L2VPN Workgroup
Internet Draft

Intended status: Standards Track

J. Rabadan
S. Palislamovic
W. Henderickx
F. Balus
Alcatel-Lucent

J. Uttaro
AT&T

K. Patel
A. Sajassi
Cisco

A. Isaac
T. Boyes
Bloomberg

Expires: December 2013

June 26, 2013

Usage and applicability of BGP MPLS based Ethernet VPN
draft-rp-l2vpn-evpn-usage-00.txt

Abstract

This document discusses the usage and applicability of BGP MPLS based Ethernet VPN (E-VPN) in a simple and fairly common deployment scenario. The different E-VPN procedures will be explained on the example scenario, analyzing the benefits and trade-offs of each option. Along with [E-VPN], this document is intended to provide a simplified guide for the deployment of E-VPN in Service Provider networks.

This Internet-Draft will expire on December 28, 2013.

Table of Contents

1. Introduction

This document complements [E-VPN] by discussing the applicability of the technology in a simple and fairly common deployment scenario, which is described in section 2.

After describing the topology of the use-case scenario and the characteristics of the service to be deployed, the following section will describe the provisioning model, comparing the E-VPN procedures with the provisioning tasks required for other VPN technologies, such as VPLS or IP-VPN.

Once the provisioning model is analyzed, the following sections will describe the control plane and data plane procedures for the traffic 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 this document.

Finally, E-VPN provides some potential traffic flow optimization tools that are also described in the context of the example scenario.

2. Use-case scenario description

The following figure depicts the scenario that will be referenced throughout the rest of the document.

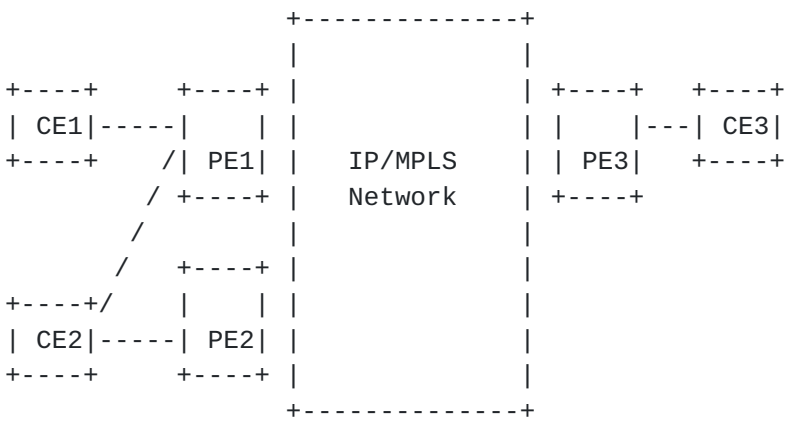

Figure 1 E-VPN use-case scenario

There are three PEs and three CEs considered in this example: PE1, PE2, PE3, as well as CE1, CE2 and CE3. Layer-2 traffic must be extended among the three CEs. The following service requirements are assumed in this scenario:

o Redundancy requirements: CE1 and CE3 are single-homed to PE1 and

PE3 respectively. CE2 requires multi-homing connectivity to PE1 and PE2, not only for redundancy purposes, but also for adding more upstream/downstream connectivity bandwidth to/from the network. If CE2 has a single CE-VID (or a few CE-VIDs) the current VPLS multi-homing solutions (based on load-balancing per CE-VID or service) do not provide the optimized link utilization required in this example. Another redundancy requirement that must be met is fast convergence. E.g.: 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, irrespectively of the number of affected services and MAC addresses. E-VPN provides the flow-based load-balancing multi-homing solution required in this scenario to optimize the upstream/downstream link utilization between CE2 and PE1-PE2. E-VPN also provides a fast convergence solution so that PE3 can immediately send the traffic to PE2 upon failure on the link between CE2 and PE1.

- Service interface requirements: service definition must be flexible in terms of CE-VID-to-broadcast-domain assignment and service contexts in the core. The following three services are required in this example:

  EVI100 - It will use VLAN-based service interfaces in the three CEs with a 1:1 mapping (VLAN-to-EVI). The CE-VIDs at the three CEs can be the same, e.g.: VID 100, or different at each CE, e.g.: 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 in [E-VPN] as the "Unique VLAN" E-VPN case. This term will be used throughout this document too.

  EVI200 - It will use VLAN-bundle service interfaces in CE1, CE2 and CE3, based on an N:1 VLAN-to-EVI mapping. In this case, the service provider just needs to assign a pre-configured number of CE-VIDs on the ingress PE to EVI200, and send the customer frames with the original CE-VIDs. The Service Provider will build a single broadcast domain for the customer. The customer will be responsible for the CE-VID handling.

  EVI300 - It will use VLAN-aware bundling service interfaces in CE1, CE2 and CE3. At the ingress PE, an N:1 VLAN-to-EVI mapping will be done, however and as opposed to EVI200, a separate core broadcast domain is required per CE-VID. In addition to that, the CE-VIDs can be different (hence CE-VID translation is required). Note that, while the requirements stated for EVI100 and EVI200 might be met with the current VPLS solutions, the VLAN-aware

bundling service interfaces required by EVI300 are not supported by the current VPLS tools.

o BUM (Broadcast, Unknown unicast, Multicast) optimization requirements: The solution must be able to support ingress replication, P2MP MPLS LSPs and MP2MP MPLS LSPs and the user must be able to decide what kind of provider tree will be used by each EVI service. For example, if we assume that EVI100 and EVI200 will not carry much BUM traffic, we can use ingress replication for those service instances. The benefit is that the core will not need to maintain any states for the multicast trees associated to EVI100 and EVI200. On the contrary, if EVI300 is presumably carrying a significant amount of multicast traffic, P2MP MPLS LSPs or MP2MP LSPs can be used for this service. Note that ingress replication and P2MP LSPs are supported by VPLS solutions (see [VPLS-MCAST]), however VPLS solutions do not support MP2MP LSPs, since the source of the tree must be identified for the data plane MAC learning, and that identification is challenging when using MP2MP LSPs. Since E-VPN uses the control plane for MAC learning, any type of provider multicast tree is supported in the core.

As already outlined above, the current VPLS solutions, based on [RFC4761][RFC4762][RFC6074], cannot meet all the above set of requirements and therefore a new solution is needed. The following sections will describe how E-VPN can be used to meet those service requirements and even optimize the network further by:

o Providing the user with an option to reduce (and even suppress) the ARP-flooding.

o Supporting ARP termination for inter-subnet forwarding

3. Provisioning Model

One of the requirements stated in [E-VPN-REQ] is the ease of provisioning. BGP parameters and service context parameters should be auto-provisioned so that the addition of a new EVI to the E-VPN requires a minimum number of single-sided provisioning touches. However this is only possible in a limited number of cases. This section describes the provisioning tasks required for the services described in section 2, i.e. EVI100 (VLAN-based service interfaces), EVI200 (VLAN-bundle service interfaces) and EVI300 (VLAN-aware bundling service interfaces).

3.1. Common provisioning tasks

Regardless of the service interface type (VLAN-based, VLAN-bundle or VLAN-aware), the following sub-sections describe the parameters to be

provisioned in the three PEs.

3.1.1. Non-service specific parameters

The multi-homing function in E-VPN requires the provisioning of certain parameters which are not service-specific and that are shared by all the EVIs using the multi-homing capabilities. In our use-case, these parameters are only provisioned in PE1 and PE2, and are listed below:

- Ethernet Segment Identifier (ESI): only the ESI associated to 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 NLRIs). In our example, a LAG is used between CE2 and PE1-PE2 (since all-active multi-homing is a requirement) therefore the ESI can be auto-derived from the LACP information as described in [E-VPN]. Note that the ESI MUST be unique across all the PEs in the network, therefore the auto-provisioning of the ESI is only recommended in case the CEs are managed by the Service Provider. Otherwise the ESI should be manually provisioned 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 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 [E-VPN].

- Multi-homing type: the user must be able to provision the multi-homing type to be used in the network. In our use-case, the multi-homing 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 sent by PE1 and PE2 along with the Ethernet A-D route for the CE2 ESI.

In our use-case, besides the above parameters, all the corresponding LAG and LACP parameters will be configured in PE1 and PE2, so that CE2 can send different flows to PE1 and PE2 for the same CE-VID as though they were forming a single system from the CE2 perspective.

3.1.2. Service specific parameters

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

EVI service:

- o EVI identifier: global identifier per EVI that is shared by all the PEs part of the EVI, i.e. PE1, PE2 and PE3 will be provisioned with EVI100, 200 and 300. The EVI identifier can be associated to (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 E-VPN BGP routes only if the service interface type (of the source PE) is VLAN-aware.

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

- o EVI Route Target(s) (EVI RT): one or more RTs can be provisioned per EVI. 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. Next section discusses the specifics of each service interface type.

- o CE-VID and port/LAG binding to EVI identifier or Ethernet Tag: see the next section.

3.2. Service interface dependent provisioning tasks

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

3.2.1. VLAN-based service interface EVI

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

EVI100 can be a "unique-VLAN" E-VPN if the CE-VID being used for this service in CE1, CE2 and CE3 is equal, e.g. 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 EVI RD and EVI RT can be auto-derived from the CE-VID:

- o The auto-derived EVI RD will be a Type 1 RD, as recommended in [E-VPN], and it will be comprised of [PE-IP]:[zero-padded-VID]; where PE-IP is the IP address of the PE (normally 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.

- o The auto-derived EVI 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 4-byte value where the low order 12 bits are the VID (VID 100 in our example) and the high order 20 bits are zero. Note that auto-deriving the EVI RT implies supporting a basic any-to-any topology in the E-VPN and using the same import and export RT in the EVI.

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

3.2.2. 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.

3.2.3. 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). E.g.: 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 core broadcast domain, i.e. Ethernet Tag, which will be encoded in the BGP E-VPN routes.

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

4. BGP E-VPN NLRI usage

[E-VPN] defines four different types of routes and four different extended communities advertised along with the different routes. However not all the PEs in a network must generate and process all the different routes and extended communities. The following table 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 E-VPN 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

PE3 is only required to export MAC and Inclusive multicast 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.

5. MAC-based forwarding model use-case

This section describes how the BGP E-VPN 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.

5.1. E-VPN Network Startup procedures

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

- Infrastructure setup: the proper MPLS infrastructure must be setup among PE1, PE2 and PE3 so that the E-VPN services can make use of P2P, P2MP and/or MP2MP LSPs. In addition to the MPLS transport, PE1 and PE2 must be properly configured to create a multi-chassis LAG to CE2. Details are provided in [E-VPN]. Once the LAG is properly setup, as discussed in section 3.1, the ESI for the CE2 Ethernet Segment, e.g. ESI12, can be auto-generated by PE1 and PE2 from the LACP information exchanged with CE2. Alternatively, the ESI can also be manually provisioned on PE1 and PE2. 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 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 3.1.1. 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 E-VPN services. PE1 and PE2 are aware of the existence of a shared Ethernet Segment, i.e. ESI12.

5.2. VLAN-based service procedures

Assuming that the E-VPN 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 EVIs is 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 3.2.1) 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.

5.2.1. 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 setup. Note that ingress replication, P2MP LSPs or MP2MP LSPs can optionally be signaled in the PMSI Tunnel attribute and the corresponding tree be created. In the described use-case, since all the EVIs have the same core topology, PMSI aggregation makes sense in order to save some multicast forwarding states in the core.

- 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 list of 4k RTs (one per EVI) and an ESI Label extended community with the active-standby flag set to zero (all-active multi-homing type) and an ESI Label different from zero (used by the non-DF for split-horizon functions). These routes will be imported by the three PEs, since the RTs match the EVI RTs locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as discussed in [E-VPN].

- 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.

5.2.2. Packet walkthrough

Once the services are setup, the traffic can start flowing. Assuming there are no MAC addresses learnt 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 packets from each CE (example for EVI1).

(1) BUM packet 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 packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI1 context and if CE1-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE1): (1) a forwarding state is added for CE1-MAC associated to the corresponding port and CE-VID, (2) the ARP-request is snooped and the tuple CE1-MAC/CE1-IP is added to the ARP proxy 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 EVI1 from the PE1 label space. Since we assume a MAC forwarding model, a label per EVI 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 ARP entry are added to their EVI1 context. From this moment on, any ARP request from CE2 or CE3 destined to CE1-IP, can be directly replied by PE1, PE2 or PE3 and ARP flooding for CE1-IP is not needed in the core.

c) Since the ARP packet is a broadcast packet, it is forwarded by PE1 using the Inclusive multicast tree for EVI1 (CE-VID=1 is kept if translation is required). Depending on the type of tree, the label stack may vary. E.g. assuming ingress replication and no aggregation, 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 packet is locally replicated to CE2.

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

Any other type of BUM packet from CE1 would follow the same procedures. BUM packets from CE3 would follow the same procedures too.

(2) BUM packet 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 packet and will forward it to e.g. PE2. Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI1 context and if CE2-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE2): (1) a forwarding state is added for CE2-MAC associated to the corresponding LAG/ESI and CE-VID, (2) the ARP-request is snooped and the tuple CE2-MAC/CE2-IP is added to the ARP proxy 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 EVI). Note that, since PE3 is not part of ESI12, it will install a forwarding state for CE2-MAC as long as the A-D route per ESI for ESI12 is 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 learnt CE2-MAC locally attached to ESI12. Otherwise it will add forwarding state for CE2-MAC.

c) Assuming PE2 does not have the ARP information for CE3-IP yet, and since the ARP is a broadcast packet and PE2 the non-DF for EVI1 on ESI12, the packet is forwarded by PE2 in the Inclusive multicast tree for EVI1, 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 if the result of the CE2 hashing had been different and the packet sent to PE1, PE1 would not have added the ESI label to the label stack (PE1 is the DF for EVI1 on ESI12).

d) The MPLS-encapsulated packet gets to PE1 and PE3. PE1 de-encapsulate the Inclusive multicast tree label(s) and based on the ESI label at the bottom of the stack, it decides to not forward the packet to the ESI12. It will pop the ESI label and will replicate it to CE1 though, since CE1 is not part of the ESI

identified by the ESI label. At PE3, the Inclusive multicast tree label(s) are popped and the packet forwarded to CE3. If a P2MP LSP is used as Inclusive multicast tree for EVI1, PE3 will find an ESI label after popping the P2MP LSP label. The ESI label will simply be ignored and popped, since CE3 is not part of ESI12.

(3) Unicast packet example from CE3 to CE1:

a) A unicast packet 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 ARP proxy table).

b) Based on the CE-VID, the packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB within the EVI1 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 to the MAC CE1-MAC is found (including the label associated to EVI1 in PE1 and the P2P LSP label to get to PE1). The unicast packet is then encapsulated and forwarded to PE1.

c) At PE1, the packet is identified to be part of EVI1 (based on the bottom of the stack label) and a destination MAC lookup is performed in the EVI1 context. The labels are popped and the packet forwarded to CE1 with CE-VID=1. Unicast packets from CE1 to CE3 or from CE2 to CE3 follow the same procedures described above.

(4) Unicast packet example from CE3 to CE2:

a) A unicast packet 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 packet is identified to be forwarded in the EVI1 context. A source MAC lookup is done in the MAC FIB within the EVI1 context and 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 PE3 finds CE2-MAC associated to 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 packet, PE3 may decide to forward the packet using the label stack associated to PE2 (label received from the MAC advertisement route) or the label stack associated to PE1 (label received from the A-D route per EVI for EVI1). Either way, the packet 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. In particular, if the packet arrives to PE2, the bottom label is assumed to be a label per EVI, hence a MAC lookup for the EVI1 context is done. If the packet arrives to PE1, the bottom label is assumed to be a label identifying ESI12, hence the packet is forwarded to ESI12.

Unicast packets from CE1 to CE2 follow the same procedures. Aliasing is possible in this case too, since ESI12 is local to PE1 and load balancing through PE1 and PE2 may happen.

5.3. VLAN-bundle service procedures

Instead of using VLAN-based interfaces, the Service Provider 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, e.g. 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.

5.3.1. Service startup procedures

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

o 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 setup. Note that ingress replication, P2MP LSPs or MP2MP 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 are part of the same EVI, a single tree is created for all of them.

o 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 multi-homing 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 EVI200 RT locally configured. The A-D routes per ESI will be used for fast

convergence and split-horizon functions, as described in [E-VPN].

o 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.

5.3.2. Packet Walkthrough

The packet walkthrough for the VLAN-bundle case is similar to the one described for EVI1 in the VLAN-based case except for some differences. The main difference is the fact that no VLAN translation is allowed and the CE-VIDs are kept untouched from CE to CE.

(1) BUM packet example from CE1:

a) An ARP-request tagged with any CE-VID is issued from source MAC CE1-MAC to find the MAC address of CE3-IP.

b) The packet is identified to be forwarded in the EVI200 context as long as its CE-VID belongs to the VLAN-bundle defined in the PE1 port to CE1. This case 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. A source MAC lookup is done next, in the MAC FIB and ARP proxy table within the EVI200 context and if CE1-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE1): (1) a forwarding state is added for CE1-MAC associated to the corresponding port (CE-VID is not taken into account), (2) the ARP-request is snooped and the tuple CE1- MAC/CE1-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE1 containing the EVI200 RD and RT, ESI=0, Ethernet-Tag=0 and CE1-MAC/CE1-IP along with an MPLS label assigned from the PE1 label space. Since we assume a MAC forwarding model, a label per EVI 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 ARP entry are added to their EVI200 context. From this moment on, any ARP request from CE2 or CE3 destined to CE1-IP, can be directly replied by PE1, PE2 or PE3 and ARP flooding for CE1-IP is not needed in the core.

c) Since the ARP is a broadcast packet, it is forwarded by PE1 using the Inclusive multicast tree for EVI200. Note that the ingress CE-VID MUST be kept at the imposition PE and the disposition PE. Depending on the type of tree, the label stack may vary. E.g. assuming ingress replication, the packet is replicated to PE2 and PE3 with the downstream allocated labels (by PE2 and PE3 respectively) and the P2P LSP transport labels. No other labels

are added to the stack.

d) Assuming PE1 is the DF for EVI200 on ESI12, the packet is locally replicated to CE2.

e) The MPLS-encapsulated packet gets to PE2 and PE3. Since PE2 is non-DF for EVI200 on ESI12 and there is no other CE connected, the packet is discarded. At PE3, the packet is de-encapsulated and replicated to CE3. The CE-VID remains untouched throughout the whole process.

Any other type of BUM packet from CE1 would follow the same procedures. BUM packets from CE3 would follow the same procedures too.

(2) BUM packet example from CE2:

a) An ARP-request, tagged with any CE-VID, is issued from source MAC CE2-MAC to find the MAC address of CE3-IP.

b) CE2 will hash the packet and will forward it to e.g. PE2. The packet CE-VID is identified to be forwarded in the EVI200 context, since the CE-VID belongs to the defined VLAN-bundle on the port. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI200 context and if CE2-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE2): (1) a forwarding state is added for CE2-MAC associated to the corresponding LAG/ESI, (2) the ARP-request is snooped and the tuple CE2-MAC/CE2-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE2 containing the EVI200 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 EVI). Note that since PE3 is not part of ESI12, it will install a forwarding state for CE2-MAC as long as the A-D route per ESI for ESI12 is 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 learnt CE2-MAC locally attached to ESI12. Otherwise it will add a forwarding state for CE2-MAC.

c) Assuming PE2 does not have the ARP information for CE3-IP yet, and since the ARP is a broadcast packet and PE2 the non-DF for EVI200 on ESI12, the packet is forwarded by PE2 in the Inclusive multicast tree for EVI200, 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 if the result of the CE2 hashing had been different and the packet sent to PE1, PE1 would not have added the ESI label to the label stack (PE1 is the DF for EVI200 on ESI12).

d) The MPLS-encapsulated packet gets to PE1 and PE3. PE1 de-encapsulate the Inclusive multicast tree label(s) and based on the ESI label at the bottom of the stack, it decides to not forward the packet to the ESI12. It will pop the ESI label and will replicate it to CE1 though, since CE1 is not part of the ESI identified by the ESI label. At PE3, the Inclusive multicast tree label(s) are popped and the packet forwarded to CE3. If a P2MP LSP is used as Inclusive multicast tree for EVI200, PE3 will find an ESI label after popping the P2MP LSP label. The ESI label will simply be ignored and popped, since CE3 is not part of ESI12.

(3) Unicast packet example from CE3 to CE1:

a) A unicast packet, tagged with any CE-VID is issued from source MAC CE3-MAC and destination MAC CE1-MAC (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 ARP proxy table).

b) The packet is identified to be forwarded in the EVI200 context, since the CE-VID belongs to the defined VLAN-bundle on the port. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI200 context and, this time, since we assume CE3-MAC and CE3-IP are 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 to the MAC CE1-MAC is found (this includes the label associated to EVI200 in PE1 and the P2P LSP label to get to PE1). The unicast packet is then encapsulated and forwarded to PE1. The CE-VID is kept.

c) At PE1, the packet is identified to be part of EVI200 (based on the bottom label) and a destination MAC lookup is performed in the EVI200 context. The labels are popped and the packet forwarded to CE1. The CE-VID remains untouched throughout the whole process.

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

(4) Unicast packet example from CE3 to CE2:

a) A unicast packet, tagged with any CE-VID, is issued from source MAC CE3-MAC and destination MAC CE2-MAC (PE3 has previously resolved an ARP request from CE3 to find the MAC of CE2-IP).

b) The packet is identified to be forwarded in the EVI200 context, since the CE-VID belongs to the defined VLAN-bundle on the ingress port. A source MAC lookup is done in the MAC FIB within the EVI200 context and 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 PE3 finds CE2-MAC associated to 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 EVI200 (from PE1 and PE2). Based on a hashing function for the packet, PE3 may decide to forward the packet using the label stack associated to PE2 (label received from the MAC advertisement route) or the label stack associated to PE1 (label received from the A-D route per EVI for EVI200). Either way, the packet is encapsulated and sent to the remote PE.

c) At PE2 (or PE1), the packet is identified to be part of EVI200 based on the bottom label, and a destination MAC lookup is performed at the MAC FIB. In particular, if the packet arrives to PE2, the bottom label is assumed to be a label per EVI, hence a MAC lookup for the EVI200 context is done. If the packet arrives to PE1, the bottom label is assumed to be a label identifying ESI12, hence the packet is forwarded to ESI12.

Unicast packets from CE1 to CE2 follow the same procedures. Aliasing is possible in this case too, since ESI12 is local to PE1 and load balancing through PE1 and PE2 may happen.

5.4. VLAN-aware bundling service procedures

The last potential service type analyzed in this document is VLAN-aware bundling. When these types of service interfaces are used to carry the 4k CE-VIDs among CE1, CE2 and CE3, all the CE-VIDs will be mapped to the same EVI, e.g. 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.

5.4.1. Service startup procedures

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

o 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 setup. Note that ingress replication, P2MP LSPs or MP2MP 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.

o 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 multi-homing 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 EVI300 RT locally configured. The A-D routes per ESI will be used for fast convergence and split-horizon functions, as described in [E-VPN].

o Ethernet A-D routes per EVI (one route): An A-D route (EVI300) will be sent by PE1 and PE2 for ESI12. 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.

5.4.2. Packet Walkthrough

The packet walkthrough for the VLAN-aware case is similar to the ones 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. Note that this model requires qualified learning on the MAC FIBs.

(1) BUM packet example from CE1:

a) An ARP-request tagged with CE-VID=x is issued from source MAC CE1-MAC to find the MAC address of CE3-IP.

b) The packet is identified to be forwarded in the EVI300 context as long as its CE-VID belongs to the range defined in the PE1 port to CE1. In addition to it, CE-VID=x is mapped to Ethernet-Tag=y at the EVI300 (where x and y might be equal if no translation is needed). A source MAC lookup is done next, in the MAC FIB and ARP proxy table within the EVI300/Ethernet-Tag=y context and if CE1-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE1): (1) a forwarding state is added for CE1-MAC associated to the corresponding port and Ethernet-Tag, (2) the ARP-request is snooped and the tuple CE1-MAC/CE1-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE1 containing the EVI300 RD and RT, ESI=0, Ethernet-Tag=y and CE1-MAC/CE1-IP along with an MPLS label assigned from the PE1 label space. Since we assume a MAC forwarding model, a label per EVI 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 ARP entry are added to their EVI300/Ethernet-Tag=y context. From this moment on, any ARP request from CE2 or CE3 destined to CE1-IP, can be directly replied by PE1, PE2 or PE3 and ARP flooding is not needed in the core.

c) Since the ARP is a broadcast packet, it is forwarded by PE1 using the Inclusive multicast tree for EVI300/Ethernet-Tag=y. Note that the ingress CE-VID=x MUST be translated to the Ethernet-Tag=y at the imposition PE, assuming x and y are not equal. Depending on the type of tree, the label stack may vary. E.g. assuming ingress replication, the packet is replicated to PE2 and PE3 with the downstream allocated labels (by PE2 and PE3 respectively) and the P2P LSP transport labels. No other labels are added to the stack.

d) Assuming PE1 is the DF for EVI300 on ESI12, the packet is locally replicated to CE2. Note that the Ethernet-Tag MUST be translated to the egress CE-VID (if they are different).

e) The MPLS-encapsulated packet gets to PE2 and PE3. Since PE2 is non-DF for EVI300 on ESI12 and there are no other CEs connected, the packet is discarded. At PE3, the packet is de-encapsulated and replicated to CE3. The Ethernet-Tag in the packet is translated to the egress CE-VID (if different).

Any other type of BUM packet from CE1 would follow the same procedures. BUM packets from CE3 would follow the same procedures too.

(2) BUM packet example from CE2:

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

b) CE2 will hash the packet and will forward the packet to e.g. PE2. The packet CE-VID=x is identified to be forwarded in the EVI300/Ethernet-Tag=y context, since the CE-VID belongs to the defined range on the port/Ethernet-Tag. A source MAC lookup is done in the MAC FIB and ARP proxy table within the EVI300/Ethernet-Tag=y context and if CE2-MAC is unknown, three actions are carried out (assuming the source MAC is accepted by PE2): (1) a forwarding state is added for CE2-MAC associated to the corresponding LAG/ESI and Ethernet-Tag, (2) the ARP-request is snooped and the tuple CE2-MAC/CE2-IP is added to the ARP proxy table and (3) a BGP MAC advertisement route is triggered from PE2 containing the EVI300 RD and RT, ESI=12, Ethernet-Tag=y and CE2-MAC/CE2-IP along with an MPLS label assigned from the PE2 label space (one label per EVI). Note that since PE3 is not part

of ESI12, it will install a forwarding state for CE2-MAC in the EVI300/Ethernet-Tag=y context as long as the A-D route per ESI for ESI12 is 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 learnt CE2-MAC locally attached to ESI12. Otherwise it will add a forwarding state for CE2-MAC.

c) Assuming PE2 does not have the ARP information for CE3-IP yet, and since the ARP is a broadcast packet and PE2 the non-DF for EVI300 on ESI12, the packet is forwarded by PE2 in the Inclusive multicast tree for EVI300/Ethernet-Tag=y, 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 if the result of the CE2 hashing had been different and the packet sent to PE1, PE1 would not have added the ESI label to the label stack (PE1 is the DF for EVI300 on ESI12).

d) The MPLS-encapsulated packet gets to PE1 and PE3. PE1 de-encapsulate the Inclusive multicast tree label(s) and based on the ESI label at the bottom of the stack, it decides to not forward the packet to the ESI12. It will pop the ESI label and will replicate it to CE1 though, since CE1 is not part of the ESI identified by the ESI label. The Ethernet-Tag will be translated, if needed, to the egress CE-VID. At PE3, the Inclusive multicast tree label(s) are popped and the packet forwarded to CE3 after translating the Ethernet-Tag to the egress CE-VID. If a P2MP LSP is used as Inclusive multicast tree for EVI300/Ethernet-Tag=y, PE3 will find an ESI label after popping the P2MP LSP label. The ESI label will be simply ignored and popped, since CE3 is not part of ESI12.

(3) Unicast packet example from CE3 to CE1:

a) A unicast packet, tagged with CE-VID=x is issued from source MAC CE3-MAC and destination MAC CE1-MAC (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 ARP proxy table).

b) The packet is identified to be forwarded in the EVI300/Ethernet-Tag=y context, since the CE-VID belongs to the defined range on the port/Ethernet-Tag. A source MAC lookup is done in the MAC FIB within the EVI300/Ethernet-Tag=y 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 to the MAC CE1-MAC is found (this includes the label associated to EVI300/Ethernet-Tag=y in PE1 and the P2P LSP label to get to PE1). The unicast packet is then encapsulated and forwarded to PE1. The CE-VID=x is translated

to the Ethernet-Tag=y value.

c) At PE1, the packet is identified to be part of EVI300 (based on the bottom of the stack label) and a destination MAC lookup is performed in the EVI300/Ethernet-Tag=y context. The labels are popped and the packet forwarded to CE1 after translating the Ethernet-Tag value to the egress CE-VID.

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

(4) Unicast packet example from CE3 to CE2:

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

b) The packet is identified to be forwarded in the EVI300/Ethernet-Tag=y context, since the CE-VID belongs to the defined range on the ingress port/Ethernet-Tag. A source MAC lookup is done in the MAC FIB table within the EVI300/Ethernet-Tag=y context and 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 PE3 finds CE2-MAC associated to PE2 on ESI12/Ethernet-Tag=y, 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 EVI300 (from PE1 and PE2). Based on a hashing function for the packet, PE3 may decide to forward the packet using the label stack associated to PE2 (label received from the MAC advertisement route) or the label stack associated to PE1 (label received from the A-D route per EVI for EVI300). Either way, the packet is encapsulated, CE-VID translated to Ethernet-Tag and sent to the remote PE.

c) At PE2 (or PE1), the packet is identified to be part of EVI300/Ethernet-Tag=y based on the bottom label and the packet Ethernet-Tag, and a destination MAC lookup is performed at the MAC FIB. In particular, if the packet arrives to PE2, the bottom label is assumed to be a label per EVI and the Ethernet-Tag=y, hence a MAC lookup for the EVI300/Ethernet-Tag=y context is done. If the packet arrives to PE1, the bottom label is assumed to be a label identifying ESI12 and the packet Ethernet-Tag the pointer at the egress CE-VID, hence the packet is forwarded to ESI12, with a translated tag from the Ethernet-Tag=y to the egress CE-VID=x.

Unicast packets from CE1 to CE2 follow the same procedures. Aliasing is possible in this case too, since ESI12 is local to PE1 and load balancing through PE1 and PE2 may happen.

6. MPLS-based forwarding model use-case

E-VPN supports an alternative forwarding model, usually referred to as MPLS-based forwarding or disposition model as opposed to the MAC-based forwarding or disposition model described in section 5. Using MPLS-based forwarding model instead of the MAC-based one might have an impact on:

- The number of forwarding states required
- The FIB where the forwarding states are handled: MAC FIB or MPLS 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 EVI. 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 might be necessary or not depending on the available next-hop forwarding states in the LFIB.

The following sub-sections highlight the impact on the control and data plane procedures described in section 5 when and 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 decision local to the PE node.

6.1. Impact of MPLS-based forwarding on the E-VPN network startup

The MPLS-based forwarding model has no impact on the procedures explained in section 5.1.

6.2. 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 number of routes, when all the service interfaces are VLAN-based. 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): no impact compared to the MAC-based model.
- Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact

compared to the MAC-based model.

- o Ethernet A-D routes per EVI (4k routes per PE/ESI): no impact compared to the MAC-based model.

- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same EVI, 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. E.g. 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. CE1-MAC1 and CE1-MAC2 (MACs being sent from CE1) would however require a different MPLS label each, label=y and label=z, even if they belong to the same EVI as CE2-MAC1/MAC2. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment.

- o 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.

6.3. 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:

- o Flooding tree setup per EVI (one route): no impact compared to the MAC-based model.

- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact compared to the MAC-based model.

- o Ethernet A-D routes per EVI (one route per PE/ESI): no impact compared to the MAC-based model since no VLAN translation is required.

- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same EVI, 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. E.g. 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. CE1-MAC1 and CE1-MAC2 (MACs being sent from CE1) would however require a different MPLS label each, label=y and label=z, even if they belong to the same EVI as CE2-MAC1/MAC2. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment.

- o 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.

6.4. Impact of MPLS-based forwarding on the VLAN-aware service procedures

Compared to the MAC-based forwarding model, the MPLS-based forwarding model has definitively an impact in terms of number of A-D routes when all the service interfaces are VLAN-aware bundle type. The differences for the use-case described in this document are summarized in the following list:

- o Flooding tree setup per EVI (4k routes per PE): no impact compared to the MAC-based model.

- o Ethernet A-D routes per ESI (one route for ESI12 per PE): no impact compared to the MAC-based model.

- o Ethernet A-D routes per EVI (4k routes per PE/ESI): PE1 and PE2 will send 4k routes for EVI300, one per <ESI, Ethernet-Tag ID> tuple. This will allow 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. The MAC-based forwarding model would allow the PEs to send a single route per PE/ESI for EVI300, since the packet with the embedded Ethernet-Tag would be used to perform a MAC lookup and find out the egress CE-VID.

- o MAC-advertisement routes: instead of allocating and advertising the same MPLS label for all the new MACs locally learnt on the same EVI, 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. E.g. 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. CE1-MAC1 and CE1-MAC2 (MACs being sent from CE1) would however require a different MPLS label each, label=y and label=z, even if they belong to the same EVI as CE2-MAC1/MAC2. It is up to the PE1 implementation to use a different label per individual MAC within the same ES Segment. Note that, in this model, the Ethernet-Tag will be set to a non-zero value for the MAC-advertisement routes. The same MAC address can be announced with different Ethernet-Tag value. This will make the advertising PE install two different forwarding states in the MPLS LFIB.

- o 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. Comparison between MAC-based and MPLS-based forwarding models

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

The MAC-based forwarding model can save A-D routes per EVI when VLAN-aware bundling services are deployed and therefore reduce the control plane overhead. 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. An A-D route per <EVI, Ethernet-Tag> is required for VLAN-aware services, as opposed to an A-D route per EVI.

The following table summarizes the implementation details of both models for the VLAN-aware bundling service type.

4k CE-VID VLANs	MAC-based Model	MPLS-based Model
A-D routes/EVI	1 per ESI/EVI	4k per ESI/EVI
Egress PE Forwarding states	1 per MAC	1 per next-hop
Egress PE Lookups	2 (MPLS+MAC)	1 (MPLS)

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.

8. Traffic flow optimization

In addition to the procedures described across sections 1 through 7, E-VPN [E-VPN] 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, E-VPN procedures allow for intelligent, closest to the source, inter-subnet forwarding and solves the commonly known sub-optimal 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 the following sections.

8.1. Control Plane Procedures

8.1.1. MAC learning options

The fundamental premise of [E-VPN] is the notion of a different approach to MAC address learning compared to traditional IEEE 802.1 bridge learning methods; specifically E-VPN 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 L2 FIB is populated with the appropriate MAC addresses.

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

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

- Remote learning defines the procedures used for learning MAC addresses of network elements remotely connected to EVI, i.e. far-end PEs. Remote learning procedures defined in [E-VPN] advocate using only control plane learning; specifically BGP. Through the

use of BGP E-VPN NLRIs, the remote PE has the capability of advertising all the MAC addresses present in its local FIB.

8.1.2. Proxy ARP

In E-VPN, MAC addresses are advertised via the MAC Advertisement Route, as discussed in [E-VPN]. Optionally an IP address can be advertised along with the MAC address announcement. However, there are certain rules put in place in terms of IP address usage: if the MAC Advertisement Route contains an IP address, and the IP Address Length is 32 bits (or 128 in the IPv6 case), this particular IP address correlates directly with the advertised MAC address. Such advertisement allows us to build a Proxy ARP table populated with the IP<>MAC bindings received from all the remote nodes.

Furthermore, based on these bindings, a local EVI can now provide Proxy-ARP functionality for all ARP requests directed to the IP address pool learned through BGP. Therefore, the amount of unnecessary L2 flooding, ARP requests in this case, can be further reduced by the introduction of Proxy-ARP functionality across all E-VPN EVIs.

8.1.3. 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 the remote EVI network will somehow signal its presence to the network. This signaling can be either done through gratuitous events. Once the remote PE acknowledges the presence of the node in the EVI, it will do two things: install its MAC address in its local FIB and advertise this MAC address to all other BGP speakers via E-VPN NLRI. Therefore, we can assume that any active MAC address is propagated and learnt through the entire E-VPN domain. Given that MAC addresses become pre-populated - 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 unless control or management plane learning is used in all the PEs for a given EVI, unknown unicast flooding MUST be enabled.

8.1.4. 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 E-VPN and IP-VPN is required. A common way of stitching the two service planes is through the use of an 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 of 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 of the IRB, the EVI follows standard bridging procedures.

A typical example of E-VPN 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 trough the IRB interface.

In addition to optimizing the traffic patterns in the network, local inter-subnet forwarding also optimizes greatly the amount of processing needed to cross the subnets: standard VPLS to IP-VPN stitching through IRB interfaces forces the traffic to pass through IRB interfaces twice, once locally, as the traffic gets into the routing domain for a given IP VPN, and once remotely as the traffic exits the routing domain and enters the remote VPLS instance at the egress PE.

Through E-VPN 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 EVI, bypassing any egress IP-VPN processing.

8.2. Packet Walkthrough Examples

Assuming that the services are setup according to figure 1 in section 2, the following flow optimization processes will take place in terms of creating, receiving and forwarding packets across the network.

8.2.1. Proxy-ARP example for CE2 to CE3 traffic

Using figure 1 in section 2, consider EVI 400 residing on PE1, PE2 and PE3 connecting CE2 and CE3 networks. Also, consider that PE1 and PE2 are part of the all-active multi-homing 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 EVI 400 context.

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

From the forwarding perspective, when a node behind CE2 sends a packet destined to a node behind CE3, it will first send an ARP request to e.g. 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.

8.2.2. Flood suppression example for CE1 to CE3 traffic

Using figure 1 in section 2, consider EVI 500 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 E-VPN Type 2 MAC Advertisement Routes and advertise all locally learned MAC addresses. The routes will also include the MPLS label associated with the corresponding egress EVI or egress next-hop, depending on the forwarding model scheme being used by PE3.

Given that PE1 automatically learns and installs all MAC addresses behind CE3, its EVI FIB will already be pre-populated 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 EVI 500, 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 EVI 500 context, or will just forward the traffic directly to the CE3. In the case that PE1 EVI 500 does not have a MAC entry for a specific destination that CE1 is trying to reach, PE1 will drop the packet since unknown unicast flooding is disabled.

Based on the assumption that all the MAC entries behind the CEs are

pre-populated through gratuitous-ARP and/or DHCP requests, if one specific MAC entry is not present in the EVI 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 be flooded and consume network bandwidth.

8.2.3. Optimization of inter-subnet forwarding example for CE3 to CE2 traffic

Using figure 1 in section 2 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, EVI 600 and EVI 60. Each IP subnet is associated to a different E-VPN context. Thus there is a single subnet, subnet 600, between CE1 and CE3 that is established through EVI 600. Similarly, there is another subnet, subnet 60, between CE2 and CE3 that is established through EVI 60. 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 E-VPN procedures, i.e. FIB lookup within the PE1 (or PE2) after adding the corresponding E-VPN label to the MPLS label stack (downstream label allocation from PE3 for EVI 60).

When it comes to crossing the subnet boundaries, the ingress PE implements local inter-subnet forwarding. For example, when a node behind CE2 (EVI 60) sends a packet to a node behind CE1 (EVI 600) the destination IP address will be in the subnet 600, but the destination MAC address will be the address of source node's default gateway, which in this case will be an IRB interface on PE1 (connecting EVI 60 to IP-VPN 666). Once PE1 sees the traffic destined to its own MAC address, it will route the packet to EVI 600, i.e. it will change the source MAC address to the one of the IRB interface in EVI 600 and change the destination MAC address to the address belonging to the node behind CE1, which is already populated in the EVI 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 a typical IP-VPN-to-VPLS scenario, once the packet leaves the L2 domain on PE1, it would be routed through the IP-VPN procedures and consequently, through a remote PE3 IRB interface, routed back into the remote VPLS domain for further processing. However, in the E-VPN case, traffic locally routed and

forwarded to the egress PE within the E-VPN EVI context.

In our use-case, therefore, when node from subnet 60 behind CE2 sends traffic to the node on subnet 600 behind CE3, the destination MAC address is the PE1 EVI 60 IRB MAC address. However, once the traffic locally crosses EVIs, to EVI 600, 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 E-VPN and already installed in EVI 600. The rest of the forwarding through PE1 is using the EVI 600 forwarding context and label space.

Another very relevant optimization is due to the fact that traffic between PEs is forwarded through E-VPN, rather than through IP-VPN. In the example described above for traffic from EVI 60 on CE2 to EVI 600 on CE3, there is no need for IP-VPN processing on the egress PE3. Traffic is forwarded either to the EVI 600 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.

9. Conventions used in this document

In the examples, the following conventions are used:

- CE-VIDs refer to the VLAN tag identifiers being used at CE1, CE2 and CE3 to tag customer traffic sent to the Service Provider E-VPN network

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

- CE1-IP, CE2-IP and CE3-IP refer to IP addresses associated to the above MAC addresses.

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

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

10. Security Considerations

11. IANA Considerations

12. References

12.1. Normative References

[RFC4761]Kompella, K., Ed., and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762]Lasserre, M., Ed., and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC6074]Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.

[RFC4364]Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

12.2. Informative References

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-03.txt, work in progress, February, 2013

[EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-02.txt

[VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-l2vpn-vpls-mcast-13.txt

13. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

14. Authors' Addresses

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

Senad Palislamovic
Alcatel-Lucent
Email: senad.palislamovic@alcatel-lucent.com

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

Florin Balus
Alcatel-Lucent
Email: Florin.Balus@alcatel-lucent.com

Keyur Patel
Cisco
Email: keyupate@cisco.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

James Uttaro
AT&T
Email: uttaro@att.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

Truman Boyes
Bloomberg
Email: tboyes@bloomberg.net