Internet Engineering Task Force (IETF)
Request for Comments: 8388
Category: Informational
ISSN: 2070-1721

J. Rabadan, Ed.
S. Palislamovic
W. Henderickx
Nokia
A. Sajassi
Cisco
J. Uttaro
AT&T

May 2018

Usage and Applicability of BGP MPLS-Based Ethernet VPN

Abstract

This document discusses the usage and applicability of BGP MPLS-based Ethernet VPN (EVPN) in a simple and fairly common deployment scenario. The different EVPN procedures are explained in the example scenario along with the benefits and trade-offs of each option. This document is intended to provide a simplified guide for the deployment of EVPN networks.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are candidates for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc8388.

Table of Contents

1. Introduction

This document complements [RFC7432] by discussing the applicability of the technology in a simple and fairly common deployment scenario, which is described in Section 3.

After describing the topology and requirements of the use case scenario, Section 4 will describe the provisioning model.

Once the provisioning model is analyzed, Sections 5, 6, and 7 will describe the control-plane and data-plane procedures in the example scenario for the two potential disposition/forwarding models: MAC-based and MPLS-based models. While both models can interoperate in the same network, each one has different trade-offs that are analyzed in Section 8.

Finally, EVPN provides some potential traffic flow optimization tools that are also described in Section 9 in the context of the example scenario.

2. Terminology

The following terminology is used:

VID: VLAN Identifier

CE: Customer Edge (device)

EVI: EVPN Instance

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

ES: An Ethernet Segment is a set of links through which a CE is connected to one or more PEs. Each ES is identified by an Ethernet Segment Identifier (ESI) in the control plane.

CE-VIDs: The VLAN Identifier tags being used at CE1, CE2, and CE3 to tag customer traffic sent to the service provider EVPN network.

CE1-MAC, CE2-MAC, and CE3-MAC: The source MAC addresses "behind" each CE, respectively. These MAC addresses can belong to the CEs themselves or to devices connected to the CEs.

CE1-IP, CE2-IP, and CE3-IP: The IP addresses associated with the above MAC addresses

LACP: Link Aggregation Control Protocol

RD: Route Distinguisher

RT: Route Target

PE: Provider Edge (router)

AS: Autonomous System

PE-IP: The IP address of a given PE

3. Use Case Scenario Description and Requirements

Figure 1 depicts the scenario that will be referenced throughout the rest of the document.

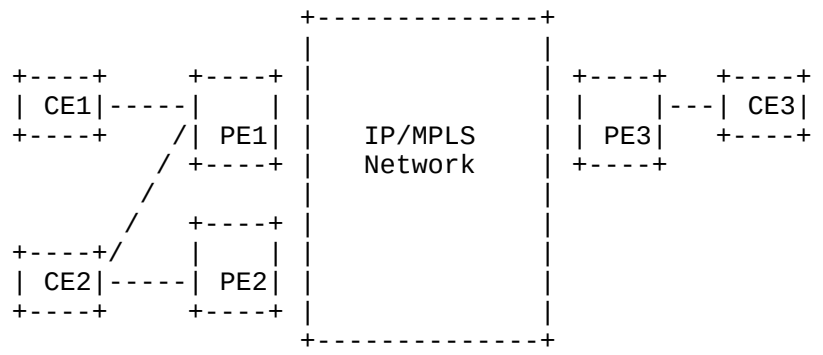

Figure 1: EVPN Use Case Scenario

There are three PEs and three CEs considered in this example: PE1, PE2, and PE3, as well as CE1, CE2, and CE3. Broadcast domains must be extended among the three CEs.

3.1. Service Requirements

The following service requirements are assumed in this scenario:

- Redundancy requirements:
  - CE2 requires multihoming connectivity to PE1 and PE2, not only for redundancy purposes but also for adding more upstream/downstream connectivity bandwidth to/from the network.
  - Fast convergence. For example, if the link between CE2 and PE1 goes down, a fast convergence mechanism must be supported so that PE3 can immediately send the traffic to PE2, irrespective of the number of affected services and MAC addresses.
- Service interface requirements:
  - The service definition must be flexible in terms of CE-VID-to-broadcast-domain assignment in the core.

- The following three EVI services are required in this example:

  EVI100 uses VLAN-based service interfaces in the three CEs with a 1:1 VLAN-to-EVI mapping. The CE-VIDs at the three CEs can be the same (for example, VID 100) or different at each CE (for instance, VID 101 in CE1, VID 102 in CE2, and VID 103 in CE3). A single broadcast domain needs to be created for EVI100 in any case; therefore, CE-VIDs will require translation at the egress PEs if they are not consistent across the three CEs. The case when the same CE-VID is used across the three CEs for EVI100 is referred to in [RFC7432] as the "Unique VLAN" EVPN case. This term will be used throughout this document too.

  EVI200 uses VLAN bundle service interfaces in CE1, CE2, and CE3 based on an N:1 VLAN-to-EVI mapping. The operator needs to preconfigure a range of CE-VIDs and its mapping to the EVI, and this mapping should be consistent in all the PEs (no translation is supported). A single broadcast domain is created for the customer. The customer is responsible for keeping the separation between users in different CE-VIDs.

  EVI300 uses VLAN-aware bundling service interfaces in CE1, CE2, and CE3. As in the EVI200 case, an N:1 VLAN-to-EVI mapping is created at the ingress PEs; however, in this case, a separate broadcast domain is required per CE-VID. The CE-VIDs can be different (hence, CE-VID translation is required).

Note that in Section 4.2.1, only EVI100 is used as an example of VLAN-based service provisioning. In Sections 6.2 and 7.2, 4k VLAN-based EVIs (EVI1 to EVI4k) are used so that the impact of MAC versus MPLS disposition models in the control plane can be evaluated. In the same way, EVI200 and EVI300 will be described with a 4k:1 mapping (CE-VIDs-to-EVI mapping) in Sections 6.3, 6.4, 7.3, and 7.4.

o Broadcast, Unknown Unicast, Multicast (BUM) optimization requirements:

  - The solution must support ingress replication or P2MP MPLS LSPs on a per EVI service. For example, we can use ingress replication for EVI100 and EVI200, assuming those EVIs will not carry much BUM traffic. On the contrary, if EVI300 is presumably carrying a significant amount of multicast traffic, P2MP MPLS LSPs can be used for this service.

  - The benefit of ingress replication compared to P2MP LSPs is that the core routers will not need to maintain any multicast states.

3.2. Why EVPN Is Chosen to Address This Use Case

Virtual Private LAN Service (VPLS) solutions based on [RFC4761], [RFC4762], and [RFC6074] cannot meet the requirements in Section 3, whereas EVPN can.

For example:

- If CE2 has a single CE-VID (or a few CE-VIDs), the current VPLS multihoming solutions (based on load-balancing per CE-VID or service) do not provide the optimized link utilization required in this example. EVPN provides the flow-based, load-balancing, multihoming solution required in this scenario to optimize the upstream/downstream link utilization between CE2 and PE1-PE2.

- EVPN provides a fast convergence solution that is independent of the CE-VIDs in the multihomed PEs. Upon failure on the link between CE2 and PE1, PE3 can immediately send the traffic to PE2 based on a single notification message being sent by PE1. This is not possible with VPLS solutions.

- With regard to service interfaces and mapping to broadcast domains, while VPLS might meet the requirements for EVI100 and EVI200, the VLAN-aware bundling service interfaces required by EVI300 are not supported by the current VPLS tools.

The rest of the document will describe how EVPN can be used to meet the service requirements described in Section 3 and even optimize the network further by:

- providing the user with an option to reduce (and even suppress) ARP (Address Resolution Protocol) flooding; and

- supporting ARP termination and inter-subnet forwarding.

4. Provisioning Model

One of the requirements stated in [RFC7209] is the ease of provisioning. BGP parameters and service context parameters should be auto-provisioned so that the addition of a new MAC-VRF to the EVI requires a minimum number of single-sided provisioning touches. However, this is possible only in a limited number of cases. This section describes the provisioning tasks required for the services described in Section 3, i.e., EVI100 (VLAN-based service interfaces), EVI200 (VLAN bundle service interfaces), and EVI300 (VLAN-aware bundling service interfaces).

4.1. Common Provisioning Tasks

Regardless of the service interface type (VLAN-based, VLAN bundle, or VLAN-aware), the following subsections describe the parameters to be provisioned in the three PEs.

4.1.1. Non-Service-Specific Parameters

The multihoming function in EVPN requires the provisioning of certain parameters that are not service specific and that are shared by all the MAC-VRFs in the node using the multihoming capabilities. In our use case, these parameters are only provisioned or auto-derived in PE1 and PE2 and are listed below:

- Ethernet Segment Identifier (ESI): Only the ESI associated with CE2 needs to be considered in our example. Single-homed CEs such as CE1 and CE3 do not require the provisioning of an ESI (the ESI will be coded as zero in the BGP Network Layer Reachability Information (NLRI)). In our example, a Link Aggregation Group (LAG) is used between CE2 and PE1-PE2 (since all-active multihoming is a requirement); therefore, the ESI can be auto-derived from the LACP information as described in [RFC7432]. Note that the ESI must be unique across all the PEs in the network; therefore, the auto-provisioning of the ESI is recommended only in case the CEs are managed by the operator. Otherwise, the ESI should be manually provisioned (Type 0, as in [RFC7432]) in order to avoid potential conflicts.

- ES-Import Route Target (ES-Import RT): This is the RT that will be sent by PE1 and PE2, along with the ES route. Regardless of how the ESI is provisioned in PE1 and PE2, the ES-Import RT must always be auto-derived from the 6-byte MAC address portion of the ESI value.

- Ethernet Segment Route Distinguisher (ES RD): This is the RD to be encoded in the ES route, and it is the Ethernet Auto-Discovery (A-D) route to be sent by PE1 and PE2 for the CE2 ESI. This RD should always be auto-derived from the PE-IP address, as described in [RFC7432].

- Multihoming type: The user must be able to provision the multihoming type to be used in the network. In our use case, the multihoming type will be set to all-active for the CE2 ESI. This piece of information is encoded in the ESI Label extended community flags and is sent by PE1 and PE2 along with the Ethernet A-D route for the CE2 ESI.

In addition, the same LACP parameters will be configured in PE1 and PE2 for the ES so that CE2 can send frames to PE1 and PE2 as though they were forming a single system.

4.1.2. Service-Specific Parameters

The following parameters must be provisioned in PE1, PE2, and PE3 per EVI service:

- EVI Identifier: The global identifier per EVI that is shared by all the PEs that are part of the EVI, i.e., PE1, PE2, and PE3 will be provisioned with EVI100, 200, and 300. The EVI identifier can be associated with (or be the same value as) the EVI default Ethernet Tag (4-byte default broadcast domain identifier for the EVI). The Ethernet Tag is different from zero in the EVPN BGP routes only if the service interface type (of the source PE) is a VLAN-aware bundle.

- EVI Route Distinguisher (EVI RD): This RD is a unique value across all the MAC-VRFs in a PE. Auto-derivation of this RD might be possible depending on the service interface type being used in the EVI. The next section discusses the specifics of each service interface type.

- EVI Route Target(s) (EVI RT): One or more RTs can be provisioned per MAC-VRF. The RT(s) imported and exported can be equal or different, just as the RT(s) in IP-VPNs. Auto-derivation of this RT(s) might be possible depending on the service interface type being used in the EVI. The next section discusses the specifics of each service interface type.

- CE-VID and port/LAG binding to EVI identifier or Ethernet Tag: For more information, please see Section 4.2.

4.2. Service-Interface-Dependent Provisioning Tasks

Depending on the service interface type being used in the EVI, a given CE-VID binding provision must be specified.

4.2.1. VLAN-Based Service Interface EVI

In our use case, EVI100 is a VLAN-based service interface EVI.

EVI100 can be a "unique-VLAN" service if the CE-VID being used for this service in CE1, CE2, and CE3 is identical (for example, VID 100). In that case, the VID 100 binding must be provisioned in PE1, PE2, and PE3 for EVI100 and the associated port or LAG. The MAC-VRF RD and RT can be auto-derived from the CE-VID:

- The auto-derived MAC-VRF RD will be a Type 1 RD, as recommended in [RFC7432], and it will be comprised of [PE-IP]:[zero-padded-VID]; where [PE-IP] is the IP address of the PE (a loopback address) and [zero-padded-VID] is a 2-byte value where the low-order 12 bits are the VID (VID 100 in our example) and the high-order 4 bits are zero.

- The auto-derived MAC-VRF RT will be composed of [AS]:[zero-padded-VID]; where [AS] is the Autonomous System that the PE belongs to and [zero-padded-VID] is a 2- or 4-byte value where the low-order 12 bits are the VID (VID 100 in our example) and the high-order bits are zero. Note that auto-deriving the RT implies supporting a basic any-to-any topology in the EVI and using the same import and export RT in the EVI.

If EVI100 is not a "unique-VLAN" instance, each individual CE-VID must be configured in each PE, and MAC-VRF RDs and RTs cannot be auto-derived; hence, they must be provisioned by the user.

4.2.2. VLAN Bundle Service Interface EVI

Assuming EVI200 is a VLAN bundle service interface EVI, and VIDs 200-250 are assigned to EVI200, the CE-VID bundle 200-250 must be provisioned on PE1, PE2, and PE3. Note that this model does not allow CE-VID translation and the CEs must use the same CE-VIDs for EVI200. No auto-derived EVI RDs or EVI RTs are possible.

4.2.3. VLAN-Aware Bundling Service Interface EVI

If EVI300 is a VLAN-aware bundling service interface EVI, CE-VID binding to EVI300 does not have to match on the three PEs (only on PE1 and PE2, since they are part of the same ES). For example, PE1 and PE2 CE-VID binding to EVI300 can be set to the range 300-310 and PE3 to 321-330. Note that each individual CE-VID will be assigned to a different broadcast domain, which will be represented by an Ethernet Tag in the control plane.

Therefore, besides the CE-VID bundle range bound to EVI300 in each PE, associations between each individual CE-VID and the corresponding EVPN Ethernet Tag must be provisioned by the user. No auto-derived EVI RDs/RTs are possible.

5. BGP EVPN NLRI Usage

[RFC7432] defines four different route types and four different extended communities. However, not all the PEs in an EVPN network must generate and process all the different routes and extended communities. Table 1 shows the routes that must be exported and imported in the use case described in this document. "Export", in this context, means that the PE must be capable of generating and exporting a given route, assuming there are no BGP policies to prevent it. In the same way, "Import" means the PE must be capable of importing and processing a given route, assuming the right RTs and policies. "N/A" means neither import nor export actions are required.

BGP EVPN Routes	PE1-PE2	PE3
ES	Export/Import	N/A
A-D per ESI	Export/Import	Import
A-D per EVI	Export/Import	Import
MAC	Export/Import	Export/Import
Inclusive Mcast	Export/Import	Export/Import

Table 1: Base EVPN Routes and Export/Import Actions

PE3 is required to export only MAC and Inclusive Multicast (Mcast) routes and be able to import and process A-D routes as well as MAC and Inclusive Multicast routes. If PE3 did not support importing and processing A-D routes per ESI and per EVI, fast convergence and aliasing functions (respectively) would not be possible in this use case.

6. MAC-Based Forwarding Model Use Case

This section describes how the BGP EVPN routes are exported and imported by the PEs in our use case as well as how traffic is forwarded assuming that PE1, PE2, and PE3 support a MAC-based forwarding model. In order to compare the control- and data-plane impact in the two forwarding models (MAC-based and MPLS-based) and different service types, we will assume that CE1, CE2, and CE3 need to exchange traffic for up to 4k CE-VIDs.

6.1. EVPN Network Startup Procedures

Before any EVI is provisioned in the network, the following procedures are required:

- Infrastructure setup: The proper MPLS infrastructure must be set up among PE1, PE2, and PE3 so that the EVPN services can make use of Point-to-Point (P2P) and P2MP LSPs. In addition to the MPLS transport, PE1 and PE2 must be properly configured with the same LACP configuration to CE2. Details are provided in [RFC7432]. Once the LAG is properly set up, the ESI for the CE2 Ethernet Segment (for example, ESI12) can be auto-generated by PE1 and PE2 from the LACP information exchanged with CE2 (ESI Type 1), as discussed in Section 4.1. Alternatively, the ESI can also be manually provisioned on PE1 and PE2 (ESI Type 0). PE1 and PE2 will auto-configure a BGP policy that will import any ES route matching the auto-derived ES-Import RT for ESI12.

- Ethernet Segment route exchange and Designated Forwarder (DF) election: PE1 and PE2 will advertise a BGP Ethernet Segment route for ESI12, where the ESI RD and ES-Import RT will be auto-generated as discussed in Section 4.1.1. PE1 and PE2 will import the ES routes of each other and will run the DF election algorithm for any existing EVI (if any, at this point). PE3 will simply discard the route. Note that the DF election algorithm can support service carving so that the downstream BUM traffic from the network to CE2 can be load-balanced across PE1 and PE2 on a per-service basis.

At the end of this process, the network infrastructure is ready to start deploying EVPN services. PE1 and PE2 are aware of the existence of a shared Ethernet Segment, i.e., ESI12.

6.2. VLAN-Based Service Procedures

Assuming that the EVPN network must carry traffic among CE1, CE2, and CE3 for up to 4k CE-VIDs, the service provider can decide to implement VLAN-based service interface EVIs to accomplish it. In this case, each CE-VID will be individually mapped to a different EVI. While this means a total number of 4k MAC-VRFs are required per PE, the advantages of this approach are the auto-provisioning of most of the service parameters if no VLAN translation is needed (see Section 4.2.1) and great control over each individual customer broadcast domain. We assume in this section that the range of EVIs from 1 to 4k is provisioned in the network.

6.2.1. Service Startup Procedures

As soon as the EVIs are created in PE1, PE2, and PE3, the following control-plane actions are carried out:

- Flooding tree setup per EVI (4k routes): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI (up to 4k routes per PE) so that the flooding tree per EVI can be set up. Note that ingress replication or P2MP LSPs can be optionally signaled in the Provider Multicast Service Interface (PMSI) Tunnel attribute and the corresponding tree can be created.

- Ethernet A-D routes per ESI (a set of routes for ESI12): A set of A-D routes with a total list of 4k RTs (one per EVI) for ESI12 will be issued from PE1 and PE2 (it has to be a set of routes so that the total number of RTs can be conveyed). As per [RFC7432], each Ethernet A-D route per ESI is differentiated from the other routes in the set by a different Route Distinguisher (ES RD). This set will also include ESI Label extended communities with the active-standby flag set to zero (all-active multihoming type) and an ESI Label different from zero (used for split-horizon functions). These routes will be imported by the three PEs, since the RTs match the locally configured EVI RTs. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as discussed in [RFC7432].

- Ethernet A-D routes per EVI (4k routes): An A-D route per EVI will be sent by PE1 and PE2 for ESI12. Each individual route includes the corresponding EVI RT and an MPLS Label to be used by PE3 for the aliasing function. These routes will be imported by the three PEs.

6.2.2. Packet Walk-Through

Once the services are set up, the traffic can start flowing. Assuming there are no MAC addresses learned yet and that MAC learning at the access is performed in the data plane in our use case, this is the process followed upon receiving frames from each CE (for example, EVI1).

BUM frame example from CE1:

a. An ARP request with CE-VID=1 is issued from source MAC CE1-MAC (MAC address coming from CE1 or from a device connected to CE1) to find the MAC address of CE3-IP.

b. Based on the CE-VID, the frame is identified to be forwarded in the MAC-VRF-1 (EVI1) context. A source MAC lookup is done in the MAC FIB, and the sender's CE1-IP is looked up in the proxy ARP table within the MAC-VRF-1 (EVI1) context. If CE1-MAC/CE1-IP are unknown in both tables, three actions are carried out (assuming the source MAC is accepted by PE1):

   1. the forwarding state is added for the CE1-MAC associated with the corresponding port and CE-VID;

   2. the ARP request is snooped and the tuple CE1-MAC/CE1-IP is added to the proxy ARP table; and

   3. a BGP MAC Advertisement route is triggered from PE1 containing the EVI1 RD and RT, ESI=0, Ethernet-Tag=0, and CE1-MAC/CE1-IP, along with an MPLS Label assigned to MAC-VRF-1 from the PE1 Label space. Note that depending on the implementation, the MAC FIB and proxy ARP learning processes can independently send two BGP MAC advertisements instead of one (one containing only the CE1-MAC and another one containing CE1-MAC/CE1-IP).

   Since we assume a MAC forwarding model, a label per MAC-VRF is normally allocated and signaled by the three PEs for MAC Advertisement routes. Based on the RT, the route is imported by PE2 and PE3, and the forwarding state plus the ARP entry are added to their MAC-VRF-1 context. From this moment on, any ARP request from CE2 or CE3 destined to CE1-IP can be directly replied to by PE1, PE2, or PE3, and ARP flooding for CE1-IP is not needed in the core.

c. Since the ARP frame is a broadcast frame, it is forwarded by PE1 using the Inclusive Multicast Tree for EVI1 (CE-VID=1 tag should be kept if translation is required). Depending on the type of tree, the label stack may vary. For example, assuming ingress replication, the packet is replicated to PE2 and PE3 with the downstream allocated labels and the P2P LSP transport labels. No other labels are added to the stack.

d. Assuming PE1 is the DF for EVI1 on ESI12, the frame is locally replicated to CE2.

e. The MPLS-encapsulated frame gets to PE2 and PE3. Since PE2 is non-DF for EVI1 on ESI12, and there is no other CE connected to PE2, the frame is discarded. At PE3, the frame is de-encapsulated and the CE-VID is translated, if needed, and forwarded to CE3.

Any other type of BUM frame from CE1 would follow the same procedures. BUM frames from CE3 would follow the same procedures too.

BUM frame example from CE2:

a. An ARP request with CE-VID=1 is issued from source MAC CE2-MAC to find the MAC address of CE3-IP.

b. CE2 will hash the frame and will forward it to, for example, PE2. Based on the CE-VID, the frame is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB and the sender's CE2-IP is looked up in the proxy ARP table within the MAC-VRF-1 context. If both are unknown, three actions are carried out (assuming the source MAC is accepted by PE2):

    1. the forwarding state is added for the CE2-MAC associated with the corresponding LAG/ESI and CE-VID;

    2. the ARP request is snooped and the tuple CE2-MAC/CE2-IP is added to the proxy ARP table; and

    3. a BGP MAC Advertisement route is triggered from PE2 containing the EVI1 RD and RT, ESI=12, Ethernet-Tag=0, and CE2-MAC/CE2-IP, along with an MPLS Label assigned from the PE2 Label space (one label per MAC-VRF). Again, depending on the implementation, the MAC FIB and proxy ARP learning processes can independently send two BGP MAC advertisements instead of one.

    Note that since PE3 is not part of ESI12, it will install the forwarding state for CE2-MAC as long as the A-D routes for ESI12 are also active on PE3. On the contrary, PE1 is part of ESI12, therefore PE1 will not modify the forwarding state for CE2-MAC if it has previously learned CE2-MAC locally attached to ESI12. Otherwise, it will add the forwarding state for CE2-MAC associated with the local ESI12 port.

c. Assuming PE2 does not have the ARP information for CE3-IP yet, and since the ARP is a broadcast frame and PE2 is the non-DF for EVI1 on ESI12, the frame is forwarded by PE2 in the Inclusive Multicast Tree for EVI1, thus adding the ESI Label for ESI12 at the bottom of the stack. The ESI Label has been previously allocated and signaled by the A-D routes for ESI12. Note that, as per [RFC7432], if the result of the CE2 hashing is different and the frame is sent to PE1, PE1 should add the ESI Label too (PE1 is the DF for EVI1 on ESI12).

d. The MPLS-encapsulated frame gets to PE1 and PE3. PE1 de-encapsulates the Inclusive Multicast Tree Label(s) and, based on the ESI Label at the bottom of the stack, it decides to not forward the frame to the ESI12. It will pop the ESI Label and will replicate it to CE1, since CE1 is not part of the ESI identified by the ESI Label. At PE3, the Inclusive Multicast Tree Label is popped and the frame forwarded to CE3. If a P2MP LSP is used as the Inclusive Multicast Tree for EVI1, PE3 will find an ESI Label after popping the P2MP LSP Label. The ESI Label will simply be popped, since CE3 is not part of ESI12.

Unicast frame example from CE3 to CE1:

a. A unicast frame with CE-VID=1 is issued from source MAC CE3-MAC and destination MAC CE1-MAC (we assume PE3 has previously resolved an ARP request from CE3 to find the MAC of CE1-IP and has added CE3-MAC/CE3-IP to its proxy ARP table).

b. Based on the CE-VID, the frame is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB within the MAC-VRF-1 context and this time, since we assume CE3-MAC is known, no further actions are carried out as a result of the source lookup. A destination MAC lookup is performed next and the label stack associated with the MAC CE1-MAC is found (including the label associated with MAC-VRF-1 in PE1 and the P2P LSP Label to get to PE1). The unicast frame is then encapsulated and forwarded to PE1.

c. At PE1, the packet is identified to be part of EVI1 and a destination MAC lookup is performed in the MAC-VRF-1 context. The labels are popped and the frame is forwarded to CE1 with CE-VID=1.

   Unicast frames from CE1 to CE3 or from CE2 to CE3 follow the same procedures described above.

Unicast frame example from CE3 to CE2:

a. A unicast frame with CE-VID=1 is issued from source MAC CE3-MAC and destination MAC CE2-MAC (we assume PE3 has previously resolved an ARP request from CE3 to find the MAC of CE2-IP).

b. Based on the CE-VID, the frame is identified to be forwarded in the MAC-VRF-1 context. We assume CE3-MAC is known. A destination MAC lookup is performed next and PE3 finds CE2-MAC associated with PE2 on ESI12, an Ethernet Segment for which PE3 has two active A-D routes per ESI (from PE1 and PE2) and two active A-D routes for EVI1 (from PE1 and PE2). Based on a

hashing function for the frame, PE3 may decide to forward the frame using the label stack associated with PE2 (label received from the MAC Advertisement route) or the label stack associated with PE1 (label received from the A-D route per EVI for EVI1). Either way, the frame is encapsulated and sent to the remote PE.

c. At PE2 (or PE1), the packet is identified to be part of EVI1 based on the bottom label, and a destination MAC lookup is performed. At either PE (PE2 or PE1), the FIB lookup yields a local ESI12 port to which the frame is sent.

Unicast frames from CE1 to CE2 follow the same procedures.

6.3. VLAN Bundle Service Procedures

Instead of using VLAN-based interfaces, the operator can choose to implement VLAN bundle interfaces to carry the traffic for the 4k CE-VIDs among CE1, CE2, and CE3. If that is the case, the 4k CE-VIDs can be mapped to the same EVI (for example, EVI200) at each PE. The main advantage of this approach is the low control-plane overhead (reduced number of routes and labels) and easiness of provisioning at the expense of no control over the customer broadcast domains, i.e., a single Inclusive Multicast Tree for all the CE-VIDs and no CE-VID translation in the provider network.

6.3.1. Service Startup Procedures

As soon as the EVI200 is created in PE1, PE2, and PE3, the following control-plane actions are carried out:

- Flooding tree setup per EVI (one route): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI (hence, only one route per PE) so that the flooding tree per EVI can be set up. Note that ingress replication or P2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree can be created.

- Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a single RT (RT for EVI200), an ESI Label extended community with the active-standby flag set to zero (all-active multihoming type), and an ESI Label different from zero (used by the non-DF for split-horizon functions). This route will be imported by the three PEs, since the RT matches the locally configured EVI200 RT. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as described in [RFC7432].

- Ethernet A-D routes per EVI (one route): An A-D route (EVI200) will be sent by PE1 and PE2 for ESI12. This route includes the EVI200 RT and an MPLS Label to be used by PE3 for the aliasing function. This route will be imported by the three PEs.

6.3.2. Packet Walk-Through

The packet walk-through for the VLAN bundle case is similar to the one described for EVI1 in the VLAN-based case except for the way the CE-VID is handled by the ingress PE and the egress PE:

- No VLAN translation is allowed and the CE-VIDs are kept untouched from CE to CE, i.e., the ingress CE-VID must be kept at the imposition PE and at the disposition PE.

- The frame is identified to be forwarded in the MAC-VRF-200 context as long as its CE-VID belongs to the VLAN bundle defined in the PE1/PE2/PE3 port to CE1/CE2/CE3. Our example is a special VLAN bundle case since the entire CE-VID range is defined in the ports; therefore, any CE-VID would be part of EVI200.

Please refer to Section 6.2.2 for more information about the control-plane and forwarding-plane interaction for BUM and unicast traffic from the different CEs.

6.4. VLAN-Aware Bundling Service Procedures

The last potential service type analyzed in this document is VLAN-aware bundling. When this type of service interface is used to carry the 4k CE-VIDs among CE1, CE2, and CE3, all the CE-VIDs will be mapped to the same EVI (for example, EVI300). The difference, compared to the VLAN bundle service type in the previous section, is that each incoming CE-VID will also be mapped to a different "normalized" Ethernet Tag in addition to EVI300. If no translation is required, the Ethernet Tag will match the CE-VID. Otherwise, a translation between CE-VID and Ethernet Tag will be needed at the imposition PE and at the disposition PE. The main advantage of this approach is the ability to control customer broadcast domains while providing a single EVI to the customer.

6.4.1. Service Startup Procedures

As soon as the EVI300 is created in PE1, PE2, and PE3, the following control-plane actions are carried out:

- Flooding tree setup per EVI per Ethernet Tag (4k routes): Each PE will send one Inclusive Multicast Ethernet Tag route per EVI and per Ethernet Tag (hence, 4k routes per PE) so that the flooding

tree per customer broadcast domain can be set up. Note that ingress replication or P2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created. In the described use case, since all the CE-VIDs and Ethernet Tags are defined on the three PEs, multicast tree aggregation might make sense in order to save forwarding states.

- Ethernet A-D routes per ESI (one route for ESI12): A single A-D route for ESI12 will be issued from PE1 and PE2. This route will include a single RT (RT for EVI300), an ESI Label extended community with the active-standby flag set to zero (all-active multihoming type), and an ESI Label different than zero (used by the non-DF for split-horizon functions). This route will be imported by the three PEs, since the RT matches the locally configured EVI300 RT. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as described in [RFC7432].

- Ethernet A-D routes per EVI: A single A-D route (EVI300) may be sent by PE1 and PE2 for ESI12 in case no CE-VID translation is required. This route includes the EVI300 RT and an MPLS Label to be used by PE3 for the aliasing function. This route will be imported by the three PEs. Note that if CE-VID translation is required, an A-D per EVI route is required per Ethernet Tag (4k).

6.4.2. Packet Walk-Through

The packet walk-through for the VLAN-aware case is similar to the one described before. Compared to the other two cases, VLAN-aware services allow for CE-VID translation and for an N:1 CE-VID to EVI mapping. Both things are not supported at once in either of the two other service interfaces. Some differences compared to the packet walk-through described in Section 6.2.2 are as follows:

- At the ingress PE, the frames are identified to be forwarded in the EVI300 context as long as their CE-VID belong to the range defined in the PE port to the CE. In addition to it, CE-VID=x is mapped to a "normalized" Ethernet-Tag=y at the MAC-VRF-300 (where x and y might be equal if no translation is needed). Qualified learning is now required (a different bridge table is allocated within MAC-VRF-300 for each Ethernet Tag). Potentially, the same MAC could be learned in two different Ethernet Tag bridge tables of the same MAC-VRF.

- Any new locally learned MAC on the MAC-VRF-300/Ethernet-Tag=y interface is advertised by the ingress PE in a MAC Advertisement route using the now Ethernet Tag field (Ethernet-Tag=y) so that the remote PE learns the MAC associated with the MAC-VRF-300/

Ethernet-Tag=y FIB. Note that the Ethernet Tag field is not used in advertisements of MACs learned on VLAN-based or VLAN-bundle service interfaces.

- At the ingress PE, BUM frames are sent to the corresponding flooding tree for the particular Ethernet Tag they are mapped to. Each individual Ethernet Tag can have a different flooding tree within the same EVI300. For instance, Ethernet-Tag=y can use ingress replication to get to the remote PEs, whereas Ethernet-Tag=z can use a P2MP LSP.

- At the egress PE, Ethernet-Tag=y (for a given broadcast domain within MAC-VRF-300) can be translated to egress CE-VID=x. That is not possible for VLAN bundle interfaces. It is possible for VLAN-based interfaces, but it requires a separate MAC-VRF per CE-VID.

7. MPLS-Based Forwarding Model Use Case

EVPN supports an alternative forwarding model, usually referred to as the MPLS-based forwarding or disposition model, as opposed to the MAC-based forwarding or disposition model described in Section 6. Using the MPLS-based forwarding model instead of the MAC-based model might have an impact on the following:

- the number of forwarding states required; and

- the FIB where the forwarding states are handled (MAC FIB or MPLS Label FIB (LFIB)).

The MPLS-based forwarding model avoids the destination MAC lookup at the egress PE MAC FIB at the expense of increasing the number of next-hop forwarding states at the egress MPLS LFIB. This also has an impact on the control plane and the label allocation model, since an MPLS-based disposition PE must send as many routes and labels as required next-hops in the egress MAC-VRF. This concept is equivalent to the forwarding models supported in IP-VPNs at the egress PE, where an IP lookup in the IP-VPN FIB may or may not be necessary depending on the available next-hop forwarding states in the LFIB.

The following subsections highlight the impact on the control- and data-plane procedures described in Section 6 when an MPLS-based forwarding model is used.

Note that both forwarding models are compatible and interoperable in the same network. The implementation of either model in each PE is a local decision to the PE node.

7.1. Impact of MPLS-Based Forwarding on the EVPN Network Startup

The MPLS-based forwarding model has no impact on the procedures explained in Section 6.1.

7.2. Impact of MPLS-Based Forwarding on the VLAN-Based Service Procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has no impact in terms of the number of routes when all the service interfaces are based on VLAN. The differences for the use case described in this document are summarized in the following list:

- Flooding tree setup per EVI (4k routes per PE): There is no impact when compared to the MAC-based model.
- Ethernet A-D routes per ESI (one set of routes for ESI12 per PE): There is no impact compared to the MAC-based model.
- Ethernet A-D routes per EVI (4k routes per PE/ESI): There is no impact compared to the MAC-based model.
- MAC Advertisement routes: Instead of allocating and advertising the same MPLS Label for all the new MACs locally learned on the same MAC-VRF, a different label must be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label (at least per CE) must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence, less scalability) is required in terms of forwarding states. For example, if CE2 sends traffic from two different MACs to PE1, CE2-MAC1, and CE2-MAC2, the same MPLS Label=x can be re-used for both MAC advertisements, since they both share the same source ESI12. It is up to the PE1 implementation to use a different label per individual MAC within the same ES (even if only one label per ESI is enough).
- PE1, PE2, and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane but will rather add forwarding states to the MPLS LFIB.

7.3. Impact of MPLS-Based Forwarding on the VLAN Bundle Service Procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has no impact in terms of number of routes when all the service interfaces are VLAN bundle type. The differences for the use case described in this document are summarized in the following list:

- Flooding tree setup per EVI (one route): There is no impact compared to the MAC-based model.
- Ethernet A-D routes per ESI (one route for ESI12 per PE): There is no impact compared to the MAC-based model.
- Ethernet A-D routes per EVI (one route per PE/ESI): There is no impact compared to the MAC-based model since no VLAN translation is required.
- MAC Advertisement routes: Instead of allocating and advertising the same MPLS Label for all the new MACs locally learned on the same MAC-VRF, a different label must be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label (at least per CE) must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence, less scalability) is required in terms of forwarding states. It is up to the PE1 implementation to use a different label per individual MAC within the same ES (even if only one label per ESI is enough).
- PE1, PE2, and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane, but will rather add forwarding states to the MPLS LFIB.

7.4. Impact of MPLS-Based Forwarding on the VLAN-Aware Service Procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has no impact in terms of the number of A-D routes when all the service interfaces are of the VLAN-aware bundle type. The differences for the use case described in this document are summarized in the following list:

- Flooding tree setup per EVI (4k routes per PE): There is no impact compared to the MAC-based model.
- Ethernet A-D routes per ESI (one route for ESI12 per PE): There is no impact compared to the MAC-based model.

- Ethernet A-D routes per EVI (1 route per ESI or 4k routes per PE/ESI): PE1 and PE2 may send one route per ESI if no CE-VID translation is needed. However, 4k routes are normally sent for EVI300, one per <ESI, Ethernet Tag ID> tuple. This allows the egress PE to find out all the forwarding information in the MPLS LFIB and even support Ethernet Tag to CE-VID translation at the egress.

- MAC Advertisement routes: Instead of allocating and advertising the same MPLS Label for all the new MACs locally learned on the same MAC-VRF, a different label must be advertised per CE next-hop or MAC so that no MAC FIB lookup is needed at the egress PE. In general, this means that a different label (at least per CE) must be advertised, although the PE can decide to implement a label per MAC if more granularity (hence, less scalability) is required in terms of forwarding states. It is up to the PE1 implementation to use a different label per individual MAC within the same ES. Note that the Ethernet Tag will be set to a non-zero value for the MAC Advertisement routes. The same MAC address can be announced with a different Ethernet Tag value. This will make the advertising PE install two different forwarding states in the MPLS LFIB.

- PE1, PE2, and PE3 will not add forwarding states to the MAC FIB upon learning new local CE MAC addresses on the data plane but will rather add forwarding states to the MPLS LFIB.

8. Comparison between MAC-Based and MPLS-Based Egress Forwarding Models

Both forwarding models are possible in a network deployment, and each one has its own trade-offs.

Both forwarding models can save A-D routes per EVI when VLAN-aware bundling services are deployed and no CE-VID translation is required. While this saves a significant amount of routes, customers normally require CE-VID translation; hence, we assume an A-D per EVI route per <ESI, Ethernet Tag> is needed.

The MAC-based model saves a significant amount of MPLS Labels compared to the MPLS-based forwarding model. All the MACs and A-D routes for the same EVI can signal the same MPLS Label, saving labels from the local PE space. A MAC FIB lookup at the egress PE is required in order to do so.

The MPLS-based forwarding model can save forwarding states at the egress PEs if labels per next-hop CE (as opposed to per MAC) are implemented. No egress MAC lookup is required. Also, a different label per next-hop CE per MAC-VRF is consumed, as opposed to a single label per MAC-VRF.

Table 2 summarizes the resource implementation details of both models.

Resources	MAC-Based Model	MPLS-Based Model
MPLS Labels Consumed	1 per MAC-VRF	1 per CE/EVI
Egress PE Forwarding States	1 per MAC	1 per Next-Hop
Egress PE Lookups	2 (MPLS+MAC)	1 (MPLS)

Table 2: Resource Comparison between MAC-Based and MPLS-Based Models

The egress forwarding model is an implementation local to the egress PE and is independent of the model supported on the rest of the PEs; i.e., in our use case, PE1, PE2, and PE3 could have either egress forwarding model without any dependencies.

9. Traffic Flow Optimization

In addition to the procedures described across Sections 3 through 8, EVPN [RFC7432] procedures allow for optimized traffic handling in order to minimize unnecessary flooding across the entire infrastructure. Optimization is provided through specific ARP termination and the ability to block unknown unicast flooding. Additionally, EVPN procedures allow for intelligent, close to the source, inter-subnet forwarding and solves the commonly known suboptimal routing problem. Besides the traffic efficiency, ingress-based inter-subnet forwarding also optimizes packet forwarding rules and implementation at the egress nodes as well. Details of these procedures are outlined in Sections 9.1 and 9.2.

9.1. Control-Plane Procedures

9.1.1. MAC Learning Options

The fundamental premise of [RFC7432] is the notion of a different approach to MAC address learning compared to traditional IEEE 802.1 bridge learning methods; specifically, EVPN differentiates between data and control-plane-driven learning mechanisms.

Data-driven learning implies that there is no separate communication channel used to advertise and propagate MAC addresses. Rather, MAC addresses are learned through IEEE-defined bridge learning procedures as well as by snooping on DHCP and ARP requests. As different MAC addresses show up on different ports, the Layer 2 (L2) FIB is populated with the appropriate MAC addresses.

Control-plane-driven learning implies a communication channel that could be either a control-plane protocol or a management-plane mechanism. In the context of EVPN, two different learning procedures are defined: local and remote procedures.

- Local learning defines the procedures used for learning the MAC addresses of network elements locally connected to a MAC-VRF. Local learning could be implemented through all three learning procedures: control plane, management plane, and data plane. However, the expectation is that for most of the use cases, local learning through the data plane should be sufficient.

- Remote learning defines the procedures used for learning MAC addresses of network elements remotely connected to a MAC-VRF, i.e., far-end PEs. Remote learning procedures defined in [RFC7432] advocate using only control-plane learning, BGP specifically. Through the use of BGP EVPN NLRIs, the remote PE has the capability of advertising all the MAC addresses present in its local FIB.

9.1.2. Proxy ARP/ND

In EVPN, MAC addresses are advertised via the MAC/IP Advertisement route, as discussed in [RFC7432]. Optionally, an IP address can be advertised along with the MAC address advertisement. However, there are certain rules put in place in terms of IP address usage: if the MAC/IP Route contains an IP address, this particular IP address correlates directly with the advertised MAC address. Such advertisement allows us to build a proxy ARP / Neighbor Discovery (ND) table populated with the IP<->MAC bindings received from all the remote nodes.

Furthermore, based on these bindings, a local MAC-VRF can now provide proxy ARP/ND functionality for all ARP requests and ND solicitations directed to the IP address pool learned through BGP. Therefore, the amount of unnecessary L2 flooding (ARP/ND requests/solicitations in this case) can be further reduced by the introduction of proxy ARP/ND functionality across all EVI MAC-VRFs.

9.1.3. Unknown Unicast Flooding Suppression

Given that all locally learned MAC addresses are advertised through BGP to all remote PEs, suppressing flooding of any unknown unicast traffic towards the remote PEs is a feasible network optimization.

The assumption in the use case is made that any network device that appears on a remote MAC-VRF will somehow signal its presence to the network. This signaling can be done through, for example, gratuitous

ARPs. Once the remote PE acknowledges the presence of the node in the MAC-VRF, it will do two things: install its MAC address in its local FIB and advertise this MAC address to all other BGP speakers via EVPN NLRI. Therefore, we can assume that any active MAC address is propagated and learned through the entire EVI. Given that MAC addresses become prepopulated -- once nodes are alive on the network -- there is no need to flood any unknown unicast towards the remote PEs. If the owner of a given destination MAC is active, the BGP route will be present in the local RIB and FIB, assuming that the BGP import policies are successfully applied; otherwise, the owner of such destination MAC is not present on the network.

It is worth noting that unknown unicast flooding must not be suppressed unless (at least) one of the following two statements is given: a) control- or management-plane learning is performed throughout the entire EVI for all the MACs or b) all the EVI-attached devices signal their presence when they come up (Gratuitous ARP (GARP) packets or similar).

9.1.4. Optimization of Inter-Subnet Forwarding

In a scenario in which both L2 and L3 services are needed over the same physical topology, some interaction between EVPN and IP-VPN is required. A common way of stitching the two service planes is through the use of an Integrated Routing and Bridging (IRB) interface, which allows for traffic to be either routed or bridged depending on its destination MAC address. If the destination MAC address is the one from the IRB interface, traffic needs to be passed through a routing module and potentially be either routed to a remote PE or forwarded to a local subnet. If the destination MAC address is not the one from the IRB interface, the MAC-VRF follows standard bridging procedures.

A typical example of EVPN inter-subnet forwarding would be a scenario in which multiple IP subnets are part of a single or multiple EVIs, and they all belong to a single IP-VPN. In such topologies, it is desired that inter-subnet traffic can be efficiently routed without any tromboning effects in the network. Due to the overlapping physical and service topology in such scenarios, all inter-subnet connectivity will be locally routed through the IRB interface.

In addition to optimizing the traffic patterns in the network, local inter-subnet forwarding also greatly optimizes the amount of processing needed to cross the subnets. Through EVPN MAC advertisements, the local PE learns the real destination MAC address associated with the remote IP address and the inter-subnet forwarding

can happen locally. When the packet is received at the egress PE, it is directly mapped to an egress MAC-VRF and bypasses any egress IP-VPN processing.

Please refer to [EVPN-INTERSUBNET] for more information about the IP inter-subnet forwarding procedures in EVPN.

9.2. Packet Walk-Through Examples

Assuming that the services are set up according to Figure 1 in Section 3, the following flow optimization processes will take place in terms of creating, receiving, and forwarding packets across the network.

9.2.1. Proxy ARP Example for CE2-to-CE3 Traffic

Using Figure 1 in Section 3, consider EVI400 residing on PE1, PE2, and PE3 connecting CE2 and CE3 networks. Also, consider that PE1 and PE2 are part of the all-active multihoming ES for CE2, and that PE2 is elected designated forwarder for EVI400. We assume that all the PEs implement the proxy ARP functionality in the MAC-VRF-400 context.

In this scenario, PE3 will not only advertise the MAC addresses through the EVPN MAC Advertisement route but also IP addresses of individual hosts (i.e., /32 prefixes) behind CE3. Upon receiving the EVPN routes, PE1 and PE2 will install the MAC addresses in the MAC-VRF-400 FIB and, based on the associated received IP addresses, PE1 and PE2 can now build a proxy ARP table within the context of MAC-VRF-400.

From the forwarding perspective, when a node behind CE2 sends a frame destined to a node behind CE3, it will first send an ARP request to, for example, PE2 (based on the result of the CE2 hashing). Assuming that PE2 has populated its proxy ARP table for all active nodes behind the CE3, and that the IP address in the ARP message matches the entry in the table, PE2 will respond to the ARP request with the actual MAC address on behalf of the node behind CE3.

Once the nodes behind CE2 learn the actual MAC address of the nodes behind CE3, all the MAC-to-MAC communications between the two networks will be unicast.

9.2.2. Flood Suppression Example for CE1-to-CE3 Traffic

Using Figure 1 in Section 3, consider EVI500 residing on PE1 and PE3 connecting CE1 and CE3 networks. Consider that both PE1 and PE3 have disabled unknown unicast flooding for this specific EVI context. Once the network devices behind CE3 come online, they will learn

their MAC addresses and create local FIB entries for these devices. Note that local FIB entries could also be created through either a control or management plane between PE and CE as well. Consequently, PE3 will automatically create EVPN Type 2 MAC Advertisement routes and advertise all locally learned MAC addresses. The routes will also include the corresponding MPLS Label.

Given that PE1 automatically learns and installs all MAC addresses behind CE3, its MAC-VRF FIB will already be prepopulated with the respective next-hops and label assignments associated with the MAC addresses behind CE3. As such, as soon as the traffic sent by CE1 to nodes behind CE3 is received into the context of EVI500, PE1 will push the MPLS Label(s) onto the original Ethernet frame and send the packet to the MPLS network. As usual, once PE3 receives this packet, and depending on the forwarding model, PE3 will either do a next-hop lookup in the EVI500 context or just forward the traffic directly to the CE3. In the case that PE1 MAC-VRF-500 does not have a MAC entry for a specific destination that CE1 is trying to reach, PE1 will drop the frame since unknown unicast flooding is disabled.

Based on the assumption that all the MAC entries behind the CEs are prepopulated through gratuitous ARP and/or DHCP requests, if one specific MAC entry is not present in the MAC-VRF-500 FIB on PE1, the owner of that MAC is not alive on the network behind the CE3; hence, the traffic can be dropped at PE1 instead of flooding and consuming network bandwidth.

9.2.3. Optimization of Inter-subnet Forwarding Example for CE3-to-CE2 Traffic

Using Figure 1 in Section 3, consider that there is an IP-VPN 666 context residing on PE1, PE2, and PE3, which connects CE1, CE2, and CE3 into a single IP-VPN domain. Also consider that there are two EVIs present on the PEs, EVI600 and EVI60. Each IP subnet is associated with a different MAC-VRF context. Thus, there is a single subnet (subnet 600) between CE1 and CE3 that is established through EVI600. Similarly, there is another subnet (subnet 60) between CE2 and CE3 that is established through EVI60. Since both subnets are part of the same IP-VPN, there is a mapping of each EVI (or individual subnet) to a local IRB interface on the three PEs.

If a node behind CE2 wants to communicate with a node on the same subnet seating behind CE3, the communication flow will follow the standard EVPN procedures, i.e., FIB lookup within the PE1 (or PE2) after adding the corresponding EVPN label to the MPLS Label stack (downstream label allocation from PE3 for EVI60).

When it comes to crossing the subnet boundaries, the ingress PE implements local inter-subnet forwarding. For example, when a node behind CE2 (EVI60) sends a packet to a node behind CE1 (EVI600), the destination IP address will be in the subnet 600, but the destination MAC address will be the address of the source node's default gateway, which in this case will be an IRB interface on PE1 (connecting EVI60 to IP-VPN 666). Once PE1 sees the traffic destined to its own MAC address, it will route the packet to EVI600, i.e., it will change the source MAC address to the one of the IRB interface in EVI600 and change the destination MAC address to the address belonging to the node behind CE1, which is already populated in the MAC-VRF-600 FIB, either through data- or control-plane learning.

An important optimization to be noted is the local inter-subnet forwarding in lieu of IP-VPN routing. If the node from subnet 60 (behind CE2) is sending a packet to the remote end-node on subnet 600 (behind CE3), the mechanism in place still honors the local inter-subnet (inter-EVI) forwarding.

In our use case, therefore, when the node from subnet 60 behind CE2 sends traffic to the node on subnet 600 behind CE3, the destination MAC address is the PE1 MAC-VRF-60 IRB MAC address. However, once the traffic locally crosses EVIs to EVI600 (via the IRB interface on PE1), the source MAC address is changed to that of the IRB interface and the destination MAC address is changed to the one advertised by PE3 via EVPN and already installed in MAC-VRF-600. The rest of the forwarding through PE1 is using the MAC-VRF-600 forwarding context and label space.

Another very relevant optimization is due to the fact that traffic between PEs is forwarded through EVPN rather than through IP-VPN. In the example described above for traffic from EVI60 on CE2 to EVI600 on CE3, there is no need for IP-VPN processing on the egress PE3. Traffic is forwarded either to the EVI600 context in PE3 for further MAC lookup and next-hop processing or directly to the node behind CE3, depending on the egress forwarding model being used.

10. Security Considerations

Please refer to the "Security Considerations" section in [RFC7432]. The standards produced by the SIDR Working Group address secure route origin authentication (e.g., RFCs 6480 through 6493) and route advertisement security (e.g., RFCs 8205 through 8211). They protect the integrity and authenticity of IP address advertisements and ASN/IP prefix bindings. This document and [RFC7432] use BGP to convey other info (e.g., MAC addresses); thus, the protections offered by the SIDR WG RFCs are not applicable in this context.

11. IANA Considerations

This document has no IANA actions.

12. References

12.1. Normative References

[RFC7209] Sajassi, A., Aggarwal, R., Uttaro, J., Bitar, N., Henderickx, W., and A. Isaac, "Requirements for Ethernet VPN (EVPN)", RFC 7209, DOI 10.17487/RFC7209, May 2014, <https://www.rfc-editor.org/info/rfc7209>.

[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>.

12.2. Informative References

[RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <https://www.rfc-editor.org/info/rfc4761>.

[RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <https://www.rfc-editor.org/info/rfc4762>.

[RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, DOI 10.17487/RFC6074, January 2011, <https://www.rfc-editor.org/info/rfc6074>.

[EVPN-INTERSUBNET] Sajassi, A., Salam, S., Thoria, S., Drake, J., Rabadan, J., and L. Yong, "Integrated Routing and Bridging in EVPN", Work in Progress, draft-ietf-bess-evpn-inter-subnet-forwarding-03, February 2017.

Acknowledgments

The authors want to thank Giles Heron for his detailed review of the document. We also thank Stefan Plug and Eric Wunan for their comments.

Contributors

The following people contributed substantially to the content of this document and should be considered coauthors:

Florin Balus
Keyur Patel
Aldrin Isaac
Truman Boyes

Authors' Addresses

Jorge Rabadan (editor)
Nokia
777 E. Middlefield Road
Mountain View, CA 94043
United States America

Email: jorge.rabadan@nokia.com

Senad Palislamovic
Nokia

Email: senad.palislamovic@nokia.com

Wim Henderickx
Nokia
Copernicuslaan 50
2018 Antwerp
Belgium

Email: wim.henderickx@nokia.com

Ali Sajassi
Cisco

Email: sajassi@cisco.com

James Uttaro
AT&T

Email: uttaro@att.com