

Internet Engineering Task Force (IETF)
Request for Comments: 9014
Category: Standards Track
ISSN: 2070-1721

J. Rabadan, Ed.
S. Sathappan
W. Henderickx
Nokia
A. Sajassi
Cisco
J. Drake
Juniper
May 2021

Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks

Abstract

This document describes how Network Virtualization Overlays (NVOs) 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 Ethernet Virtual Private Networks (EVPNs) and other Layer 2 VPN (L2VPN) technologies used in the WAN, such as Virtual Private LAN Services (VPLSs), VPLS extensions for Provider Backbone Bridging (PBB-VPLS), EVPN, or PBB-EVPN. It also describes how the existing technical specifications apply to the interconnection and extends the EVPN procedures needed in some cases. In particular, this document describes how EVPN routes are processed on Gateways (GWs) that interconnect EVPN-Overlay and EVPN-MPLS networks, as well as the Interconnect Ethernet Segment (I-ES), to provide multihoming. This document also describes the use of the Unknown MAC Route (UMR) to avoid issues of a Media Access Control (MAC) scale on Data Center Network Virtualization Edge (NVE) devices.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at <https://www.rfc-editor.org/info/rfc9014>.

Copyright Notice

Copyright (c) 2021 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 (<https://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

1. Introduction
2. Conventions and Terminology
3. Decoupled Interconnect Solution for EVPN-Overlay Networks
 - 3.1. Interconnect Requirements

- 3.2. VLAN-Based Handoff
- 3.3. PW-Based Handoff
- 3.4. Multihoming Solution on the GWs
- 3.5. Gateway Optimizations
 - 3.5.1. MAC Address Advertisement Control
 - 3.5.2. ARP/ND Flooding Control
 - 3.5.3. Handling Failures between GW and WAN Edge Routers
- 4. Integrated Interconnect Solution for EVPN-Overlay Networks
 - 4.1. Interconnect Requirements
 - 4.2. VPLS Interconnect for EVPN-Overlay Networks
 - 4.2.1. Control/Data Plane Setup Procedures on the GWs
 - 4.2.2. Multihoming Procedures on the GWs
 - 4.3. PBB-VPLS Interconnect for EVPN-Overlay Networks
 - 4.3.1. Control/Data Plane Setup Procedures on the GWs
 - 4.3.2. Multihoming Procedures on the GWs
 - 4.4. EVPN-MPLS Interconnect for EVPN-Overlay Networks
 - 4.4.1. Control plane Setup Procedures on the GWs
 - 4.4.2. Data Plane Setup Procedures on the GWs
 - 4.4.3. Multihoming Procedure Extensions on the GWs
 - 4.4.4. Impact on MAC Mobility Procedures
 - 4.4.5. Gateway Optimizations
 - 4.4.6. Benefits of the EVPN-MPLS Interconnect Solution
 - 4.5. PBB-EVPN Interconnect for EVPN-Overlay Networks
 - 4.5.1. Control/Data Plane Setup Procedures on the GWs
 - 4.5.2. Multihoming Procedures on the GWs
 - 4.5.3. Impact on MAC Mobility Procedures
 - 4.5.4. Gateway Optimizations
 - 4.6. EVPN-VXLAN Interconnect for EVPN-Overlay Networks
 - 4.6.1. Globally Unique VNIs in the Interconnect Network
 - 4.6.2. Downstream-Assigned VNIs in the Interconnect Network
- 5. Security Considerations
- 6. IANA Considerations
- 7. References
 - 7.1. Normative References
 - 7.2. Informative References

Acknowledgments

Contributors

Authors' Addresses

1. Introduction

[RFC8365] discusses the use of Ethernet Virtual Private Networks (EVPNs) [RFC7432] as the control plane for Network Virtualization Overlays (NVOs), where VXLAN [RFC7348], NVGRE [RFC7637], or MPLS over GRE [RFC4023] can be used as possible data plane encapsulation options.

While this model provides a scalable and efficient multitenant solution within the Data Center, it might not be easily extended to the Wide Area Network (WAN) in some cases, due to the requirements and existing deployed technologies. For instance, a Service Provider might have an already deployed Virtual Private LAN Service (VPLS) [RFC4761] [RFC4762], VPLS extensions for Provider Backbone Bridging (PBB-VPLS) [RFC7041], EVPN [RFC7432], or PBB-EVPN [RFC7623] network that has to be used to interconnect Data Centers and WAN VPN users. A Gateway (GW) function is required in these cases. In fact, [RFC8365] discusses two main Data Center Interconnect (DCI) solution groups: "DCI using GWs" and "DCI using ASBRs". This document specifies the solutions that correspond to the "DCI using GWs" group.

It is assumed that the NVO GW and the WAN Edge functions can be decoupled into two separate systems or integrated into the same system. The former option will be referred to as "decoupled interconnect solution" throughout the document, whereas the latter one will be referred to as "integrated interconnect solution".

The specified procedures are local to the redundant GWs connecting a DC to the WAN. The document does not preclude any combination across different DCs for the same tenant. For instance, a "Decoupled" solution can be used in GW1 and GW2 (for DC1), and an "Integrated" solution can be used in GW3 and GW4 (for DC2).

While the Gateways and WAN Provider Edges (PEs) use existing specifications in some cases, the document also defines extensions that are specific to DCI. In particular, those extensions are:

- * The Interconnect Ethernet Segment (I-ES), an Ethernet Segment that can be associated to a set of pseudowires (PWs) or other tunnels. The I-ES defined in this document is not associated with a set of Ethernet links, as per [RFC7432], but rather with a set of virtual tunnels (e.g., a set of PWs). This set of virtual tunnels is referred to as vES [VIRTUAL-ES].
- * The use of the Unknown MAC Route (UMR) in a DCI scenario.
- * The processing of EVPN routes on Gateways with MAC-VRFs connecting EVPN-Overlay and EVPN-MPLS networks, or EVPN-Overlay and EVPN-Overlay networks.

2. Conventions and Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

AC: Attachment Circuit

ARP: Address Resolution Protocol

BUM: Broadcast, Unknown Unicast and Multicast (traffic)

CE: Customer Equipment

CFM: Connectivity Fault Management

DC: Data Center

DCI: Data Center Interconnect

DF: Designated Forwarder

EVI: EVPN Instance

EVPN: Ethernet Virtual Private Network, as in [RFC7432]

EVPN Tunnel binding: refers to a tunnel to a remote PE/NVE for a given EVI. Ethernet packets in these bindings are encapsulated with the Overlay or MPLS encapsulation and the EVPN label at the bottom of the stack.

ES: Ethernet Segment

ESI: Ethernet Segment Identifier

GW: Gateway or Data Center Gateway

I-ES and I-ESI: Interconnect Ethernet Segment and Interconnect Ethernet Segment Identifier. An I-ES is defined on the GWs for multihoming to/from the WAN.

MAC Media Access Control

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

MP2P and LSM tunnels: refer to multipoint-to-point and label switched multicast tunnels

ND: Neighbor Discovery

NDF: Non-Designated Forwarder

NVE: Network Virtualization Edge
 NVGRE: Network Virtualization using Generic Routing Encapsulation
 NVO: Network Virtualization Overlay
 OAM: Operations, Administration, and Maintenance
 PBB: Provider Backbone Bridging
 PE: Provider Edge
 PW: Pseudowire
 RD: Route Distinguisher
 RR: Route Reflector
 RT: Route Target
 S/C-TAG: refers to a combination of Service Tag and Customer Tag in a 802.1Q frame
 TOR: Top-Of-Rack
 UMR: Unknown MAC Route
 vES: virtual Ethernet Segment
 VNI/VSID: refers to VXLAN/NVGRE virtual identifiers
 VPLS: Virtual Private LAN Service
 VSI: Virtual Switch Instance or VPLS instance in a particular PE
 VXLAN: Virtual eXtensible LAN

3. 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. Note that, although not shown in Figure 1, GWs may have local Attachment Circuits (ACs).

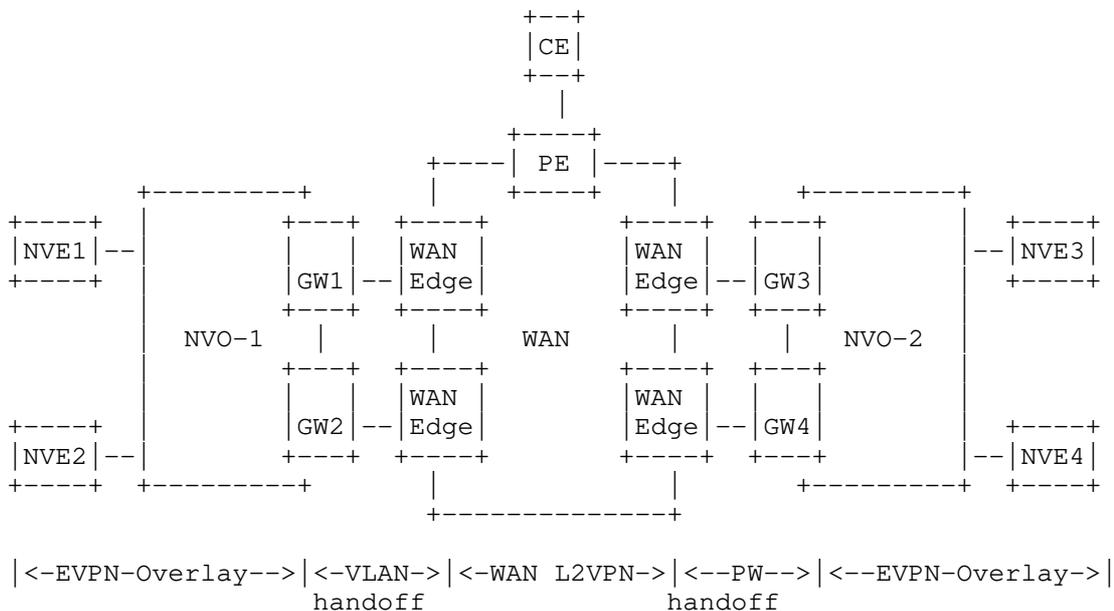


Figure 1: Decoupled Interconnect Model

The following section describes the interconnect requirements for

this model.

3.1. Interconnect Requirements

The decoupled interconnect architecture is intended to be deployed in networks where the EVPN-Overlay and WAN providers are different entities and a clear demarcation is needed. This solution solves the following requirements:

- * A simple connectivity handoff between the EVPN-Overlay network provider and the WAN provider so that QoS and security enforcement are easily accomplished.
- * Independence of the L2VPN technology deployed in the WAN.
- * Multihoming between GW and WAN Edge routers, including per-service load balancing. Per-flow load balancing is not a strong requirement, since a deterministic path per service is usually required for an easy QoS and security enforcement.
- * Support of Ethernet OAM and Connectivity Fault Management (CFM) [IEEE.802.1AG] [Y.1731] functions between the GW and the WAN Edge router to detect individual AC failures.
- * Support for the following optimizations at the GW:
 - Flooding reduction of unknown unicast traffic sourced from the DC Network Virtualization Edge (NVE) devices.
 - Control of the WAN MAC addresses advertised to the DC.
 - Address Resolution Protocol (ARP) and Neighbor Discovery (ND) flooding control for the requests coming from the WAN.

3.2. VLAN-Based Handoff

In this option, the handoff between the GWs and the WAN Edge routers is based on VLANs [IEEE.802.1Q]. 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 [RFC8365].

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], [RFC7432], and [RFC7623].

3.3. PW-Based Handoff

If MPLS between the GW and the WAN Edge router is an option, a PW-based interconnect solution can be deployed. In this option, the handoff between both routers is based on FEC128-based PWs [RFC4762] or FEC129-based PWs (for a greater level of network automation) [RFC6074]. Note that this model still provides a clear demarcation 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 handoff and less provisioning overhead than a combined C/S-TAG handoff. The PW-based handoff 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 [RFC3031], the GW MUST support an interworking function in each MAC-VRF that requires extension to the WAN:

- * If a FEC128-based PW is used between the MAC-VRF (GW) and the VSI (WAN Edge), the corresponding Virtual Connection Identifier (VCID) MUST be provisioned on the MAC-VRF and match the VCID used in the peer VSI at the WAN Edge router.
- * 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.

If a PW-based handoff is used, the GW's AC (or point of attachment to the EVPN instance) uses a combination of a PW label and VLAN IDs. PWs are treated as service interfaces, defined in [RFC7432].

3.4. Multihoming Solution on the GWs

EVPN single-active multihoming -- i.e., per-service load-balancing multihoming -- is required in this type of interconnect.

The GWs will be provisioned with a unique ES for each WAN interconnect, and the handoff attachment circuits or PWs between the GW and the WAN Edge router will be assigned an ESI for each such ES. The ESI will be administratively configured on the GWs according to the procedures in [RFC7432]. This I-ES will be referred to as "I-ES" hereafter, and its identifier will be referred to as "I-ESI". Different ESI types are described in [RFC7432]. The use of Type 0 for the I-ESI is RECOMMENDED in this document.

The solution (on the GWs) MUST follow the single-active multihoming procedures as described in [RFC8365] for the provisioned I-ESI -- i.e., Ethernet A-D routes per ES and per EVI will be advertised to the DC NVEs for the multihoming functions, and ES routes will be advertised so that ES discovery and Designated Forwarder (DF) procedures can be followed. The MAC addresses learned (in the data plane) on the handoff links will be advertised with the I-ESI encoded in the ESI field.

3.5. Gateway Optimizations

The following GW features are optional and optimize the control plane and data plane in the DC.

3.5.1. MAC Address Advertisement Control

The use of EVPN in NVO networks brings a significant number of benefits, as described in [RFC8365]. 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 [RFC4684] 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 (UMR) that is advertised into a given DC by each of the DC's GWs. This route is defined in [RFC7543] and is a regular EVPN MAC/IP Advertisement route in which the MAC Address Length is set to 48, the MAC address is set to 0, and the ESI field is set to the DC GW's I-ESI.

An NVE within that DC that understands and processes the UMR will send unknown unicast frames to one of the DC's GWs, which will then forward that packet to the correct egress PE. Note that, because the ESI is set to the DC GW's I-ESI, all-active multihoming can be applied to unknown unicast MAC addresses. An NVE that does not understand the Unknown MAC Route will handle unknown unicast as

described in [RFC7432].

This document proposes that local policy determine whether MAC addresses and/or the UMR 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 UMR. Advertising all the DC MAC addresses in the control/management plane is usually the case when the NVEs reside in hypervisors. Refer to [RFC8365], Section 7.

It is worth noting that the UMR usage in [RFC7543] and the UMR usage in this document are different. In the former, a Virtual Spoke (V-spoke) does not necessarily learn all the MAC addresses pertaining to hosts in other V-spokes of the same network. The communication between two V-spokes is done through the Default MAC Gateway (DMG) until the V-spokes learn each other's MAC addresses. In this document, two leaf switches in the same DC are recommended for V-spokes to learn each other's MAC addresses for the same EVI. The leaf-to-leaf communication is always direct and does not go through the GW.

3.5.2. ARP/ND 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.

3.5.3. Handling Failures between GW and WAN Edge Routers

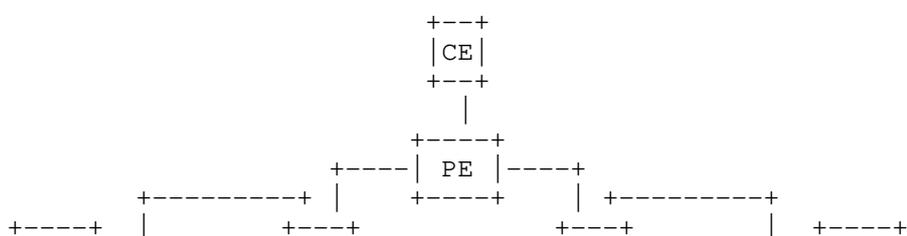
Link/PE failures are handled on the GWs as specified in [RFC7432]. The GW detecting the failure will withdraw the EVPN routes, as per [RFC7432].

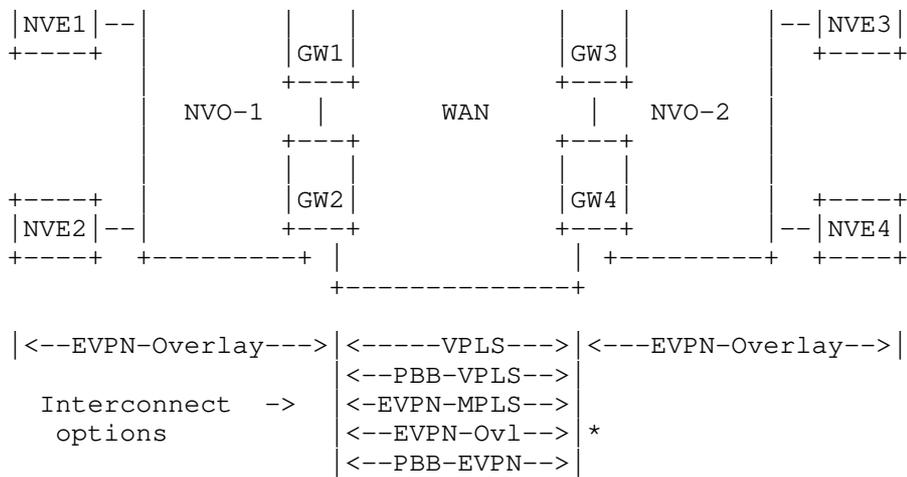
Individual AC/PW failures may be detected by OAM mechanisms. For instance:

- * If the interconnect solution is based on a VLAN handoff, Ethernet-CFM [IEEE.802.1AG] [Y.1731] 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.
- * If the interconnect solution is based on a PW handoff, the Label Distribution Protocol (LDP) PW Status bits TLV [RFC6870] may be used to detect individual PW failures on both the GW and WAN Edge router.

4. Integrated Interconnect Solution for EVPN-Overlay Networks

When the DC and the WAN are operated by the same administrative entity, the Service Provider can decide to integrate the GW and WAN Edge PE functions in the same router for obvious reasons to save as relates to Capital Expenditure (CAPEX) and Operating Expenses (OPEX). This is illustrated in Figure 2. Note that this model does not provide an explicit demarcation link between DC and WAN anymore. Although not shown in Figure 2, note that the GWs may have local ACs.





* EVPN-Ovl stands for EVPN-Overlay (and it's an interconnect option).

Figure 2: Integrated Interconnect Model

4.1. Interconnect Requirements

The integrated interconnect solution meets the following requirements:

- * Control plane and data plane interworking between the EVPN-Overlay network and the L2VPN technology supported in the WAN, irrespective of the technology choice -- i.e., (PBB-)VPLS or (PBB-)EVPN, as depicted in Figure 2.
- * Multihoming, including single-active multihoming with per-service load balancing or all-active multihoming -- i.e., per-flow load-balancing -- as long as the technology deployed in the WAN supports it.
- * Support for end-to-end MAC Mobility, Static MAC protection and other procedures (e.g., proxy-arp) described in [RFC7432] as long as EVPN-MPLS is the technology of choice in the WAN.
- * Independent inclusive multicast trees in the WAN and in the DC. That is, the inclusive multicast tree type defined in the WAN does not need to be the same as in the DC.

4.2. VPLS Interconnect for EVPN-Overlay Networks

4.2.1. Control/Data Plane Setup Procedures on the GWs

Regular MPLS tunnels and Targeted LDP (tLDP) / BGP sessions will be set up to the WAN PEs and RRs as per [RFC4761], [RFC4762], and [RFC6074], and overlay tunnels and EVPN will be set up as per [RFC8365]. Note that different route targets for the DC and the WAN are normally required (unless [RFC4762] is used in the WAN, in which case no WAN route target is needed). A single type-1 RD per service may be used.

In order to support multihoming, the GWs will be provisioned with an I-ESI (see Section 3.4), which will be unique for each interconnection. In this case, the I-ES will represent the group of PWs to the WAN PEs and GWs. All the [RFC7432] procedures are still followed for the I-ES -- 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:

- * VPLS PWs will be instantiated in the WAN split-horizon group.
- * 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 (although not shown in Figure 2), 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 composed of the sublist created by the inclusive multicast routes and the sublist created for VPLS in the WAN. BUM frames received from a local Attachment Circuit (AC) 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, so that loops and packet duplication are avoided. In case a GW can successfully establish both an EVPN binding and a PW to the same far-end PE, the EVPN binding will prevail, and the PW will be brought down operationally.

The optimization procedures described in Section 3.5 can also be applied to this model.

4.2.2. Multihoming Procedures on the GWs

This model supports single-active multihoming on the GWs. All-active multihoming is not supported by VPLS; therefore, it cannot be used on the GWs.

In this case, for a given EVI, all the PWs in the WAN split-horizon group are assigned to I-ES. All the single-active multihoming procedures as described by [RFC8365] will be followed for the I-ES.

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

4.3. PBB-VPLS Interconnect for EVPN-Overlay Networks

4.3.1. 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 4.2.1 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:

- * Attempts to set up a PW between the two GWs within the B-component context will never be blocked.
- * If a PW exists between two GWs for the B-component and an attempt is made to set up an EVPN binding on an I-component linked to that B-component, the EVPN binding will be kept down operationally. Note that the BGP EVPN routes will still be valid but not used.
- * 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 optimization procedures described in Section 3.5 can also be

applied to this interconnect option.

4.3.2. Multihoming Procedures on the GWs

This model supports single-active multihoming on the GWs. All-active multihoming is not supported by this scenario.

The single-active multihoming procedures as described by [RFC8365] will be followed for the I-ES for each EVI instance connected to the B-component. Note that in this case, for a given EVI, all the EVPN bindings in the I-component are assigned to the I-ES. The non-DF GW for the I-ES will block the transmission and reception of all the I-component EVPN bindings for BUM and unicast traffic. When learning MACs from the WAN, the non-DF MUST NOT advertise EVPN MAC/IP routes for those MACs.

4.4. EVPN-MPLS Interconnect for EVPN-Overlay Networks

If EVPN for MPLS tunnels (referred to as "EVPN-MPLS" hereafter) are supported in the WAN, an end-to-end EVPN solution can be deployed. The following sections describe the proposed solution as well as its impact on the procedures from [RFC7432].

4.4.1. 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 set up one BGP EVPN session to the DC RR (or two BGP EVPN sessions if there are redundant DC RRs) and one session to the WAN RR (or two sessions if there are redundant WAN RRs).

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 [RFC9012] 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-ES and its assigned I-ESI will be configured on the GWs for multihoming. This I-ES represents the WAN EVPN-MPLS PEs to the DC but also the DC EVPN-Overlay NVEs to the WAN. Optionally, different I-ESI values are 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 instead will be consumed and re-advertised (if needed):

- * Ethernet A-D routes, ES routes, and Inclusive Multicast routes are consumed by the GWs and processed locally for the corresponding [RFC7432] procedures.
- * MAC/IP advertisement routes will be received and 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 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 for

encapsulations using a VNI).

- The appropriate Route Targets (RTs) and [RFC9012] BGP encapsulation extended community will be used according to [RFC8365].

The GWs will also generate the following local EVPN routes that will be sent to the DC and WAN, with their corresponding RTs and [RFC9012] BGP encapsulation extended community values:

- * ES route(s) for the I-ESI(s).
- * Ethernet A-D routes per ES 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.
- * Inclusive Multicast routes with independent tunnel-type value for the WAN and DC. For example, a P2MP Label Switched Path (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.
- * MAC/IP advertisement routes for MAC addresses learned in local attachment circuits. Note that these routes will not include the I-ESI value in the ESI field. These routes will include a zero ESI or a non-zero ESI for local multihomed 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 the above local service routes: set-DC will be sent to the DC RRs and will include an 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:

- * 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, giving the set-WAN route higher priority. An administrative option MAY change this preference so that the set-DC route is selected first.
- * MAC/IP advertisement routes for local attachment circuits: As above, the GW will select only one, giving the route from the set-WAN a higher priority. As with the Inclusive multicast routes, an administrative option MAY change this priority.

4.4.2. 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 the case of VLAN-aware bundle services) will be set up 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:

- * EVPN-Overlay flooding list (composed of bindings to the remote NVEs or multicast tunnel to the NVEs).
- * 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 receives 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 nonzero ESI).

4.4.3. Multihoming Procedure Extensions on the GWs

This model supports single-active as well as all-active multihoming.

All the [RFC7432] multihoming procedures for the DF election on I-ES(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] multihomed Ethernet Segments.

The forwarding behavior on the GWs is described below:

- * Single-active multihoming; 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-ES. 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 its 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-ES, 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 reoriginate 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 to know who the primary GW is, based on the reception of the MAC/IP routes.
- * All-active multihoming; 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.

Consider the diagram in Figure 2 for EVPN-MPLS interconnect and all-active multihoming, 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, creating packet duplication. 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.

4.4.6. Benefits of the EVPN-MPLS Interconnect Solution

The "DCI using ASBRs" solution described in [RFC8365] and the GW solution with EVPN-MPLS interconnect may be seen as similar, since they both retain the EVPN attributes between Data Centers and throughout the WAN. However, the EVPN-MPLS interconnect solution on the GWs has significant benefits compared to the "DCI using ASBRs" solution:

- * As in any of the described GW models, this solution supports the connectivity of local attachment circuits on the GWs. This is not possible in a "DCI using ASBRs" solution.
- * Different data plane encapsulations can be supported in the DC and the WAN, while a uniform encapsulation is needed in the "DCI using ASBRs" solution.
- * Optimized multicast solution, with independent inclusive multicast trees in DC and WAN.
- * 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.
- * 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.
- * Proxy-ARP/ND function on the DC GWs can be leveraged to reduce ARP/ND flooding in the DC or/and the WAN.

4.5. 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.

4.5.1. 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 customer

MACs (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-ES will be configured on the GWs for multihoming, but its I-ESI will only be used in the EVPN control plane for the I-component EVI. No unreserved ESIs will be used in the control plane of the B-component EVI, as per [RFC7623]. That is, the I-ES will be represented to the WAN PBB-EVPN PEs using shared or dedicated B-MACs.

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 4.3.1.

4.5.2. Multihoming Procedures on the GWs

This model supports single-active as well as all-active multihoming.

The forwarding behavior of the DF and non-DF will be changed based on the description outlined in Section 4.4.3, substituting the WAN split-horizon group for the B-component, and using [RFC7623] procedures for the traffic sent or received on the B-component.

4.5.3. 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 the perspective of Mobility and the related procedures described in [RFC7432], the C-MACs learned from the B-component are considered local.

4.5.4. Gateway Optimizations

All the considerations explained in Section 4.4.5 are applicable to the PBB-EVPN interconnect option.

4.6. 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. While multiple Overlay tunnel combinations at the WAN and the DC are possible (MPLSoGRE, NVGRE, etc.), VXLAN is described here, given its popularity in the industry. This section focuses on the specific case of EVPN for VXLAN (EVPN-VXLAN hereafter) and the impact on the [RFC7432] procedures.

The procedures described in Section 4.4 apply to this section, too, only substituting EVPN-MPLS for EVPN-VXLAN control plane specifics and using [RFC8365] "Local Bias" procedures instead of Section 4.4.3. Since there are no ESI labels in VXLAN, GWs need to rely on "Local Bias" to apply split horizon on packets generated from the I-ES and sent to the peer GW.

This use case assumes that NVEs need to use the VNIs or VSIDs as 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 [RFC8365].

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

1. 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 autoderived from the agreed-upon interconnect VNI. The VNI used inside each DC MAY be the same as the interconnect VNI.
2. 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 (a), 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.

4.6.1. 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 4.4 will be followed, considering that the VNIs advertised/received by the GWs will be translated accordingly.

4.6.2. 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 4.6.1. 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 upon on all the GWs and PEs of the interconnect network.

The procedures described in Section 4.4 will be followed, taking into account the considerations above for the VNI translation.

5. Security Considerations

This document applies existing specifications to a number of interconnect models. The security considerations included in those documents, such as [RFC7432], [RFC8365], [RFC7623], [RFC4761], and [RFC4762] apply to this document whenever those technologies are used.

As discussed, [RFC8365] discusses two main DCI solution groups: "DCI

using GWs" and "DCI using ASBRs". This document specifies the solutions that correspond to the "DCI using GWs" group. It is important to note that the use of GWs provides a superior level of security on a per-tenant basis, compared to the use of ASBRs. This is due to the fact that GWs need to perform a MAC lookup on the frames being received from the WAN, and they apply security procedures, such as filtering of undesired frames, filtering of frames with a source MAC that matches a protected MAC in the DC, or application of MAC-duplication procedures defined in [RFC7432]. On ASBRs, though, traffic is forwarded based on a label or VNI swap, and there is usually no visibility of the encapsulated frames, which can carry malicious traffic.

In addition, the GW optimizations specified in this document provide additional protection of the DC tenant systems. For instance, the MAC-address advertisement control and Unknown MAC Route defined in Section 3.5.1 protect the DC NVEs from being overwhelmed with an excessive number of MAC/IP routes being learned on the GWs from the WAN. The ARP/ND flooding control described in Section 3.5.2 can reduce/suppress broadcast storms being injected from the WAN.

Finally, the reader should be aware of the potential security implications of designing a DCI with the decoupled interconnect solution (Section 3) or the integrated interconnect solution (Section 4). In the decoupled interconnect solution, the DC is typically easier to protect from the WAN, since each GW has a single logical link to one WAN PE, whereas in the Integrated solution, the GW has logical links to all the WAN PEs that are attached to the tenant. In either model, proper control plane and data plane policies should be put in place in the GWs in order to protect the DC from potential attacks coming from the WAN.

6. IANA Considerations

This document has no IANA actions.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <<https://www.rfc-editor.org/info/rfc4761>>.
- [RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <<https://www.rfc-editor.org/info/rfc4762>>.
- [RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, DOI 10.17487/RFC6074, January 2011, <<https://www.rfc-editor.org/info/rfc6074>>.
- [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, <<https://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, <<https://www.rfc-editor.org/info/rfc7432>>.

- [RFC7543] Jeng, H., Jalil, L., Bonica, R., Patel, K., and L. Yong, "Covering Prefixes Outbound Route Filter for BGP-4", RFC 7543, DOI 10.17487/RFC7543, May 2015, <<https://www.rfc-editor.org/info/rfc7543>>.
- [RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <<https://www.rfc-editor.org/info/rfc7623>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <<https://www.rfc-editor.org/info/rfc8365>>.
- [RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <<https://www.rfc-editor.org/info/rfc9012>>.

7.2. Informative References

- [IEEE.802.1AG] IEEE, "IEEE Standard for Local and Metropolitan Area Networks Virtual Bridged Local Area Networks Amendment 5: Connectivity Fault Management", IEEE standard 802.1ag-2007, January 2008.
- [IEEE.802.1Q] IEEE, "IEEE Standard for Local and metropolitan area networks--Bridges and Bridged Networks", IEEE standard 802.1Q-2014, DOI 10.1109/IEEESTD.2014.6991462, December 2014, <<https://doi.org/10.1109/IEEESTD.2014.6991462>>.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <<https://www.rfc-editor.org/info/rfc3031>>.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, DOI 10.17487/RFC4023, March 2005, <<https://www.rfc-editor.org/info/rfc4023>>.
- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <<https://www.rfc-editor.org/info/rfc4684>>.
- [RFC6870] Muley, P., Ed. and M. Aissaoui, Ed., "Pseudowire Preferential Forwarding Status Bit", RFC 6870, DOI 10.17487/RFC6870, February 2013, <<https://www.rfc-editor.org/info/rfc6870>>.
- [RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <<https://www.rfc-editor.org/info/rfc7348>>.
- [RFC7637] Garg, P., Ed. and Y. Wang, Ed., "NVGRE: Network Virtualization Using Generic Routing Encapsulation",

[VIRTUAL-ES]

Sajassi, A., Brissette, P., Schell, R., Drake, J. E., and J. Rabadan, "EVPN Virtual Ethernet Segment", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-virtual-eth-segment-06, 9 March 2020,
<<https://tools.ietf.org/html/draft-ietf-bess-evpn-virtual-eth-segment-06>>.

[Y.1731] ITU-T, "OAM functions and mechanisms for Ethernet based networks", ITU-T Recommendation Y.1731, August 2019.

Acknowledgments

The authors would like to thank Neil Hart, Vinod Prabhu, and Kiran Nagaraj for their valuable comments and feedback. We would also like to thank Martin Vigoureux and Alvaro Retana for their detailed reviews and comments.

Contributors

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

Ravi Shekhar
Juniper Networks

Anil Lohiya
Juniper Networks

Wen Lin
Juniper Networks

Florin Balus
Cisco

Patrice Brissette
Cisco

Senad Palislamovic
Nokia

Dennis Cai
Alibaba

Authors' Addresses

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

Email: jorge.rabadan@nokia.com

Senthil Sathappan
Nokia

Email: senthil.sathappan@nokia.com

Wim Henderickx
Nokia

Email: wim.henderickx@nokia.com

Ali Sajassi
Cisco

Email: sajassi@cisco.com

John Drake
Juniper

Email: jdrake@juniper.net