for EVPN
draft-malhotra-bess-evpn-lsoe-00

Abstract

In an EVPN network, EVPN PEs provide VPN bridging and routing service to connected CE devices based on BGP EVPN control plane. At present, there is no PE-CE control plane defined for an EVPN PE to learn CE MAC, IP, and any other routes from a CE that may be distributed in EVPN control plane to enable unicast flows between CE devices. As a result, EVPN PEs rely on data plane based gleaning of source MACs for CE MAC learning, ARP/ND snooping for CE IPv4/IPv6 learning, and in some cases, local configuration for learning prefix routes behind a CE. A PE-CE control plane alternative to this traditional learning approach, where applicable, offers certain distinct advantages that in turn result in simplified EVPN operation.

This document defines a PE-CE control plane as an optional alternative to traditional non-control-plane based PE-CE learning in an EVPN network. It defines PE-CE control plane procedures and TLVs based on LSoE as the base protocol, enumerates advantages that may be achieved by using this PE-CE control plane, and discusses in detail EVPN use cases that are simplified as a result.

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

In an EVPN network, CE devices typically connect to an EVPN PE via layer-2 interfaces that terminate in a BD on the PE. Multi-homed LAG interfaces together with EVPN all-active multi-homing procedures are used to achieve PE-CE link and PE node redundancy for fault-tolerance and load-balancing. PEs provide overlay bridging and, optionally, first-hop routing service for these CE devices based on an EVPN control plane that is used to distribute CE MAC, IP, and prefix reachability across PEs.

At present, there is no PE-CE control plane defined for an EVPN PE to learn connected CE host MACs and IPs. As a result, EVPN PEs rely on:

- data plane based gleaning of source MAC for MAC learning,
- ARP snooping for IPv4 + MAC learning, and
- ND snooping for IPv6 + MAC learning.

A PE-CE control plane alternative to this traditional learning approach, where applicable, can offer some distinct advantages across various boot-up, mobility, and convergence scenarios:

- PE-CE learning is decoupled from non-deterministic hashing of data, ARP, and ND packets from CEs over all-active multi-homed LAG interfaces.
- PE-CE learning is decoupled from non-deterministic periodicity of data traffic from CEs or, in an extreme scenario, from CE device being silent for an extended period.
- PE-CE learning is decoupled from non-deterministic CE behavior with respect to unsolicited ARPs and NAs following boot-up and moves.
- PE-CE learning is decoupled from latencies associated with data packet triggered ARP and ND gleaning.

This in-turn results in simplification of certain EVPN operations such as aliasing, MAC and IP syncing across multi-homing PEs, and probing on MAC/IP moves. In addition, it helps achieve a deterministic convergence behavior across various boot-up, mobility, and failure scenarios.

A PE may also use local policy configuration for learning prefixes behind a CE that does not run a dynamic routing protocol. A PE-CE control plane can provide an operationally simpler alternative to local configuration for such use cases, where CE and PE devices are not under the same configuration management entity.

This document defines a new PE-CE control plane as an alternative to traditional data-plane and ARP/ND snooping based PE-CE host learning and to local configuration-based PE-CE prefix learning. It defines PE-CE control plane procedures and TLVs based on [LSOE] as the base protocol, enumerates advantages that may be achieved by using this PE-CE control plane, and discusses in detail EVPN operations that are simplified as a result. Use of PE-CE control plane defined in this document is intended to be optional and backwards compatible with CEs that use traditional PE-CE learning within the same BD.

1.1 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.

The following terms are used in this document:

- LSoE: Link State over Ethernet Protocol defined in [LSOE]
- EVPN-IRB: A BGP-EVPN distributed control plane based integrated routing and bridging fabric overlay discussed in [EVPN-IRB]
- Underlay: IP or MPLS fabric core network that provides IP or MPLS routed reachability between EVPN PEs.
- Overlay: VPN or service layer network consisting of EVPN PEs OR VPN provider-edge (PE) switch-router devices that runs on top of an underlay routed core.
- EVPN PE: A PE switch-router in a data-center fabric that runs overlay BGP-EVPN control plane and connects to overlay CE host devices. An EVPN PE may also be the first-hop layer-3 gateway for CE/host devices. This document refers to EVPN PE as a logical function in a data-center fabric. This EVPN PE function may be physically hosted on a top-of-rack switching device (ToR) OR at layer(s) above the ToR in the Clos fabric. An EVPN PE is typically also an IP or MPLS tunnel end-point for overlay VPN flows.
- CE: A tenant host device that has layer 2 connectivity to an EVPN PE switch-router, either directly OR via intermediate switching device(s).
- Symmetric EVPN-IRB: An overlay fabric first-hop routing architecture as defined in [EVPN-IRB], wherein, overlay host-to-host routed inter-subnet flows are routed at both ingress and egress EVPN PEs.
- Asymmetric EVPN-IRB: An overlay fabric first-hop routing architecture as defined in [EVPN-IRB], wherein, overlay host-to-host routed inter-subnet flows are routed and bridged at ingress PE and bridged at egress PEs.
- Centralized EVPN-IRB: An overlay fabric first-hop routing architecture, wherein, overlay host-to-host routed inter-subnet

flows are routed at a centralized gateway, typically at the one of the spine layers, and where EVPN PEs are pure bridging devices.

- ARP: Address Resolution Protocol [RFC 826].
- ND: IPv6 Neighbor Discovery Protocol [RFC 4861].
- Ethernet-Segment: physical Ethernet or LAG port that connects an access device to an EVPN PE, as defined in [RFC 7432].
- ESI: Ethernet Segment Identifier as defined in [RFC 7432].
- LAG: Layer-2 link-aggregation, also known as layer-2 bundle port-channel, or bond interface.
- EVPN all-active multi-homing: PE-CE all-active multi-homing achieved via a multi-homed layer-2 LAG interface on a CE with member links to multiple PEs and related EVPN procedures on the PEs.
- EVPN Aliasing: multi-homing procedure as defined in [RFC 7432].
- BD: Broadcast Domain.
- Bridge Table: An instantiation of a broadcast domain on a MAC-VRF.
- AC: A PE Attachment Circuit. This may be an access (untagged) or trunk (tagged) layer-2 interface that is a member of a local VLAN or a BD.

2. PE <-> CE Control Plane Overview

The Link State over Ethernet (LSoE) protocol is defined in [LSOE] as a protocol over Ethernet links to auto-discover connected neighbor’s layer 2, layer 3 attributes, and encapsulations for the purpose of bringing up upper layer routing protocols. This document leverages LSoE as a PE-CE protocol in an EVPN network fabric on access links between an EVPN PE and CE. Specifically,

- o PE-CE control plane based on LSoE protocol is proposed for CE MAC learning as an alternative to data-plane based source MAC learning.
- o PE-CE control plane based on LSoE protocol is proposed for CE MAC-IP adjacency learning as an alternative to MAC-IP learning based on ARP/ND snooping.
- o PE-CE control plane based on LSoE is proposed for learning of IP Prefixes and associated overlay indexes, as an alternative to local configuration on the PE for use case defined in section 4.1 of [EVPN-PREFIX-ADV].

Note that any specification related to base LSoE protocol itself is considered out of scope for this document and will continue to be covered in the base protocol spec. This document will instead focus on procedures and TLV extensions needed to achieve the above learning on PE-CE links in an EVPN network. Any text that relates to the base protocol included in this document is simply background information in the context of use cases covered in this document. The reader should refer to the base LSoE protocol document for the exact LSoE protocol specification.

```
                         +------------------------+
                         | Underlay Network Fabric|
                         +------------------------+

                              BGP-EVPN Peering
                      <-------------------------------->

                 +------+             +------+      +------+
                 |  PE1 |  .....      |  PE2 |      |  PE3 |
                 +------+             +------+      +------+
                    |                     \            /
               LSoE Session                \   ESI    /
                    |                 LSoE  \       / LSoE
                 CE-host              to PE2 CE-Host  to PE3
```

Figure 1

An LSoE session is established on layer-2 access interfaces between the EVPN PE and each connected CE host device. A session end-point is identified by a peer device MAC address on a layer-2 interface. LSoE HELLO messages are used for end-point discovery and OPEN messages are exchanged between two end-points to establish an LSoE peering. Once LSoE peering is established, encapsulation TLVs are exchanged for learning.

In the context of an EVPN network, CE Attachment Circuits (AC logical interfaces) typically terminate in a BD on the PE, with multi-homed LAG interfaces used for EVPN all-active multi-homing. CE hosts may be directly connected to EVPN PEs via access ports, or may be connected on trunk-ports via another switch. In a common EVPN-IRB design, EVPN PEs also function as distributed first-hop gateways for hosts in a BD. While symmetric and asymmetric IRB designs are possible as discussed in [EVPN IRB], procedures described in subsequent sections assume symmetric IRB with distributed any-cast gateways on EVPN PEs. Any deviations from these procedures for asymmetric IRB design or a centralized IRB design will be covered in future updates to this document.

The next few sections will focus on additional LSoE TLVs and procedures needed for PE-CE learning on EVPN PE ACs without and with all-active multi-homing.

3. TLVs

This section defines new TLVs that are used by PE-CE control plane defined in this document.

3.1 Overlay IPv4 Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv4 and MAC bindings. Alternatively, it may also be used to carry an overlay MAC with a NULL IPv4 address in a non-IRB use case.

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |    Type = 8   |          PDU Length           |     Count     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |                 IPv4 Address                  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |   PrefixLen   |E|     Rsvd    |               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+               +
   |                         MAC Address                           |
   +               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |                 IPv4 Address                  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |   PrefixLen   |E|     Rsvd    |               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+               +
   |                         MAC Address                           |
   +               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |                   more ...                    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2

- o A new LSoE PDU type (8) is requested for this PDU.
- o The IPv4 Address is that of an overlay.
- o MAC address carries the MAC binding for the particular IPv4 address if one is set in the PDU. If an IPv4 address is not set, it simply signals an overlay MAC address.
- o EVPN flag ’E’ indicates if this encapsulation is being sent on behalf of a remote host learnt via EVPN. Use of this flag is covered in a later section.

This PDU is used to carry PE’s any-cast gateway IPv4 address and MAC bindings to a CE host device. Optionally, it may also be used to relay a remote CE’s IPv4 address and MAC bindings to a local CE host within a subnet, as well as to send local CE IPv4 address and MAC binding to the PE. Procedures related to use of this PDU are discussed in subsequent sections.

In comparison to IPv4 Encapsulation PDU defined in [LSOE], this PDU allows you to explicitly signal a MAC binding that MAY be different from the device MAC used to establish an LSoE peering via HELLO/OPEN messages exchange.

The encapsulation list in this PDU MUST follow full replace semantics as in the LSoE protocol specification.

3.2 Overlay IPv6 Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv6 and MAC bindings:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |    Type = 9   |           PDU Length          |     Count     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |                                               |
   +-+-+-+-+-+-+-+-+                                               +
   |                                                               |
   +                                                               +
   |                                                               |
   +                                                               +
   |                          IPv6 Address                         |
   +               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |   PrefixLen   |E|R|O|   Rsvd  |               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+               +
   |                          MAC Address                          |
   +               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |                   more ...                    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3

- o A new LSoE PDU type (9) is requested for this PDU.
- o The IPv6 Address is that of an overlay.
- o MAC address carries the MAC binding for IPv6 address in the PDU.
- o An EVPN flag ’E’ indicates if this encapsulation is being sent on behalf of a remote host learnt via EVPN. Usage of this flag is covered in a later section.
- o A Router flag ’R’ is used to carry "Router Flag" or "R-bit" as defined in [RFC4861]. Usage of this flag for the purpose of installing ND cache entries based on learning via this TLV is as defined in [RFC4861]

- An Override flag 'O' is used to carry "Override Flag" or "O-bit" as defined in [RFC4861]. Usage of this flag for the purpose of installing ND cache entries based on learning via this TLV is as defined in [RFC4861]

This PDU is used to carry PE's any-cast gateway IPv6 address and MAC bindings to a CE host device. Optionally, it may also be used to relay a remote CE's IPv6 address and MAC bindings to a local CE within a subnet, as well as to send local CE IPv6 address and MAC bindings to the PE. Procedures related to usage of this PDU are discussed in subsequent sections.

The encapsulation list contained in this PDU MUST follow full replace semantics as in the LSoE protocol specification.

3.3 Overlay IPv4 Prefix Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv4 prefix routes for prefixes behind a CE that does not run a dynamic routing protocol for use-case as defined in section 4.1 of [EVPN-PREFIX-ADV]:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   Type = 10   |          PDU Length           |     Count     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |         Prefix Count          |               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                  IPv4 Prefix                  |   PrefixLen   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                          IPv4 Prefix                          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   PrefixLen   |                    More...                    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                             GW IP                             |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     Rsvd      |                    More...                    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4

A CE device as defined in [EVPN-PREFIX-ADV], with prefixes behind it MAY use the above PDU to send these prefixes to an EVPN PE with itself as the GW. An EVPN PE MAY then advertise prefixes received via this PDU as RT-5, with TS as the GW, as defined in [EVPN-PREFIX-ADV].

o A new LSoE PDU type (10) is requested for this PDU.
o IPv4 Prefix is set to a prefix behind a CE.
o PrefixLen is set to IPv4 prefix length for the advertised prefix.
o GW-IP is set to the CE IPv4 address (advertised via Type 8 PDU).

Multiple prefixes may be set for a single GW IP. The encapsulation list contained in this PDU MUST follow full replace semantics as in the LSoE protocol specification.

3.4 Overlay IPv6 Prefix Encapsulation PDU

A new encapsulation PDU type is defined for the purpose of carrying overlay IPv6 prefix routes for prefixes behind a CE that does not run a dynamic routing protocol for use-case as defined in section 4.1 of [EVPN-PREFIX-ADV]:

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   Type = 10   |           PDU Length          |     Count     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               |          Prefix Count         |               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+               +
   |                                                               |
   +                                                               +
   |                                                               |
   +                                                               +
   |                          IPv6 Prefix                          |
   +                                               +-+-+-+-+-+-+-+-+
   |                                               |   PrefixLen   |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                                                               |
   +                                                               +
   |                                                               |
   +                          IPv4 Prefix                          +
   |                                                               |
   +                                                               +
   |                                                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |   PrefixLen   |                   more...                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                                                               |
   +                                                               +
   |                                                               |
   +                             GW IP                             +
   |                                                               |
   +                                                               +
   |                                                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     Rsvd      |                   more...                     |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5

A CE device as defined in [EVPN-PREFIX-ADV], with prefixes behind it

MAY use the above PDU to send these prefixes to an EVPN PE with itself as the GW. An EVPN PE MAY then advertise prefixes received via this PDU as RT-5, with TS as the GW, as defined in [EVPN-PREFIX-ADV].

- A new LSoE PDU type (11) is requested for this PDU.
- IPv6 Prefix is set to an IPv6 prefix behind a CE.
- PrefixLen is set to IPv6 prefix length for the advertised prefix.
- GW-IP is set to the CE IPv6 address (advertised via Type 9 PDU).

Multiple prefixes may be set for a single GW IP. The encapsulation list contained in this PDU MUST follow full replace semantics as in the LSoE protocol specification.

4. CE MAC/IP Learning on a PE AC

This section defines procedures for learning a connected CE MAC and IP on a PE local attachment circuit (AC).

4.1 PE <-> CE LSoE Session Establishment

On an EVPN PE,

- A HELLO and/or OPEN PDU sent from a CE host source MAC is received on a tagged or untagged interface that is member of a local BD, referred here to as an AC.
- OPEN messages are exchanged with the host on the AC.
- LSoE session is established to the host source MAC and bound to a local AC.

4.2 CE MAC/IP Learning

Overlay IPv4 and IPv6 encapsulation PDU types 8/9 from a CE are used for the purpose of CE MAC/IP learning on a PE:

- The EVPN flag 'E' MUST NOT be set in type 8/9 PDU from a CE.
- A MAC entry for the MAC received in a type 8/9 PDU MUST be installed in the MAC-VRF table pointing to the AC to which the session is bound.
- If an IPv4/IPv6 address is set in the PDU, an IPv4/IPv6 neighbor binding MUST be established for the IPv4/IPv6 address in the PDU to the MAC address in the PDU. In other words, a next-hop re-write for these IPv4/IPv6 neighbor entries MUST be installed using the MAC address in the PDU, and if required by forwarding logic, bound to the AC associated with the LSoE session.
- Note that an IPv4/IPv6 address MAY NOT be set in a type 8/9 PDU received from a CE, in which case this PDU is only used for MAC learning. This MAY be the case in a non-IRB EVPN network, wherein, an EVPN PE is not a first-hop router for the attached CEs.

5. PE Any-cast GW MAC/IP Learning on CE

If LSoE based host learning is enabled on a PE with a distributed any-cast gateway on the EVPN PE,

- o EVPN PE MUST send type 8/9 Overlay Encapsulation PDUs on associated ACs with LSoE sessions toward CE hosts.
- o Type 8/9 PDUs from an EVPN PE MUST be encoded with the any-cast gateway IPv4/IPv6 address and any-cast gateway MAC address.
- o EVPN flag 'E' MUST NOT be set in this PDU.
- o A CE MAY process type 8/9 PDUs to establish GW IP to MAC bindings and learn gateway MAC to LAG AC bindings, similar to handling of type 8/9 PDUs on the PE described above.

Handling of type 8/9 PDUs for the purpose of gateway learning on the host is desirable but optional. A CE MAY continue to use ARP and ND for this purpose.

6. Remote CE MAC/IP Learning on CE

For CE to CE intra-subnet flows across the overlay, CE needs to learn and install a neighbor IP to MAC binding for remote CEs. This is handled today either by flooding ARP/ND requests across the overlay bridge and optionally implementing an ARP/ND suppression cache on the PE that is populated via MAC+IP EVPN route-type 2. ARP/ND request frames are trapped on the PE that does a local ARP/ND reply on behalf of the remote CE. If LSoE based learning is enabled in the fabric, LSoE may be used for this purpose to avoid overlay ARP/ND flooding, data frame triggered ARP learning, and to avoid maintaining an ARP suppression cache on the PE.

- o Remote MAC-IP routes learned via BGP EVPN route-type 2 that are imported to a local MAC-VRF MAY also be sent as type 8/9 PDUs on LSoE sessions to CEs over local ACs in that BD.
- o EVPN flag 'E' MUST be set in this encapsulation in the PDU.
- o A CE MAY install IPv4/IPv6 neighbor MAC bindings for remote CEs within a subnet based on 'E' flagged type 8/9 PDUs received from the PE.

Handling of type 8/9 PDUs for this purpose is optional but desirable to get full benefit of a fabric that is completely setup on boot-up, avoids overlay flooding, and is decoupled from latencies associated with data plane driven ARP and ND learning.

7. PE <-> CE Control Plane with EVPN All-active Multi-Homing

```
                          +------------------------+
                          | Underlay Network Fabric|
                          +------------------------+

                                BGP-EVPN Peering
             <---------------------------------------------------->
       +------+        +------+                +------+        +------+
       |  PE1 |        |  PE2 |     .....      |  PEx |        |  PEy |
       +------+        +------+                +------+        +------+
          \               /                       \               /
           \             /                         \             /
            \           /                           \           /
             \  ESI-a  /                             \  ESI-b  /
         LSoE \       / LSoE                     LSoE \       / LSoE
         to PE1\     /  to PE2                   to PEx\     /  to PEy
                CE-Host                                 CE-Host
```

Figure 6

In an EVPN all-active multi-homing setup, a LAG interface on the CE includes member physical ports that connect to multiple PE devices. A subset of these member ports that terminate at a PE are configured as members of a local LAG interface at that PE. A LAG AC at the PE is a logical interface in a BD, identified by this LAG interface and optionally, an Ethernet Tag in case of trunk ports.

In order for LSoE based learning to work with EVPN all-active multi-homing, a separate LSoE peering MUST be established between the CE host and each PE device. For this reason, while an EVPN PE MAY form an LSoE peering to a CE host on its local LAG AC, the CE host MUST form an LSoE peering to a PE on a local LAG "member physical port".

A configurable All-active Multi-Homing mode is defined below in order to be able to bind an LSoE peering to a LAG member-port as opposed to a LAG interface.

7.1 All-active Multi-Homing Mode

When configured to run on a local LAG port in this mode,

- o LSoE HELLO messages MUST be replicated on ALL LAG member ports.
- o An LSoE OPEN message sent in response to a HELLO MUST be sent on the LAG member port on which the HELLO was received.
- o An LSoE session MUST be bound to the local LAG member port on

which the OPEN message was received.

- LSoE encapsulation PDUs MUST be sent on the local LAG member port on which the session was bound.
- LSoE Keep-Alives MUST be sent on the local LAG member port on which the session was bound.

Note that this may result in a PE receiving multiple HELLO PDUs from a CE MAC. This however is harmless, as per the [LSOE] specification. A PE simply drops redundant HELLOs from a MAC that it has already replied to with an OPEN, within a retry time window.

7.2 Source MAC

LSoE relies on the source MAC address in the Ethernet frame to establish a peering. When running LSoE on a LAG port (in all-active multi-homing mode or regular mode), LSoE frames MUST use the LAG interface MAC as the source MAC address in the Ethernet frame.

7.3 CE MAC/IP Learning with EVPN All-active Multi-Homing

In order to accomplish MAC/IP learning of CE host devices multi-homed to EVPN fabric PEs via EVPN All-active Multi-Homing:

- A multi-homed CE device MUST be configured to run LSoE on a local LAG interfaces in All-active Multi-Homing mode defined above.
- EVPN PE MAY run LSoE on local LAG interfaces to multi-homed CE devices in regular mode.
- EVPN PEs that share the same Ethernet Segment MUST use unique source MACs (that of the local LAG) in HELLO/OPEN messages to establish separate LSoE sessions to a CE.

With the above rules in place,

- An LSoE session on the CE is bound to a local LAG member-port.
- An LSoE session on the PE is bound to a local LAG AC port.
- A single LSoE session is established at the PE to a CE on the local LAG AC.
- 'N' LSoE sessions are established at the CE, one to each PE on a local LAG member interface, where N = number of multi-homing PEs in an Ethernet Segment.

Once an LSoE session is established as above, all other host learning procedures defined earlier for CE MAC/IP learning on a PE's AC port apply as is to a LAG AC in an EVPN all-active multi-homing setup.

7.4 LAG Member Link Failure

On a CE that is running in all-active multi-homing mode, an LSoE session to a PE is bound to a LAG member interface. If the link that the LSoE session is bound to fails, LSoE session will get torn down at the CE by virtue of the session interface going down. If the CE has additional active member link(s) to this PE, a new LSoE session must be established on one of the active member links via HELLO PDUs sent by the CE on its remaining active member links to the PE.

7.4.1 Session Re-establishment

LSoE session at the CE is torn down immediately following the session interface failure. While the LAG interface at the PE is still operationally UP, LSoE session at the PE is subject to Keep Alive PDUs received from the CE. Once the session expires at the PE because of missed Keep Alive PDUs from the CE, PE will respond to HELLO on one of the active member link with an OPEN to re-establish a new session. Note that the new session is still bound to the LAG AC at the PE and to a new member link at the CE.

7.4.2 TLV Retention

TLVs learnt from a CE over a failed session MUST be retained at the PE if the PE LAG AC is still operationally up following a member link failure because of active member link(s) in the LAG. TLV retention logic at the PE MAY be based on an age-out time, that is a local matter at the PE. TLV age-out time MUST be higher than the missed Keep Alive duration, after which the session is considered closed. Once a new LSoE session is established, PE MUST implement a mark and sweep logic to reconcile retained TLVs from the CE peer with the new set of TLVs received from this CE.

7.4 LAG Failure

When a LAG member link failure results in the LAG interface being operationally down, TLV age-out logic discussed above MUST NOT be in effect. LSoE session MAY be be considered as DOWN immediately on the LAG being down at the PE. This is so that, in the event of a total connectivity loss between a PE and CE, CE learnt routes can be withdrawn immediately.

7.5 Example PE <-> CE Control Plane Flow with All-active Multi-Homing

An example LSoE over all-active multi-homing session flow is discussed below for clarity.

```
                 +-------------+      +-------------+
                 |             |      |             |
                 |     PE2     |      |     PE3     |
                 |             |      |             |
                 +-+-----------+      +-+-----------+
                   |    LAG    |        |    LAG    |
                   ++--+---+--++        ++--+---+--++
                    |  |   |  |          |  |   |  |
                    |  |   |  |          |  |   |  |
                    |  |   |  |          |  |   |  |
                    |  |   |  |          |  |   |  |
                    |  |   |  |          |  |   |  |
                 +--+--+---+--+----------+--+---+--++
                 |               LAG                |
                 +----------------------------------+-+
                 |                                    |
                 |                 H1                 |
                 |                                    |
                 +------------------------------------+
```

Figure 7

Example topology with CE H1 multi-homed to PE2 and PE3 via EVPN all-active multi-homing LAG with four member ports to each PE:

```
H1 member ports to PE2:    i121, i122, i123, i124
                            |     |     |    |
PE2 member ports to H1:    i211, i212, i213, i214

H1 member ports to PE3:    i131, i132, i133, i134
                            |     |     |    |
PE3 member ports to H1:    i311, i312, i313, i314
```

H1 LAG port to PE2/PE3:	MLAG1
PE2 LAG port to H1:	LAG2
PE3 LAG port to H1:	LAG3
H1 LAG MAC:	LMAC1
PE2 LAG MAC:	LMAC2
PE3 LAG MAC:	LMAC3

H1 running LSoE on MLAG1 in All-active Multi-Homing mode
PE2 running LSoE on LAG2 in regular mode
PE3 running LSoE on LAG3 in regular mode

```
               PE2                  H1                  PE3

                |        HELLOs      |        HELLOs      |
            LAG2|<-------------------|------------------->|LAG3
            LAG2|<-------------------|------------------->|LAG3
            LAG2|<-------------------|------------------->|LAG3
            LAG2|<-------------------|------------------->|LAG3
                |                    |                    |
                |        OPEN        |        OPEN        |
                |------------------->|<-------------------|
            LAG2|                i122|i132                |LAG3
                |                    |                    |
                |        OPEN        |        OPEN        |
                |<-------------------|------------------->|
            LAG2|                i122|i132                |LAG3
                |                    |                    |
   Session to   |       Session to   |Session to          |Session to
   LMAC1 on LAG2|       LMAC2 on i122|LMAC3 on i132       |LMAC1 on LAG3
                |                    |                    |
                |      Encap-PDU     |     Encap-PDU      |
                |<-------------------|------------------->|
            LAG2|                i122|i132                |LAG3
                |         ACK        |        ACK         |
                |------------------->|<-------------------|
            LAG2|                    |                    |LAG3
                |                    |                    |
                |      Overlay-PDU   |     Overlay-PDU    |
                |------------------->|<-------------------|
            LAG2|                    |                    |LAG3
                |         ACK        |        ACK         |
                |<-------------------|------------------->|
            LAG2|                i122|i132                |LAG3
                |                    |                    |
```

Figure 8

In an example flow shown above:

- o H1: originates HELLO(SMAC=LMAC2) on all MLAG member ports
- o PE2: Multiple HELLO(SMAC=LMAC2) copies received on port LAG2
- o PE3: Multiple HELLO(SMAC=LMAC2) copies received on port LAG3
- o PE2: A single OPEN(SMAC=LMAC2, DMAC=LMAC1) sent on port LAG2
- o PE3: A single OPEN(SMAC=LMAC3, DMAC=LMAC1) sent on port LAG3
- o PE2/PE3:duplicate HELLOs from same source LMAC2 are ignored
- o H1: OPEN(SMAC=LMAC2, DMAC=LMAC1) received on member port i122
- o H1: OPEN(SMAC=LMAC1, DMAC=LMAC2) sent on member port i122
- o H1: Session established to LMAC2 on MLAG1 member port i122

- PE2: Session established to LMAC1 on LAG AC LAG2
- H1: OPEN(SMAC=LMAC3, DMAC=LMAC1) received on member port i132
- H1: OPEN(SMAC=LMAC1, DMAC=LMAC3) sent on member port i132
- H1: Session established to LMAC3 on MLAG member port i132
- PE3: Session established to LMAC1 on LAG AC LAG3
- H1: IP encapsulation PDUs (type 4/5) sent to LMAC2 and LMAC3
- PE2/PE3: H1 MAC and IP are learned
- PE2/PE3: overlay IP encapsulation PDUs (type 8/9) sent to LMAC1
- H1: Any-cast GW MAC and IP are learned
- H1: Remote host MAC and IP are learned

8. Software Neighbor Tables

Some networking stack implementations rely on ARP and ND populated neighbor tables for software forwarding. In order to inter-work with such an implementation, an LsoE learned IPv4/IPv6 neighbor entry MAY also be installed in ARP and ND neighbor table as a static / permanent entry.

In addition,

- Pre-installing LSoE learned neighbor entries may help reduce potential conflict with ARP or ND learned neighbor entries.
- Pre-installing LSoE learned neighbor entries may help reduce reliance on data traffic triggered ARP requests / ND solicitations and associated learning latency.

With respect to installing IPv6 entries learnt via LSoE in IPv6 ND cache, Router flag (R-bit) and Override flag (O-bit) received in LSoE PDU should be handled as defined in [RFC4861].

9. MAC/IP Learning Conflict Resolution

If LSoE learned neighbor entries are not already installed as static entries in ARP/ND neighbor table, it is possible that a neighbor IPv4/IPv6 adjacency may be learned both via LSoE and ARP/ND. Even if LSoE learned entries were pre-installed in neighbor table, a race condition is still possible leading to a potential conflict between ARP/ND learned and LSoE learned neighbor IP adjacency. In such scenarios, LSoE learned entry should be preferred for the purpose of programming neighbor IP adjacencies in forwarding.

With respect to MAC-VRF entries, it is recommended that data plane learning be turned off when LSoE based learning is enabled. However, if it is not, data plane learned entries MUST be reconciled with LSoE learned entries in software and, in case of a conflict, LSoE learned entries preferred if LSoE based learning is enabled.

10. PE-CE Overlay Prefix Learning

[EVPN-PREFIX-ADV] section 4.1 defines a use case, wherein, a PE may advertise IP prefixes and subnets behind a CE. In this use case, CE device does not run a dynamic routing protocol. Instead, these prefixes are learnt on the PE via local policy or configuration. Prefixes are then advertised by PE as RT-5 with the CE as the GW.

PE-CE control plane defined in this document MAY be used to learn these prefixes from a CE as an alternative to local configuration on the PE. Once an LSoE session is established between a CE and a PE, as discussed earlier,

- o A CE MAY send type 10/11 PDUs with these IPv4/IPv6 prefixes over an LSoE session to a PE with the CE IP as the GW IP.
- o A PE MAY advertise prefixes learnt via type 10/11 PDUs as RT-5 with CE IP as the GW IP.

To summarize, A PE would advertise:

- o RT-2 for the CE MAC-IP learnt via type 8/9 PDU
- o RT-5 for Prefixes learnt via type 10/11 PDU with GW IP = CE IP

11. Asymmetric EVPN-IRB

Any deviations from the above procedures proposed in this document for asymmetric IRB design will be covered in subsequent updates to this document.

12. Centralized Gateway EVPN-IRB

Any deviations from the above procedures proposed in this document for centralized GW based IRB design will be covered in subsequent updates to this document.

13. Use Cases

13.1 Simplified EVPN Operations

This section will discuss in detail, benefits and simplifications that may be achieved in the context of an EVPN network, if one chooses to implement PE-CE control plane defined in this document as opposed to using traditional data-plane and ARP/ND snooping based PE-CE learning.

13.1.1 EVPN All-active Multi-Homing

```
                          +------------------------+
                          | Underlay Network Fabric|
                          +------------------------+

                                BGP-EVPN Peering
             <--------------------------------------------------->
       +------+        +------+                +------+        +------+
       | PE1  |        |  PE2 |     .....      |  PEx |        |  PEy |
       +------+        +------+                +------+        +------+
          \               /                       \               /
           \             /                         \             /
            \           /                           \           /
             \  ESI-a  /                             \  ESI-b  /

              LAG Bundle                              LAG Bundle
              to CE Host                              to CE Host
```

Figure 9

Data plane and ARP/ND snooping based MAC/IP learning on PE-CE all-active multi-homed LAG ports is subject to unpredictable hashing of ARP, ND, and data frames from host to PE. As an example, an ARP request for a connected host might originate at PE1 but the resulting ARP response from the host might be received at PE2. Redundant EVPN PEs in all-active multi-homing mode typically handle this unpredictability via combination of methods below:

- o PEs can handle unsolicited ARP and ND response frames.
- o PEs can implement additional mechanism to SYNC ARP, ND, and MAC tables across all PEs in a redundancy group for optimal forwarding to locally connected hosts.
- o PEs can implement EVPN aliasing procedures discussed in [RFC 7432] OR re-originate SYNCed MAC-IP adjacencies as local RT-2 to achieve MAC ECMP across the overlay.
- o PEs can also re-originate SYNCed MAC-IP adjacencies as local RT-2 to achieve IP ECMP across the overlay OR implement IP aliasing procedures discussed in [EVPN-IP-ALIASING].
- o PEs can also ensure EVPN sequence number SYNC for local MAC entries for EVPN mobility procedures to work correctly, as discussed in [EVPN-IRB-MOBILITY].

The PE-CE control plane learning alternative defined in this document fully decouples MAC and IP learning over MLAG ports from unpredictable hashing of data, AR, ND frames on all-active multi-homed LAG member links. As a result, above procedures that essentially result from data-plane PE-CE learning on all-active multi-homed LAGs can be simplified via the PE-CE control plane alternative defined in this document.

13.1.2 Convergence on CE Host Moves

```
                    +-------------------------+
                    | Underlay Network Fabric|
                    +-------------------------+

                         BGP-EVPN Peering
                 <-------------------------------->

              +-------+        +-------+        +-------+
              |  PE1 |        |  PE2 |  ..... |  PEx |
              +-------+        +-------+        +-------+
                  |                |                |
                Hosts            Hosts            Hosts
```

Figure 10

Host mobility across EVPN PE switches is a common occurrence in a data center fabric for flexibility in work load placement across a DC. Further, a host move must result in minimal, if any, disruption to traffic flows / services to / from the device.

Data plane and ARP/ND snooping based PE-CE learning may result in unpredictable convergence times, following host moves for the following cases:

- o A host may or may not send any data packet immediately following a move.
- o A host may or may not send an unsolicited ARP following a move.

While probing procedures, discussed in the next sub-sections are typically used to minimize convergence time, certain scenarios discussed below may still result in extended convergence times and flooding.

13.1.2.1 Silent Hosts

If a host is silent for an extended period following a move from PE1 to PE2, any bridged traffic flow destined to this host will continue to be black-holed by PE1 until the MAC ages out at PE1. Once the the MAC ages out at PE1, any bridged traffic flow destined to the host is flooded across the overlay bridge. Flooding of unknown unicast traffic on the overlay is enabled for this purpose. In summary, PE-CE learning that is based on data-plane and AR/ND snooping may be subject to non-deterministic convergence time and flooding following host moves because of being heavily dependent on unpredictable CE behavior.

PE-CE control plane based learning defined in this document fully decouples convergence in such scenarios from non-deterministic data flows and unsolicited ARP/ND behavior on a CE.

13.1.2.2 Probing

ARP and ND probing procedures are typically used to achieve host re-learning and convergence following host moves across the overlay:

- Following a host move from PE1 to PE2, the host’s MAC is discovered at PE2 as a local MAC via a data frames received from the host. If PE2 has a prior REMOTE MAC-IP host route for this MAC from PE1, an ARP probe is typically triggered at PE2 to learn the MAC-IP as a local IP adjacency and triggers EVPN RT-2 advertisement for this MAC-IP across the overlay with new reachability via PE2.

- Following a host move from PE1 to PE2, once PE1 receives a MAC or MAC-IP route from PE2 with a higher sequence number, an ARP probe is triggered at PE1 to clear the stale local MAC-IP neighbor adjacency OR re-learn the local MAC-IP in case the host has moved back or is duplicate.

- Following a local MAC age-out, if there is a local IP adjacency with this MAC, an ARP probe is triggered for this IP to either re-learn the local MAC and maintain local l3 and l2 reachability to this host OR to clear the ARP entry in case the host is indeed no longer local. Note that clearing of stale ARP entries, following a move is required for traffic to converge in the event that the host was silent and not discovered at its new location. Once stale ARP entry for the host is cleared, routed traffic flow destined for the host can re-trigger ARP discovery for this host at the new location. ARP flooding on the overlay MUST also be done to enable ARP discovery via routed flows.

- Alternatively, ARP probing timer may be tuned to be smaller than the MAC aging timer to avoid MAC age-out.

PE-CE control plane learning alternative defined in this document decouples host learning following moves from unpredictable host behavior with respect to sending data traffic and unsolicited ARPs, and as a result from ARP probing and MAC aging timer settings. Host move handling is hence greatly simplified to a very predictable and deterministic behavior.

13.1.3 ARP Gleaning Latency

If a CE's ARP binding is not already learned on a PE via an unsolicited ARP sent by the CE following events such as boot-up, flaps, and moves, a data frame that needs to be routed to the CE triggers ARP or ND discovery process on the PE. On a typical hardware switching platform, an IP packet that does not resolve to a link layer re-write would be punted to host stack that delivers packets with incomplete link-layer resolution to ARP or ND for resolution. An ARP request / ND Solicitation is generated for the CE IP and an ARP response or NA results in installing a link-layer re-write for the CE IP. In an EVPN multi-homing environment, this procedure is further complicated as the response is only received by one of the PEs that may or may not be the one that generated the ARP or ND request. Learned neighbor binding is SYNCed to other PEs that share the multi-homed Ethernet Segment. Routed flows can now be forwarded to the host via all PEs. Latency associated with such data frame driven ARP discovery may result in significant initial convergence hit, following triggers that warrant re-gleaning of CE IP to MAC binding.

PE-CE control plane learning alternative defined in this document results in proactive host learning following these scenarios, potentially avoiding a convergence hit on initial data packets.

13.2 Applicability to non-EVPN Use Cases

While the LSoE based host learning procedure described in this document focuses on EVPN-IRB overlay fabric use case, it may also have benefits and applicability in non-EVPN use cases. Applicability of procedures described in this document to non-EVPN use cases is a topic for further study.

14. Summary

PE-CE control plane is proposed as an alternative to data plane and ARP/ND snooping based PE-CE host MAC/IP learning and for PE-CE prefix learning. With a PE-CE control plane, CE host MAC and IP are deterministically learned on host boot-up, on host configuration, across host moves, on convergence triggers such as link failures, flaps, and PE re-boots and on all-active multi-homing LAG links. A PE-CE control plane decouples CE MAC and IP learning from traffic flows sourced by a CE, from varying CE behavior with respect to sending unsolicited ARP/ND frames, and from hashing of CE sourced frames over all-active multi-homed LAG links. As a result, it helps achieve a predictable and reliable convergence behavior across these triggers and helps simplify certain EVPN procedures that are otherwise needed with a data-plane and ARP/ND snooping based PE-CE learning. In addition, it may also be used for non-host learning use cases such as prefix learning.

15. References

15.1 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, <http://www.rfc-editor.org/info/rfc7432>.

[LSOE] Bush, R., Austein R., Patel, K., "Link State Over Ethernet", Feb 2019, <https://tools.ietf.org/html/draft-ietf-lsvr-lsoe-01>.

[EVPN-IRB] Sajassi, A., Salem, S., Thoria S., Drake J., Rabadan J., "Integrated Routing and Bridging in EVPN", July 2018, <https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-05>.

[EVPN-PREFIX-ADV] Rabadan J., Henderickx W., Drake J., Lin W., Sajassi, A., "IP Prefix Advertisement in EVPN", May 2018, <https://tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement-11>.

[EVPN-IRB-MOBILITY] Malhotra, N., Sajassi, A., Rabadan, J., Drake J., Lingala A., Patekar A., "Extended Mobility Procedures for EVPN-IRB", Jan 2019, <https://tools.ietf.org/html/draft-malhotra-bess-evpn-irb-extended-mobility-04>.

[EVPN-IP-ALIASING] Sajassi, A., Badoni, G., "L3 Aliasing and Mass Withdrawal Support for EVPN", July 2017, <https://tools.ietf.org/html/draft-sajassi-bess-evpn-ip-aliasing-00>.

[RFC2119] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", March 1997, <https://tools.ietf.org/html/rfc2119>.

[RFC8174] B. Leiba, "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", May 2017, <https://tools.ietf.org/html/rfc8174>.

15.2 Informative References

16. Acknowledgements

Authors would like to thank Randy Bush and Rob Austein for detailed review and feedback to ensure consistency with base LSOE protocol specification, as well as for helping build detailed LSOE flows included in this document.

Authors would like to thank Ali Sajassi and John Drake for detailed review and very valuable input on PE-CE protocol design for EVPN use cases as well as structuring this document for EVPN use cases.

Contributors

Randy Bush
Arrcus & IIJ
5147 Crystal Springs
Bainbridge Island, WA 98110
United States of America

Email: randy@psg.com

Authors' Addresses

Neeraj Malhotra (Editor)
Arrcus
2077 Gateway Place, Suite #400
San Jose, CA 95119, USA

Email: neeraj.ietf@gmail.com

Keyur Patel
Arrcus
2077 Gateway Place, Suite #400
San Jose, CA 95119, USA

Email: keyur@arrcus.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043, USA

Email: jorge.rabadan@nokia.com