

BESS Workgroup  
Internet Draft  
Intended status: Standards Track

J. Rabadan (Ed.)  
S. Sathappan  
W. Henderickx  
Nokia

A. Sajassi  
Cisco

J. Drake  
Juniper

Expires: January 19, 2018

July 18, 2017

**Interconnect Solution for EVPN Overlay networks**  
**draft-ietf-bess-dci-evpn-overlay-05**

Abstract

This document describes how Network Virtualization Overlays (NVO) can be connected to a Wide Area Network (WAN) in order to extend the layer-2 connectivity required for some tenants. The solution analyzes the interaction between NVO networks running EVPN and other L2VPN technologies used in the WAN, such as VPLS/PBB-VPLS or EVPN/PBB-EVPN, and proposes a solution for the interworking between both.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/1id-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at

<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on January 19, 2018.

## Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

<a href="#">1. Introduction</a>	<a href="#">3</a>
<a href="#">2. Decoupled Interconnect solution for EVPN overlay networks</a>	<a href="#">3</a>
<a href="#">2.1. Interconnect requirements</a>	<a href="#">4</a>
<a href="#">2.2. VLAN-based hand-off</a>	<a href="#">5</a>
<a href="#">2.3. PW-based (Pseudowire-based) hand-off</a>	<a href="#">5</a>
<a href="#">2.4. Multi-homing solution on the GWs</a>	<a href="#">6</a>
<a href="#">2.5. Gateway Optimizations</a>	<a href="#">6</a>
<a href="#">2.5.1. MAC Address Advertisement Control</a>	<a href="#">6</a>
<a href="#">2.5.2. ARP flooding control</a>	<a href="#">7</a>
<a href="#">2.5.3. Handling failures between GW and WAN Edge routers</a>	<a href="#">7</a>
<a href="#">3. Integrated Interconnect solution for EVPN overlay networks</a>	<a href="#">8</a>
<a href="#">3.1. Interconnect requirements</a>	<a href="#">8</a>
<a href="#">3.2. VPLS Interconnect for EVPN-Overlay networks</a>	<a href="#">9</a>
<a href="#">3.2.1. Control/Data Plane setup procedures on the GWs</a>	<a href="#">9</a>
<a href="#">3.2.2. Multi-homing procedures on the GWs</a>	<a href="#">10</a>
<a href="#">3.3. PBB-VPLS Interconnect for EVPN-Overlay networks</a>	<a href="#">10</a>
<a href="#">3.3.1. Control/Data Plane setup procedures on the GWs</a>	<a href="#">10</a>
<a href="#">3.3.2. Multi-homing procedures on the GWs</a>	<a href="#">11</a>
<a href="#">3.4. EVPN-MPLS Interconnect for EVPN-Overlay networks</a>	<a href="#">11</a>
<a href="#">3.4.1. Control Plane setup procedures on the GWs</a>	<a href="#">11</a>
<a href="#">3.4.2. Data Plane setup procedures on the GWs</a>	<a href="#">13</a>
<a href="#">3.4.3. Multi-homing procedure extensions on the GWs</a>	<a href="#">14</a>
<a href="#">3.4.4. Impact on MAC Mobility procedures</a>	<a href="#">16</a>
<a href="#">3.4.5. Gateway optimizations</a>	<a href="#">16</a>
<a href="#">3.4.6. Benefits of the EVPN-MPLS Interconnect solution</a>	<a href="#">17</a>
<a href="#">3.5. PBB-EVPN Interconnect for EVPN-Overlay networks</a>	<a href="#">18</a>



<a href="#">3.5.1.</a>	Control/Data Plane setup procedures on the GWs . . . . .	<a href="#">18</a>
<a href="#">3.5.2.</a>	Multi-homing procedures on the GWs . . . . .	<a href="#">18</a>
<a href="#">3.5.3.</a>	Impact on MAC Mobility procedures . . . . .	<a href="#">18</a>
<a href="#">3.5.4.</a>	Gateway optimizations . . . . .	<a href="#">19</a>
<a href="#">3.6.</a>	EVPN-VXLAN Interconnect for EVPN-Overlay networks . . . . .	<a href="#">19</a>
<a href="#">3.6.1.</a>	Globally unique VNIs in the Interconnect network . . . . .	<a href="#">20</a>
3.6.2.	Downstream assigned VNIs in the Interconnect network . . . . .	20
<a href="#">5.</a>	Conventions and Terminology . . . . .	<a href="#">20</a>
<a href="#">6.</a>	Security Considerations . . . . .	<a href="#">21</a>
<a href="#">7.</a>	IANA Considerations . . . . .	<a href="#">21</a>
<a href="#">8.</a>	References . . . . .	<a href="#">21</a>
<a href="#">8.1.</a>	Normative References . . . . .	<a href="#">22</a>
<a href="#">8.2.</a>	Informative References . . . . .	<a href="#">22</a>
<a href="#">9.</a>	Acknowledgments . . . . .	<a href="#">23</a>
<a href="#">10.</a>	Contributors . . . . .	<a href="#">23</a>
<a href="#">11.</a>	Authors' Addresses . . . . .	<a href="#">23</a>

## [1.](#) Introduction

[EVPN-Overlays] discusses the use of EVPN as the control plane for Network Virtualization Overlays (NVO), where VXLAN, NVGRE or MPLS over GRE can be used as possible data plane encapsulation options.

While this model provides a scalable and efficient multi-tenant solution within the Data Center, it might not be easily extended to the WAN in some cases due to the requirements and existing deployed technologies. For instance, a Service Provider might have an already deployed (PBB-)VPLS or (PBB-)EVPN network that has to be used to interconnect Data Centers and WAN VPN users. A Gateway (GW) function is required in these cases.

This document describes a Interconnect solution for EVPN overlay networks, assuming that the NVO Gateway (GW) and the WAN Edge functions can be decoupled in two separate systems or integrated into the same system. The former option will be referred as "Decoupled Interconnect solution" throughout the document, whereas the latter one will be referred as "Integrated Interconnect solution".

## [2.](#) Decoupled Interconnect solution for EVPN overlay networks

This section describes the interconnect solution when the GW and WAN Edge functions are implemented in different systems. Figure 1 depicts the reference model described in this section.



- o A simple connectivity hand-off needs to be provided between the EVPN-Overlay network provider and the WAN provider so that QoS and security enforcements are easily accomplished.
- o The solution has to be independent of the L2VPN technology deployed in the WAN.
- o Multi-homing between GW and WAN Edge routers is required. Per-service load balancing MUST be supported. Per-flow load balancing MAY be supported but it is not a strong requirement since a deterministic path per service is usually required for an easy QoS and security enforcement.
- o Ethernet OAM and Connectivity Fault Management (CFM) functions



needs to be supported between the EVPN-Overlay network and the WAN network.

- o The following optimizations MAY be supported at the GW:
  - + Flooding reduction of unknown unicast traffic sourced from the DC Network Virtualization Edge devices (NVEs).
  - + Control of the WAN MAC addresses advertised to the DC.
  - + ARP flooding control for the requests coming from the WAN.

## **2.2. VLAN-based hand-off**

In this option, the hand-off between the GWs and the WAN Edge routers is based on 802.1Q VLANs. This is illustrated in Figure 1 (between the GWs in NVO-1 and the WAN Edge routers). Each MAC-VRF in the GW is connected to a different VSI/MAC-VRF instance in the WAN Edge router by using a different C-TAG VLAN ID or a different combination of S/C-TAG VLAN IDs that matches at both sides.

This option provides the best possible demarcation between the DC and WAN providers and it does not require control plane interaction between both providers. The disadvantage of this model is the provisioning overhead since the service has to be mapped to a C-TAG or S/C-TAG VLAN ID combination at both GW and WAN Edge routers.

In this model, the GW acts as a regular Network Virtualization Edge (NVE) towards the DC. Its control plane, data plane procedures and interactions are described in [[EVPN-Overlays](#)].

The WAN Edge router acts as a (PBB-)VPLS or (PBB-)EVPN PE with attachment circuits (ACs) to the GWs. Its functions are described in [[RFC4761](#)], [[RFC4762](#)], [[RFC6074](#)] or [[RFC7432](#)], [[RFC7623](#)].

## **2.3. PW-based (Pseudowire-based) hand-off**

If MPLS can be enabled between the GW and the WAN Edge router, a PW-based Interconnect solution can be deployed. In this option the hand-off between both routers is based on FEC128-based PWs or FEC129-based PWs (for a greater level of network automation). Note that this model still provides a clear demarcation boundary between DC and WAN (since there is a single PW between each MAC-VRF and peer VSI), and security/QoS policies may be applied on a per PW basis. This model provides better scalability than a C-TAG based hand-off and less provisioning overhead than a combined C/S-TAG hand-off. The PW-based hand-off interconnect is illustrated in Figure 1 (between the NVO-2 GWs and the WAN Edge routers).

In this model, besides the usual MPLS procedures between GW and WAN Edge router, the GW MUST support an interworking function in each





MAC-VRF that requires extension to the WAN:

- o If a FEC128-based PW is used between the MAC-VRF (GW) and the VSI (WAN Edge), the provisioning of the VCID for such PW MUST be supported on the MAC-VRF and MUST match the VCID used in the peer VSI at the WAN Edge router.
- o If BGP Auto-discovery [[RFC6074](#)] and FEC129-based PWs are used between the GW MAC-VRF and the WAN Edge VSI, the provisioning of the VPLS-ID MUST be supported on the MAC-VRF and MUST match the VPLS-ID used in the WAN Edge VSI.

#### **2.4. Multi-homing solution on the GWs**

As already discussed, single-active multi-homing, i.e. per-service load-balancing multi-homing MUST be supported in this type of interconnect.

The GWs will be provisioned with a unique ESI per WAN interconnect and the hand-off attachment circuits or PWs between the GW and the WAN Edge router will be assigned to such ESI. The ESI will be administratively configured on the GWs according to the procedures in [[RFC7432](#)]. This Interconnect ESI will be referred as "I-ESI" hereafter.

The solution (on the GWs) MUST follow the single-active multi-homing procedures as described in [[EVPN-Overlays](#)] for the provisioned I-ESI, i.e. Ethernet A-D routes per ESI and per EVI will be advertised to the DC NVEs. The MAC addresses learned (in the data plane) on the hand-off links will be advertised with the I-ESI encoded in the ESI field.

#### **2.5. Gateway Optimizations**

The following features MAY be supported on the GW in order to optimize the control plane and data plane in the DC.

##### **2.5.1. MAC Address Advertisement Control**

The use of EVPN in the NVO networks brings a significant number of benefits as described in [[EVPN-Overlays](#)]. However, if multiple DCs are interconnected into a single EVI, each DC will have to import all of the MAC addresses from each of the other DCs.

Even if optimized BGP techniques like RT-constraint are used, the number of MAC addresses to advertise or withdraw (in case of failure) by the GWs of a given DC could overwhelm the NVEs within that DC, particularly when the NVEs reside in the hypervisors.



The solution specified in this document uses the 'Unknown MAC' route which is advertised into a given DC by each of the DC's GWs. This route is a regular EVPN MAC/IP Advertisement route in which the MAC Address Length is set to 48, the MAC address is set to 00:00:00:00:00:00, the IP length is set to 0, and the ESI field is set to the DC GW's I-ESI.

An NVE within that DC that understands the Unknown MAC route will send (unicast) a packet with an unknown unicast MAC address to one of the DCs GWs which will then forward that packet to the correct egress PE. I.e., because the ESI is set to the DC GW's I-ESI, all-active multi-homing can be applied to unknown unicast MAC addresses.

This document proposes that local policy determines whether MAC addresses and/or the Unknown MAC route are advertised into a given DC. As an example, when all the DC MAC addresses are learned in the control/management plane, it may be appropriate to advertise only the Unknown MAC route.

#### **2.5.2. ARP flooding control**

Another optimization mechanism, naturally provided by EVPN in the GWs, is the Proxy ARP/ND function. The GWs SHOULD build a Proxy ARP/ND cache table as per [RFC7432]. When the active GW receives an ARP/ND request/solicitation coming from the WAN, the GW does a Proxy ARP/ND table lookup and replies as long as the information is available in its table.

This mechanism is especially recommended on the GWs since it protects the DC network from external ARP/ND-flooding storms.

#### **2.5.3. Handling failures between GW and WAN Edge routers**

Link/PE failures are handled on the GWs as specified in [RFC7432]. The GW detecting the failure will withdraw the EVPN routes as per [RFC7432].

Individual AC/PW failures MAY be detected by OAM mechanisms. For instance:

- o If the Interconnect solution is based on a VLAN hand-off, 802.1ag/Y.1731 Ethernet-CFM MAY be used to detect individual AC failures on both, the GW and WAN Edge router. An individual AC failure will trigger the withdrawal of the corresponding A-D per EVI route as well as the MACs learned on that AC.
- o If the Interconnect solution is based on a PW hand-off, the LDP PW Status bits TLV MAY be used to detect individual PW failures on



- o The GW function MUST provide control plane and data plane interworking between the EVPN-overlay network and the L2VPN technology supported in the WAN, i.e. (PBB-)VPLS or (PBB-)EVPN, as depicted in Figure 2.
- o Multi-homing MUST be supported. Single-active multi-homing with



per-service load balancing MUST be implemented. All-active multi-homing, i.e. per-flow load-balancing, SHOULD be implemented as long as the technology deployed in the WAN supports it.

- o If EVPN is deployed in the WAN, the MAC Mobility, Static MAC protection and other procedures (e.g. proxy-arp) described in [\[RFC7432\]](#) MUST be supported end-to-end.
- o Any type of inclusive multicast tree MUST be independently supported in the WAN as per [\[RFC7432\]](#), and in the DC as per [EVPN-Overlays].

### **[3.2. VPLS Interconnect for EVPN-Overlay networks](#)**

#### **[3.2.1. Control/Data Plane setup procedures on the GWs](#)**

Regular MPLS tunnels and TLDP/BGP sessions will be setup to the WAN PEs and RRs as per [\[RFC4761\]](#), [\[RFC4762\]](#), [\[RFC6074\]](#) and overlay tunnels and EVPN will be setup as per [\[EVPN-Overlays\]](#). Note that different route-targets for the DC and for the WAN are normally required. A single type-1 RD per service may be used.

In order to support multi-homing, the GWs will be provisioned with an I-ESI (see [section 2.4](#)), that will be unique per interconnection. All the [\[RFC7432\]](#) procedures are still followed for the I-ESI, e.g. any MAC address learned from the WAN will be advertised to the DC with the I-ESI in the ESI field.

A MAC-VRF per EVI will be created in each GW. The MAC-VRF will have two different types of tunnel bindings instantiated in two different split-horizon-groups:

- o VPLS PWs will be instantiated in the "WAN split-horizon-group".
- o Overlay tunnel bindings (e.g. VXLAN, NVGRE) will be instantiated in the "DC split-horizon-group".

Attachment circuits are also supported on the same MAC-VRF, but they will not be part of any of the above split-horizon-groups.

Traffic received in a given split-horizon-group will never be forwarded to a member of the same split-horizon-group.

As far as BUM flooding is concerned, a flooding list will be created with the sub-list created by the inclusive multicast routes and the sub-list created for VPLS in the WAN. BUM frames received from a local attachment circuit will be forwarded to the flooding list. BUM frames received from the DC or the WAN will be forwarded to the





flooding list observing the split-horizon-group rule described above.

Note that the GWs are not allowed to have an EVPN binding and a PW to the same far-end within the same MAC-VRF in order to avoid loops and packet duplication. This is described in [[EVPN-VPLS-INTEGRATION](#)].

The optimizations procedures described in [section 2.5](#) can also be applied to this model.

### **[3.2.2](#). Multi-homing procedures on the GWs**

Single-active multi-homing MUST be supported on the GWs. All-active multi-homing is not supported by VPLS.

All the single-active multi-homing procedures as described by [EVPN-Overlays] will be followed for the I-ESI.

The non-DF GW for the I-ESI will block the transmission and reception of all the bindings in the "WAN split-horizon-group" for BUM and unicast traffic.

## **[3.3](#). PBB-VPLS Interconnect for EVPN-Overlay networks**

### **[3.3.1](#). Control/Data Plane setup procedures on the GWs**

In this case, there is no impact on the procedures described in [[RFC7041](#)] for the B-component. However the I-component instances become EVI instances with EVPN-Overlay bindings and potentially local attachment circuits. A number of MAC-VRF instances can be multiplexed into the same B-component instance. This option provides significant savings in terms of PWs to be maintained in the WAN.

The I-ESI concept described in [section 3.2.1](#) will also be used for the PBB-VPLS-based Interconnect.

B-component PWs and I-component EVPN-overlay bindings established to the same far-end will be compared. The following rules will be observed:

- o Attempts to setup a PW between the two GWs within the B-component context will never be blocked.
- o If a PW exists between two GWs for the B-component and an attempt is made to setup an EVPN binding on an I-component linked to that B-component, the EVPN binding will be kept operationally down. Note that the BGP EVPN routes will still be valid but not used.



- o The EVPN binding will only be up and used as long as there is no PW to the same far-end in the corresponding B-component. The EVPN bindings in the I-components will be brought down before the PW in the B-component is brought up.

The optimizations procedures described in [section 2.5](#) can also be applied to this Interconnect option.

### **[3.3.2](#). Multi-homing procedures on the GWs**

Single-active multi-homing MUST be supported on the GWs. All-active multi-homing is not supported by this scenario.

All the single-active multi-homing procedures as described by [EVPN-Overlays] will be followed for the I-ESI for each EVI instance connected to B-component.

### **[3.4](#). EVPN-MPLS Interconnect for EVPN-Overlay networks**

If EVPN for MPLS tunnels, EVPN-MPLS hereafter, is supported in the WAN, an end-to-end EVPN solution can be deployed. The following sections describe the proposed solution as well as the impact required on the [\[RFC7432\]](#) procedures.

#### **[3.4.1](#). Control Plane setup procedures on the GWs**

The GWs MUST establish separate BGP sessions for sending/receiving EVPN routes to/from the DC and to/from the WAN. Normally each GW will setup one (two) BGP EVPN session(s) to the DC RR(s) and one(two) session(s) to the WAN RR(s).

In order to facilitate separate BGP processes for DC and WAN, EVPN routes sent to the WAN SHOULD carry a different route-distinguisher (RD) than the EVPN routes sent to the DC. In addition, although reusing the same value is possible, different route-targets are expected to be handled for the same EVI in the WAN and the DC. Note that the EVPN service routes sent to the DC RRs will normally include a [\[RFC5512\]](#) BGP encapsulation extended community with a different tunnel type than the one sent to the WAN RRs.

As in the other discussed options, an I-ESI will be configured on the GWs for multi-homing. This I-ESI represents the WAN to the DC but also the DC to the WAN. Optionally, different I-ESI values MAY be configured for representing the WAN and the DC. If different EVPN-Overlay networks are connected to the same group of GWs, each EVPN-Overlay network MUST get assigned a different I-ESI.

Received EVPN routes will never be reflected on the GWs but consumed



and re-advertised (if needed):

- o Ethernet A-D routes, ES routes and Inclusive Multicast routes are consumed by the GWs and processed locally for the corresponding [\[RFC7432\]](#) procedures.
- o MAC/IP advertisement routes will be received, imported and if they become active in the MAC-VRF, the information will be re-advertised as new routes with the following fields:
  - + The RD will be the GW's RD for the MAC-VRF.
  - + The ESI will be set to the I-ESI.
  - + The Ethernet-tag value will be kept from the received NLRI.
  - + The MAC length, MAC address, IP Length and IP address values will be kept from the received NLRI.
  - + The MPLS label will be a local 20-bit value (when sent to the WAN) or a DC-global 24-bit value (when sent to the DC).
  - + The appropriate Route-Targets (RTs) and [\[RFC5512\]](#) BGP Encapsulation extended community will be used according to [\[EVPN-Overlays\]](#).

The GWs will also generate the following local EVPN routes that will be sent to the DC and WAN, with their corresponding RTs and [\[RFC5512\]](#) BGP Encapsulation extended community values:

- o ES route(s) for the I-ESI(s).
- o Ethernet A-D routes per ESI and EVI for the I-ESI(s). The A-D per-EVI routes sent to the WAN and the DC will have consistent Ethernet-Tag values.
- o Inclusive Multicast routes with independent tunnel type value for the WAN and DC. E.g. a P2MP LSP may be used in the WAN whereas ingress replication may be used in the DC. The routes sent to the WAN and the DC will have a consistent Ethernet-Tag.
- o MAC/IP advertisement routes for MAC addresses learned in local attachment circuits. Note that these routes will not include the I-ESI, but ESI=0 or different from 0 for local multi-homed Ethernet Segments (ES). The routes sent to the WAN and the DC will have a consistent Ethernet-Tag.

Assuming GW1 and GW2 are peer GWs of the same DC, each GW will



generate two sets of local service routes: Set-DC will be sent to the DC RRs and will include A-D per EVI, Inclusive Multicast and MAC/IP routes for the DC encapsulation and RT. Set-WAN will be sent to the WAN RRs and will include the same routes but using the WAN RT and encapsulation. GW1 and GW2 will receive each other's set-DC and set-WAN. This is the expected behavior on GW1 and GW2 for locally generated routes:

- o Inclusive multicast routes: when setting up the flooding lists for a given MAC-VRF, each GW will include its DC peer GW only in the EVPN-MPLS flooding list (by default) and not the EVPN-Overlay flooding list. That is, GW2 will import two Inclusive Multicast routes from GW1 (from set-DC and set-WAN) but will only consider one of the two, having the set-WAN route higher priority. An administrative option MAY change this preference so that the set-DC route is selected first.
- o MAC/IP advertisement routes for local attachment circuits: as above, the GW will select only one, having the route from the set-WAN a higher priority. As for the Inclusive multicast routes, an administrative option MAY change this priority.

Note that, irrespective of the encapsulation, EVPN routes always have higher priority than VPLS AD routes as per [[EVPN-VPLS-INTEGRATION](#)].

### **3.4.2. Data Plane setup procedures on the GWs**

The procedure explained at the end of the previous section will make sure there are no loops or packet duplication between the GWs of the same EVPN-Overlay network (for frames generated from local ACs) since only one EVPN binding per EVI (or per Ethernet Tag in case of VLAN-aware bundle services) will be setup in the data plane between the two nodes. That binding will by default be added to the EVPN-MPLS flooding list.

As for the rest of the EVPN tunnel bindings, they will be added to one of the two flooding lists that each GW sets up for the same MAC-VRF:

- o EVPN-overlay flooding list (composed of bindings to the remote NVEs or multicast tunnel to the NVEs).
- o EVPN-MPLS flooding list (composed of MP2P or LSM tunnel to the remote PEs)

Each flooding list will be part of a separate split-horizon-group: the WAN split-horizon-group or the DC split-horizon-group. Traffic generated from a local AC can be flooded to both





split-horizon-groups. Traffic from a binding of a split-horizon-group can be flooded to the other split-horizon-group and local ACs, but never to a member of its own split-horizon-group.

When either GW1 or GW2 receive a BUM frame on an MPLS tunnel including an ESI label at the bottom of the stack, they will perform an ESI label lookup and split-horizon filtering as per [\[RFC7432\]](#) in case the ESI label identifies a local ESI (I-ESI or any other non-zero ESI).

### **[3.4.3](#). Multi-homing procedure extensions on the GWs**

Single-active as well as all-active multi-homing MUST be supported.

All the [\[RFC7432\]](#) multi-homing procedures for the DF election on I-ESI(s) as well as the backup-path (single-active) and aliasing (all-active) procedures will be followed on the GWs. Remote PEs in the EVPN-MPLS network will follow regular [\[RFC7432\]](#) aliasing or backup-path procedures for MAC/IP routes received from the GWs for the same I-ESI. So will NVEs in the EVPN-Overlay network for MAC/IP routes received with the same I-ESI.

As far as the forwarding plane is concerned, by default, the EVPN-Overlay network will have an analogous behavior to the access ACs in [\[RFC7432\]](#) multi-homed Ethernet Segments.

The forwarding behavior on the GWs is described below:

- o Single-active multi-homing; assuming a WAN split-horizon-group (comprised of EVPN-MPLS bindings), a DC split-horizon-group (comprised of EVPN-Overlay bindings) and local ACs on the GWs:
  - + Forwarding behavior on the non-DF: the non-DF MUST block ingress and egress forwarding on the EVPN-Overlay bindings associated to the I-ESI. The EVPN-MPLS network is considered to be the core network and the EVPN-MPLS bindings to the remote PEs and GWs will be active.
  - + Forwarding behavior on the DF: the DF MUST NOT forward BUM or unicast traffic received from a given split-horizon-group to a member of his own split-horizon group. Forwarding to other split-horizon-groups and local ACs is allowed (as long as the ACs are not part of an ES for which the node is non-DF). As per [\[RFC7432\]](#) and for split-horizon purposes, when receiving BUM traffic on the EVPN-Overlay bindings associated to an I-ESI, the DF GW SHOULD add the I-ESI label when forwarding to the peer GW over EVPN-MPLS.



- + When receiving EVPN MAC/IP routes from the WAN, the non-DF MUST NOT re-originate the EVPN routes and advertise them to the DC peers. In the same way, EVPN MAC/IP routes received from the DC MUST NOT be advertised to the WAN peers. This is consistent with [\[RFC7432\]](#) and allows the remote PE/NVEs know who the primary GW is, based on the reception of the MAC/IP routes.
- o All-active multi-homing; assuming a WAN split-horizon-group (comprised of EVPN-MPLS bindings), a DC split-horizon-group (comprised of EVPN-Overlay bindings) and local ACs on the GWs:
  - + Forwarding behavior on the non-DF: the non-DF follows the same behavior as the non-DF in the single-active case but only for BUM traffic. Unicast traffic received from a split-horizon-group MUST NOT be forwarded to a member of its own split-horizon-group but can be forwarded normally to the other split-horizon-groups and local ACs. If a known unicast packet is identified as a "flooded" packet, the procedures for BUM traffic MUST be followed.
  - + Forwarding behavior on the DF: the DF follows the same behavior as the DF in the single-active case but only for BUM traffic. Unicast traffic received from a split-horizon-group MUST NOT be forwarded to a member of its own split-horizon-group but can be forwarded normally to the other split-horizon-group and local ACs. If a known unicast packet is identified as a "flooded" packet, the procedures for BUM traffic MUST be followed. As per [\[RFC7432\]](#) and for split-horizon purposes, when receiving BUM traffic on the EVPN-Overlay bindings associated to an I-ESI, the DF GW MUST add the I-ESI label when forwarding to the peer GW over EVPN-MPLS.
  - + Contrary to the single-active multi-homing case, both DF and non-DF re-originate and advertise MAC/IP routes received from the WAN/DC peers, adding the corresponding I-ESI so that the remote PE/NVEs can perform regular aliasing as per [\[RFC7432\]](#).

The example in Figure 3 illustrates the forwarding of BUM traffic originated from an NVE on a pair of all-active multi-homing GWs.



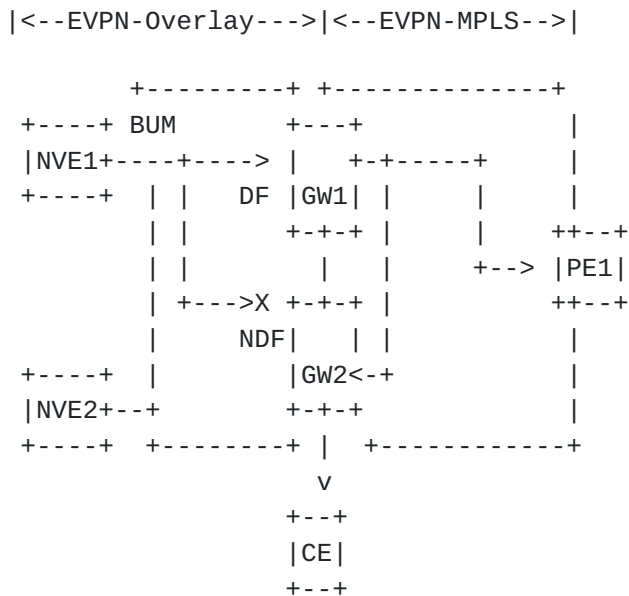


Figure 3 Multi-homing BUM forwarding

GW2 is the non-DF for the I-ESI and blocks the BUM forwarding. GW1 is the DF and forwards the traffic to PE1 and GW2. GW2 will only forward the packets to local ACs (CE in the example).

#### 3.4.4. Impact on MAC Mobility procedures

MAC Mobility procedures described in [\[RFC7432\]](#) are not modified by this document.

Note that an intra-DC MAC move still leaves the MAC attached to the same I-ESI, so under the rules of [\[RFC7432\]](#) this is not considered a MAC mobility event. Only when the MAC moves from the WAN domain to the DC domain (or from one DC to another) the MAC will be learned from a different ES and the MAC Mobility procedures will kick in.

The sticky bit indication in the MAC Mobility extended community MUST be propagated between domains.

#### 3.4.5. Gateway optimizations

All the Gateway optimizations described in [section 2.5](#) MAY be applied to the GWs when the Interconnect is based on EVPN-MPLS.

In particular, the use of the Unknown MAC route, as described in [section 2.5.1](#), solves some transient packet duplication issues in cases of all-active multi-homing, as explained below.

Consider the diagram in Figure 2 for EVPN-MPLS Interconnect and all-



active multi-homing, and the following sequence:

- a) MAC Address M1 is advertised from NVE3 in EVI-1.
- b) GW3 and GW4 learn M1 for EVI-1 and re-advertise M1 to the WAN with I-ESI-2 in the ESI field.
- c) GW1 and GW2 learn M1 and install GW3/GW4 as next-hops following the EVPN aliasing procedures.
- d) Before NVE1 learns M1, a packet arrives at NVE1 with destination M1. If the Unknown MAC route had not been advertised into the DC, NVE1 would have flooded the packet throughout the DC, in particular to both GW1 and GW2. If the same VNI/VSID is used for both known unicast and BUM traffic, as is typically the case, there is no indication in the packet that it is a BUM packet and both GW1 and GW2 would have forwarded it. However, because the Unknown MAC route had been advertised into the DC, NVE1 will unicast the packet to either GW1 or GW2.
- e) Since both GW1 and GW2 know M1, the GW receiving the packet will forward it to either GW3 or GW4.

#### **3.4.6. Benefits of the EVPN-MPLS Interconnect solution**

Besides retaining the EVPN attributes between Data Centers and throughout the WAN, the EVPN-MPLS Interconnect solution on the GWs has some benefits compared to pure BGP EVPN RR or Inter-AS model B solutions without a gateway:

- o The solution supports the connectivity of local attachment circuits on the GWs.
- o Different data plane encapsulations can be supported in the DC and the WAN.
- o Optimized multicast solution, with independent inclusive multicast trees in DC and WAN.
- o MPLS Label aggregation: for the case where MPLS labels are signaled from the NVEs for MAC/IP Advertisement routes, this solution provides label aggregation. A remote PE MAY receive a single label per GW MAC-VRF as opposed to a label per NVE/MAC-VRF connected to the GW MAC-VRF. For instance, in Figure 2, PE would receive only one label for all the routes advertised for a given MAC-VRF from GW1, as opposed to a label per NVE/MAC-VRF.





- o The GW will not propagate MAC mobility for the MACs moving within a DC. Mobility intra-DC is solved by all the NVEs in the DC. The MAC Mobility procedures on the GWs are only required in case of mobility across DCs.
- o Proxy-ARP/ND function on the DC GWs can be leveraged to reduce ARP/ND flooding in the DC or/and in the WAN.

### **3.5. PBB-EVPN Interconnect for EVPN-Overlay networks**

PBB-EVPN [[RFC7623](#)] is yet another Interconnect option. It requires the use of GWs where I-components and associated B-components are part of EVI instances.

#### **3.5.1. Control/Data Plane setup procedures on the GWs**

EVPN will run independently in both components, the I-component MAC-VRF and B-component MAC-VRF. Compared to [[RFC7623](#)], the DC C-MACs are no longer learned in the data plane on the GW but in the control plane through EVPN running on the I-component. Remote C-MACs coming from remote PEs are still learned in the data plane. B-MACs in the B-component will be assigned and advertised following the procedures described in [[RFC7623](#)].

An I-ESI will be configured on the GWs for multi-homing, but it will only be used in the EVPN control plane for the I-component EVI. No non-reserved ESIs will be used in the control plane of the B-component EVI as per [[RFC7623](#)].

The rest of the control plane procedures will follow [[RFC7432](#)] for the I-component EVI and [[RFC7623](#)] for the B-component EVI.

From the data plane perspective, the I-component and B-component EVPN bindings established to the same far-end will be compared and the I-component EVPN-overlay binding will be kept down following the rules described in [section 3.3.1](#).

#### **3.5.2. Multi-homing procedures on the GWs**

Single-active as well as all-active multi-homing MUST be supported.

The forwarding behavior of the DF and non-DF will be changed based on the description outlined in [section 3.4.3](#), only replacing the "WAN split-horizon-group" for the B-component.

#### **3.5.3. Impact on MAC Mobility procedures**

C-MACs learned from the B-component will be advertised in EVPN within



the I-component EVI scope. If the C-MAC was previously known in the I-component database, EVPN would advertise the C-MAC with a higher sequence number, as per [RFC7432]. From a Mobility perspective and the related procedures described in [RFC7432], the C-MACs learned from the B-component are considered local.

#### **3.5.4. Gateway optimizations**

All the considerations explained in [section 3.4.5](#) are applicable to the PBB-EVPN Interconnect option.

#### **3.6. EVPN-VXLAN Interconnect for EVPN-Overlay networks**

If EVPN for Overlay tunnels is supported in the WAN and a GW function is required, an end-to-end EVPN solution can be deployed. This section focuses on the specific case of EVPN for VXLAN (EVPN-VXLAN hereafter) and the impact on the [RFC7432] procedures.

This use-case assumes that NVEs need to use the VNIs or VSIDs as a globally unique identifiers within a data center, and a Gateway needs to be employed at the edge of the data center network to translate the VNI or VSID when crossing the network boundaries. This GW function provides VNI and tunnel IP address translation. The use-case in which local downstream assigned VNIs or VSIDs can be used (like MPLS labels) is described by [EVPN-Overlays].

While VNIs are globally significant within each DC, there are two possibilities in the Interconnect network:

- a) Globally unique VNIs in the Interconnect network:  
In this case, the GWs and PEs in the Interconnect network will agree on a common VNI for a given EVI. The RT to be used in the Interconnect network can be auto-derived from the agreed Interconnect VNI. The VNI used inside each DC MAY be the same as the Interconnect VNI.
- b) Downstream assigned VNIs in the Interconnect network.  
In this case, the GWs and PEs MUST use the proper RTs to import/export the EVPN routes. Note that even if the VNI is downstream assigned in the Interconnect network, and unlike option B, it only identifies the <Ethernet Tag, GW> pair and not the <Ethernet Tag, egress PE> pair. The VNI used inside each DC MAY be the same as the Interconnect VNI. GWs SHOULD support multiple VNI spaces per EVI (one per Interconnect network they are connected to).

In both options, NVEs inside a DC only have to be aware of a single VNI space, and only GWs will handle the complexity of managing



multiple VNI spaces. In addition to VNI translation above, the GWs will provide translation of the tunnel source IP for the packets generated from the NVEs, using their own IP address. GWs will use that IP address as the BGP next-hop in all the EVPN updates to the Interconnect network.

The following sections provide more details about these two options.

### **3.6.1. Globally unique VNIs in the Interconnect network**

Considering Figure 2, if a host H1 in NVO-1 needs to communicate with a host H2 in NVO-2, and assuming that different VNIs are used in each DC for the same EVI, e.g. VNI-10 in NVO-1 and VNI-20 in NVO-2, then the VNIs MUST be translated to a common Interconnect VNI (e.g. VNI-100) on the GWs. Each GW is provisioned with a VNI translation mapping so that it can translate the VNI in the control plane when sending BGP EVPN route updates to the Interconnect network. In other words, GW1 and GW2 MUST be configured to map VNI-10 to VNI-100 in the BGP update messages for H1's MAC route. This mapping is also used to translate the VNI in the data plane in both directions, that is, VNI-10 to VNI-100 when the packet is received from NVO-1 and the reverse mapping from VNI-100 to VNI-10 when the packet is received from the remote NVO-2 network and needs to be forwarded to NVO-1.

The procedures described in [section 3.4](#) will be followed, considering that the VNIs advertised/received by the GWs will be translated accordingly.

### **3.6.2. Downstream assigned VNIs in the Interconnect network**

In this case, if a host H1 in NVO-1 needs to communicate with a host H2 in NVO-2, and assuming that different VNIs are used in each DC for the same EVI, e.g. VNI-10 in NVO-1 and VNI-20 in NVO-2, then the VNIs MUST be translated as in [section 3.6.1](#). However, in this case, there is no need to translate to a common Interconnect VNI on the GWs. Each GW can translate the VNI received in an EVPN update to a locally assigned VNI advertised to the Interconnect network. Each GW can use a different Interconnect VNI, hence this VNI does not need to be agreed on all the GWs and PE of the Interconnect network.

The procedures described in [section 3.4](#) will be followed, taking the considerations above for the VNI translation.

## **5. Conventions and Terminology**

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](#)].



AC: Attachment Circuit

BUM: it refers to the Broadcast, Unknown unicast and Multicast traffic

DF: Designated Forwarder

GW: Gateway or Data Center Gateway

DCI: Data Center Interconnect

ES: Ethernet Segment

ESI: Ethernet Segment Identifier

I-ESI: Interconnect ESI defined on the GWs for multi-homing to/from the WAN

EVI: EVPN Instance

MAC-VRF: it refers to an EVI instance in a particular node

NVE: Network Virtualization Edge

PW: Pseudowire

RD: Route-Distinguisher

RT: Route-Target

TOR: Top-Of-Rack switch

VNI/VSID: refers to VXLAN/NVGRE virtual identifiers

VSI: Virtual Switch Instance or VPLS instance in a particular PE

## **6. Security Considerations**

Security considerations included in [[RFC7432](#)], [[RFC4761](#)] and [[RFC4762](#)] apply to this document.

## **7. IANA Considerations**

## **8. References**





## 8.1. Normative References

[RFC4761]Kompella, K., Ed., and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", [RFC 4761](#), DOI 10.17487/RFC4761, January 2007, <<http://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, <<http://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, <<http://www.rfc-editor.org/info/rfc6074>>.

[RFC7041]Balus, F., Ed., Sajassi, A., Ed., and N. Bitar, Ed., "Extensions to the Virtual Private LAN Service (VPLS) Provider Edge (PE) Model for Provider Backbone Bridging", [RFC 7041](#), DOI 10.17487/RFC7041, November 2013, <<http://www.rfc-editor.org/info/rfc7041>>.

[RFC7432]Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", [RFC 7432](#), DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.

[RFC2119]Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.

[RFC5512]Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", [RFC 5512](#), DOI 10.17487/RFC5512, April 2009, <<http://www.rfc-editor.org/info/rfc5512>>.

[RFC7623] Sajassi et al., "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", [RFC 7623](#), September, 2015, <<http://www.rfc-editor.org/info/rfc7623>>.

## 8.2. Informative References

[EVPN-Overlays] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using EVPN", [draft-ietf-bess-evpn-overlay-08.txt](#), work in progress, March, 2017



[EVPN-VPLS-INTEGRATION] Sajassi et al., "(PBB-)EVPN Seamless Integration with (PBB-)VPLS", [draft-ietf-bess-evpn-vpls-integration-00.txt](#), work in progress, February, 2015

## **9. Acknowledgments**

The authors would like to thank Neil Hart, Vinod Prabhu and Kiran Nagaraj for their valuable comments and feedback.

## **10. Contributors**

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

Ravi Shekhar  
Anil Lohiya  
Wen Lin  
Juniper Networks

Florin Balus  
Patrice Brissette  
Cisco

Senad Palislamovic  
Nokia

Dennis Cai  
Alibaba

## **11. Authors' Addresses**

Jorge Rabadan  
Nokia  
777 E. Middlefield Road  
Mountain View, CA 94043 USA  
Email: [jorge.rabadan@nokia.com](mailto:jorge.rabadan@nokia.com)

Senthil Sathappan  
Nokia  
Email: [senthil.sathappan@nokia.com](mailto:senthil.sathappan@nokia.com)

Wim Henderickx  
Nokia  
Email: [wim.henderickx@nokia.com](mailto:wim.henderickx@nokia.com)

Ali Sajassi



Cisco

Email: [sajassi@cisco.com](mailto:sajassi@cisco.com)

John Drake

Juniper

Email: [jdrake@juniper.net](mailto:jdrake@juniper.net)