L2VPN Workgroup
Internet Draft
Intended status: Standards Track

J. Rabadan
W. Henderickx
S. Sathappan
S. Palislamovic
Alcatel-Lucent

F. Balus
Nuage Networks

Expires: January 16, 2014

July 15, 2013

Data Center Interconnect Solution for E-VPN Overlay networks
draft-rabadan-l2vpn-dci-evpn-overlay-00.txt

Abstract

This document describes how Network Virtualization Overlay networks (NVO3) can be connected to a Wide Area Network (WAN) in order to extend the layer-2 connectivity required for some tenants. The solution will analyze the interaction between NVO3 networks running E-VPN and other technologies used in the WAN, such as VPLS/PBB-VPLS or E-VPN/PBB-EVPN.

Table of Contents

1. Introduction

[E-VPN-Overlays] discusses the use of E-VPN as the control plane for Network Virtualization Overlay (NVO) networks, 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 existing deployed technologies. For instance, a Service Provider might have an already deployed VPLS network that must be used to interconnect Data Centers.

This document describes a Data Center Interconnect (DCI) solution for E-VPN overlay Data Center networks, assuming that the L2VPN technology deployed in the WAN can be based on:

1. VPLS as defined in [RFC4761][RFC4762][RFC6074] or even PBB-VPLS, as defined in [PBB-VPLS]

2. E-VPN as defined in [E-VPN]

3. PBB-EVPN as defined in [PBB-EVPN]

Each of these DCI models is analyzed in the following sections.

2. VPLS/PBB-VPLS based DCI for E-VPN overlay networks

VPLS and PBB-VPLS are deployed in many Service Providers as the multi-point L2VPN service technology in the WAN. Those Service Providers will require integrating the new virtualized data center services with the L2VPN technology existing in the WAN, so that there is a minimum impact on the Service Provider operations.

By implementing the Data Center Gateway (DC GW) functions described in this section, a Service Provider PE will be able to connect a DC tenant segment to an existing VPLS or PBB-VPLS service, for DC-to-DC layer-2 extension and for user-to-DC layer-2 connectivity.

2.1. VPLS/PBB-VPLS DCI Solution Overview

Figure 1 depicts the reference model described in this section.

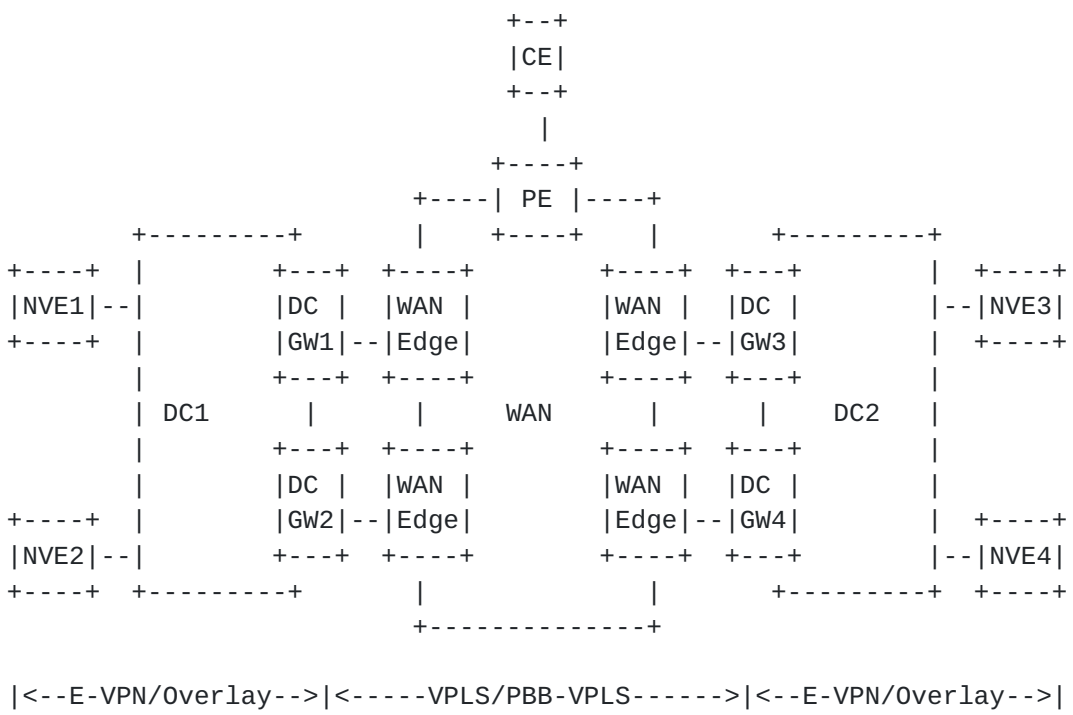


Figure 1 VPLS DCI model

In this model, the WAN Service Provider requires the use of its existing VPLS procedures to extend the layer-2 connectivity for the tenants. There are four potential options in this model:

o VPLS DCI with VLAN-based hand-off

o VPLS DCI with Pseudowire-based hand-off

o VPLS DCI with integrated Gateway and WAN Edge functions

o PBB-VPLS DCI

Section 2.2 describes each specific option.

2.2. VPLS/PBB-VPLS DCI options

2.2.1. VPLS DCI with VLAN-based hand-off

In this option, the hand-off between the DC GWs and the WAN Edge routers is based on 802.1Q VLANs. Each E-VPN Instance (EVI) in the DC GW is connected to a different VPLS Instance (VSI) in the WAN Edge router by using a different C-TAG VLAN ID or a different combination of S-TAG/C-TAG VLAN IDs that match at both sides. In this use-case, the WAN Edge router becomes a VPLS PE with regular VLAN-based Attachment Circuits.

This option is required in those cases where the WAN and DC networks are operated by different entities and a secure demarcation between both is needed (no control plane protocols are run between DC GW and WAN Edge router, and each network can apply its own security and QoS policies independently based on the incoming/outgoing VLAN ID). The disadvantages of this model are the provisioning overhead and the reduced scalability (limited to the VLAN-ID space). The provisioning in the DC GWs can be automated though by the cloud management system.

In this model, the DC GW acts as a regular Network Virtualization Edge (NVE) towards the D. Its control plane, data plane procedures and interactions are described in [E-VPN-Overlays]. From an E-VPN perspective, the connectivity to the WAN Edge routers is treated as VLAN-based service interfaces, therefore there is a 1:1 relation between DCI VLAN ID and EVI. If the data plane encapsulation in the NVO network supports VLAN tags in the encapsulated frames, a VLAN Bundle Service interface is possible in the DCI. As described in [E-VPN-Overlays] this interface type is possible if VXLAN is used and not for NVGRE. NVGRE only supports VLAN-based service interfaces.

The WAN Edge router acts as a VPLS PE. Its functions are described in [RFC4761][RFC4762][RFC6074].

The DC GW multi-homing functions for this model are described in section 2.6.

2.2.2. VPLS DCI with Pseudowire-based hand-off

If MPLS can be enabled between the DC GW and the WAN Edge router, a more scalable DCI solution can be deployed. In this option the hand-off between both routers is based on FEC128-based pseudowires or, alternatively, FEC129-based pseudowires for a greater level of network automation. Note that this model still provides a clear demarcation boundary between DC and WAN, and security/QoS policies may be applied on a per pseudowire basis.

In this model, besides the usual MPLS procedures between DC GW and WAN Edge router, the DC GW MUST support an interworking function in each EVI that requires extension to the WAN:

- If a FEC128-based pseudowire is used between the EVI (DC GW) and the VSI (WAN Edge), the provisioning of the VCID for such pseudowire MUST be supported on the EVI and must match the VCID used in the peer VSI at the WAN Edge router.

- If BGP Auto-discovery [RFC6074] and FEC129-based pseudowires are used between the DC GW EVI and the WAN Edge VSI, the provisioning of the VPLS-ID MUST be supported on the EVI and must match the

VPLS-ID used in the WAN Edge VSI. Note that the Route Distinguisher (RD) and Route Target (RT) already provisioned for its use in E-VPN, can be re-used for VPLS. The WAN Edge VSI will have to be configured with two different RT extended communities. For example, if EVI-1 in DC GW-1 (figure 1) uses RT1, the peer WAN Edge VSI will use RT1 to import/export routes from/to the DC GW and RT2 to import/export routes from/to the remote WAN Edge VSIs. The WAN Edge router will import RT1 and RT2 in two different split-horizon groups, so that traffic to/from the DC GW can be switched to/from the WAN.

The DC GW multi-homing functions for this model are described in section 2.6.

2.2.3. VPLS DCI with integrated Gateway and WAN Edge functions

When the DC and the WAN are operated by the same administrative entity, the Service Provider can decide to integrate the DC GW and WAN Edge PE functions in the same router for obvious CAPEX and OPEX saving reasons. In the example depicted in figure 1 that would mean the WAN Edge routers would be P routers and will not maintain any tenant state. Note that this model does not provide an explicit demarcation between DC and WAN anymore, and ACLs or QoS policies between both networks become a very complex task.

In this option, any EVI instance in the DC GW requiring layer-2 extension to the WAN MUST support an interworking function to VPLS. The EVI will become a VSI from the WAN perspective and will setup a full mesh of pseudowires to all the remote PEs and DC GWs (except to the DC GW of its own DC) and according to the procedures described in [RFC4761][RFC4762][RFC6074].

The DC GW multi-homing functions for this model are described in section 2.6.

2.2.4. PBB-VPLS DCI

This case is a variation of the one described in section 2.2.3. When the DC GW and WAN Edge PE functions can be integrated, PBB-VPLS can also be used as the DCI technology of choice. In this case, the DC GW EVIs will become I-components multiplexed into a B-component that will be connected to the WAN.

Since many EVIs can be multiplexed into the same B-component, this option provides significant savings in terms of pseudowires to be maintained in the WAN.

The DC GW multi-homing functions for this model are described in

section 2.6.

2.3. Unknown MAC route on the DC GWs

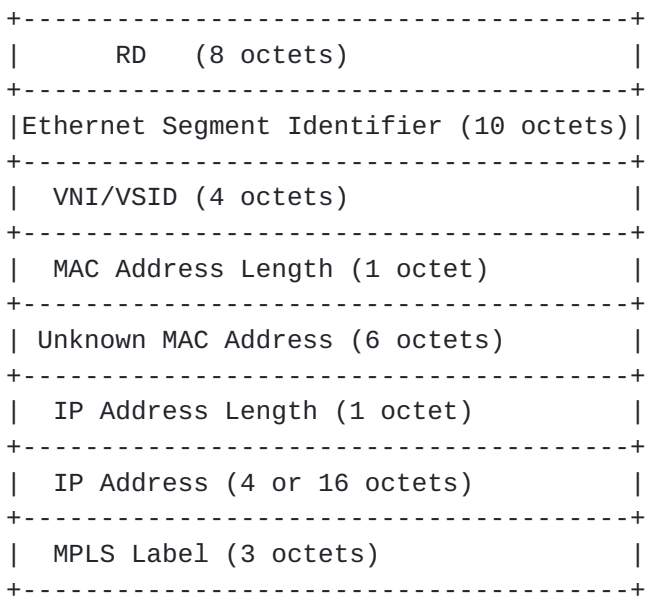
The use of E-VPN, as the control plane of Network Virtualization Networks in the DC, brings a significant number of benefits as described in [E-VPN-Overlays]. There are however two potential issues that SHOULD be addressed when a VPLS DCI is used for a NVO3 DC:

- All the MAC addresses learnt from the WAN side of the VSI must be advertised by BGP E-VPN updates. Even if optimized BGP techniques like RT-constraint are used, the amount of MAC addresses to advertise or withdraw (in case of failure) from the DC GWs can be difficult to control and overwhelming for the DC network, especially when the NVEs reside in the hypervisors.

- As described in [E-VPN-Overlays], when the NVEs reside in the hypervisors, the E-VPN BGP routes and attributes associated with multi-homing are no longer required. The simple reason is the fact that, in a hypervisor environment, there is no need for multi-homing between VMs and NVEs since both, VMs and NVEs, are part of the same hardware. This reduces the required routes to be generated and processed to only two: the MAC Advertisement Route and the Inclusive Multicast Ethernet Tag Route. While this simplification greatly helps the implementation of E-VPN in the DC, it brings back some of the issues related to Multi-Homing that were solved by the removed procedures and that are still applicable for the specific use-case of the DC, since Multi-Homing is required at the DC GWs.

The solution suggested in this document for the VPLS DCI use case is based on the use of an "Unknown MAC route" that is advertised by the two DC GWs. By using this Unknown MAC Route advertisement the user may optionally turn off the advertisement of WAN MAC addresses in the DC GW, hence reducing the control plane overhead and the size of the FDB tables in the NVEs. In addition to this, the Unknown MAC Route may provide a fast convergence solution valid for TORs and hypervisor NVEs, even if they do not support the Ethernet A-D route procedures.

If this procedure is used, when an EVI is created in the DC GWs and the Designated Forwarder (DF) is elected, the DF will send a BGP update containing an "Unknown MAC" address. The Unknown MAC address will be conveyed in an "Unknown MAC" Advertisement Route:



Where the ESI identifies the WAN Ethernet Segment, the VNI/VSID is encoded in the Ethernet Tag Field as explained in [E-VPN-Overlays], the MAC address length is set to 48 and the Unknown MAC address value will be set to 00:00:00:00:00:00. The IP address length will be zero, the IP address value omitted and the MPLS label will be set to zero.

If the DC GW is DF for more than one ES within the same EVI, it will advertise an Unknown MAC route per ES, each one tagged with its corresponding ESI.

As outlined before, there are two main functions that can be carried out by using this Unknown MAC Route: fast convergence for hypervisor NVEs (described in section 2.6.2.3) and disabling unknown unicast flooding in the DC (described in section 2.4).

2.4. Disabling unknown unicast flooding in a DC with VPLS DCI

In DCs where MAC addresses are learnt through the control plane, the use of flooding for unknown destination MAC addresses can be disabled. However, when we use a VPLS DCI, the DC GW will normally learn the WAN MAC addresses in the data plane, therefore, even if the rest of the NVEs in the DC do control plane learning, disabling the unknown unicast flooding is no longer an administrative choice.

The use of the Unknown MAC route in DC GWs allows two configuration options:

a) Disable the unknown flooding in all the NVEs in the DC (except on the DC GWs) if Data Center MACs are learnt through the

control/management plane.

b) Disable the advertisement of the WAN MAC addresses from the DC GWs, so that the control plane overhead and the forwarding table sizes in the NVEs are both reduced.

Both options SHOULD be an administrative configuration choice supported on the DC GWs.

If option b) is enabled, the DC GW will advertise only the Unknown MAC Route for the EVIs on which it is the Designated Forwarder (DF). The NVEs will learn their local MACs through the control/management plane and advertise them in BGP. If any NVE receives a packet to an unknown destination MAC address, and option a) is enabled, the NVE will send the packet to the next-hop associated to the Unknown MAC Route (for each ESI if there is more than one), since the packet is assumed to be destined to the WAN. This assumption is valid since all the DC MACs are learnt in the control/management plane. The DC GW will receive the packet and will do an FDB lookup to find out what VPLS pseudowire or attachment circuit it has to send the packet to. If the destination MAC is unknown for the DC GW, it will flood the packet to the WAN, following standard VPLS procedures.

2.5. ARP-flooding control

The use of the Unknown MAC route may eliminate the unknown flooding within the DC and provide an extra security protection mechanism against an excessive explosion of MAC addresses in the WAN that would trigger the advertisement of a significant number of MAC addresses in the DC.

Another security mechanism, naturally provided by E-VPN in the DC GWs, is the Proxy ARP function. The DC GWs SHOULD build a Proxy ARP table with the IP and MAC address information coded in the MAC advertisement routes coming from the DC NVEs. When the active DC GW receives an ARP request coming from the WAN, the DC GW should check the Proxy ARP table for the EVI and reply to the ARP request as long as the information is available.

This mechanism is specially recommended when VPLS DCI is used on the DC GWs since it protects the DC network from external ARP-flooding.

2.6. Multi-homing solution for VPLS DCI

2.6.1. Multi-homing solution requirements for VPLS DCI

As it can be easily inferred from the scenario in figure 1, a multi-homing solution MUST be provided in the DC. The Multi-homing

requirements on the DC GWs are listed here:

- o A Multi-homing solution MUST be supported by the DC GWs independently of the capabilities of the WAN Edge routers (since they can be managed by a different Service Provider).

- o The Multi-homing solution MUST support service-based (EVI) load-balancing. No flow-based load-balancing is required when VPLS DCI is used.

- o The Multi-homing solution MUST support single-active redundancy mode per E-VPN on the DC GWs, as per [E-VPN]. All-active multi-homing is neither possible if VPLS is used in the DCI nor required since the number of EVIs on the DC GWs is supposed to be large enough so that the traffic between DC and WAN can be fairly distributed.

2.6.2. Multi-homing solution description

When the DCI model is the one described in the section 2.1, a single-active Multi-homing solution is required. Note that, since all-active Multi-homing is not required, only a subset of E-VPN routes and extended communities will be needed to be generated from the DC GWs:

- o Ethernet Segment route and ES-Import route target: required for the Ethernet Segment Auto-Discovery and Designated Forwarder (DF) election between the DC GWs. The DC GWs MUST generate an ES route per WAN network to which they are directly connected, and MUST be able to process the inbound ES routes as per [E-VPN].

- o Ethernet Auto-Discovery (A-D) route per ESI: required for fast convergence and back-up path. The DC GWs MUST generate an A-D route per ESI with an ESI Label extended community. The active-standby flag will be always set and the label field will be zero (no Split-Horizon procedures are required on the DC GWs as per [E-VPN]). The DC GWs will be able to process the received A-D routes per ESI.

- o Ethernet Auto-Discovery (A-D) route per EVI: the DC GWs will NOT generate A-D routes per EVI, since no aliasing functions are required for single-active Multi-homing. The DC GWs however MUST support processing A-D routes per EVI, since there might be some TORs in the DC supporting all-active Multi-homing.

- o MAC Advertisement route and MAC Mobility extended community: they MUST be supported at generation and reception as per [E-VPN-Overlays].

- o Inclusive Multicast Ethernet Tag route and PMSI Tunnel attribute:

they MUST be supported at generation and reception as per [E-VPN-Overlays].

The above routes and communities will be used for the following Multi-homing functions:

2.6.2.1. Multi-homed Ethernet Segment Auto-Discovery

The DC GWs will advertise an Ethernet Segment route per WAN Ethernet Segment (ES), with the corresponding ES-Import extended community. There will be a single ESI per WAN network, i.e. DC GW1 and DC GW2 will only advertise one ESI in the example of figure 1, and only the DC GWs of the DC will import the ES route for the WAN ESI, as per [E-VPN].

2.6.2.2. Designated Forwarder (DF) election and forwarding

The DF election will be carried out as described in [E-VPN]. Service carving is recommended so that there can be per EVI load-balancing to/from the WAN. Assuming DC GW1 is elected as DF for a given EVI1, DC GW1 will be the only DC GW sending/receiving traffic to/from the WAN for EVI1. DC GW2 will block the transmission and reception of any traffic (including unicast and multi-destination traffic) to/from the WAN for EVI1.

The use of OAM is recommended from the non-DF to the VPLS network, so that the VPLS PEs do not send any traffic to the non-DF DC GW for the EVI in which the DC GW is non-DF:

- If the VPLS DCI solution is based on a VLAN hand-off, 802.1ag/Y.1731 Ethernet-CFM can be used by the non-DF DC GW so that the peer WAN Edge router do not send any traffic to the DC GW for that particular EVI.

- If the VPLS DCI solution is based on a pseudowire hand-off, the LDP PW Status bits TLV can be used by the non-DF to signal "Standby status" to the WAN Edge router for that particular EVI.

- If the VPLS DCI is based on an integrated DC GW and WAN Edge router solution where the EVI is part of the VPLS full mesh of pseudowires, the non-DF DC GW can also make use of the LDP PW Status bits TLV to let the remote PEs know that it is not forwarding traffic for that EVI/VSI.

2.6.2.3. Fast Convergence using the Unknown MAC Route

[E-VPN] proposes a Fast Convergence mechanism, so that when there is an ES failure on the DF router, the failover time can be uniform and

independent of the number of MACs and EVI services in the DC GWs. This is done by having the DC GWs advertising an A-D route per WAN Ethernet Segment. Upon a failure in connectivity to the WAN, the DF withdraws the Ethernet A-D route for the WAN Ethernet Segment so that the NVEs in the DC receiving the BGP withdraw can update their FDB for all the MAC addresses associated to the WAN ES.

This mechanism is valid as long as the NVEs in the DC support the Ethernet A-D route per ESI. However, as described in [E-VPN-Overlays], in the Data Center there will be a mix of NVEs supporting Ethernet A-D routes (TORs with dual-homed servers) and NVEs NOT supporting Ethernet A-D routes (hypervisors), hence a complementary fast convergence mechanism is required for those.

While the existing E-VPN Mass Withdraw procedure will be used for NVEs supporting the processing of Ethernet A-D routes, this document describes a complementary procedure for NVEs not supporting the processing of Ethernet A-D routes. The new procedure does not require the addition of any new route or extended community. It is just based on the interpretation of the Unknown MAC Route described in section 2.3 which will be sent by the DC GWs in regular MAC advertisement routes. The user MAY decide whether the Unknown MAC Route procedure is used only by the hypervisors or by the hypervisors and the TORs too.

Only one of the DC GWs will advertise the Unknown MAC Route per EVI and per WAN ESI. The DF will also advertise all the MAC addresses being learnt from the WAN Ethernet Segment (assuming option b in section 2.4 is not turned on). The hypervisor NVEs will import the Unknown MAC route as well as the rest of the WAN MAC addresses associated to the active DC GW. The Unknown MAC route is used by the active DC GW as a way of signaling that it owns the reachability to the WAN Ethernet Segment (ES) for a given EVI. The Unknown MAC address (00:...:00/48) conveyed in the Unknown MAC route will be added to the corresponding EVI forwarding table at the remote NVE.

When the WAN Ethernet Segment active path fails (due to a port or link failure), the DC GW will withdraw the Unknown MAC route on all the EVIs for which it is the DF. This triggers all the hypervisor NVEs that receive the withdraw advertisement to immediately invalidate all the MAC addresses associated to the Ethernet Segment, as opposed to having to wait for each individual MAC to be withdrawn.

This function is compatible with the E-VPN Fast Convergence procedure carried out by the use of the Ethernet A-D route. The Ethernet A-D route can still be used for TOR NVEs supporting all the E-VPN routes.

Note that while the E-VPN mass withdraw provides a fast convergence mechanism independent of the number of services and MACs, the Unknown MAC withdraw provides a fast convergence mechanism per service, independent of the number of MACs in each service, i.e. convergence is not expected to be uniform for all the services, but uniform for all the hosts within a service. The use of the Unknown MAC route can significantly speed up the convergence in hypervisor NVEs, especially in services with a fair amount of MACs.

3. E-VPN DCI for E-VPN overlay networks

Another potential DCI technology that can be used in the WAN is E-VPN. Assuming E-VPN for MPLS tunnels is used in the WAN, the use of a DC GW is required if the overlay tunneling technology deployed within the DC is not MPLS over GRE, i.e. if VXLAN or NVGRE are used.

Figure 2 illustrates this E-VPN DCI model.

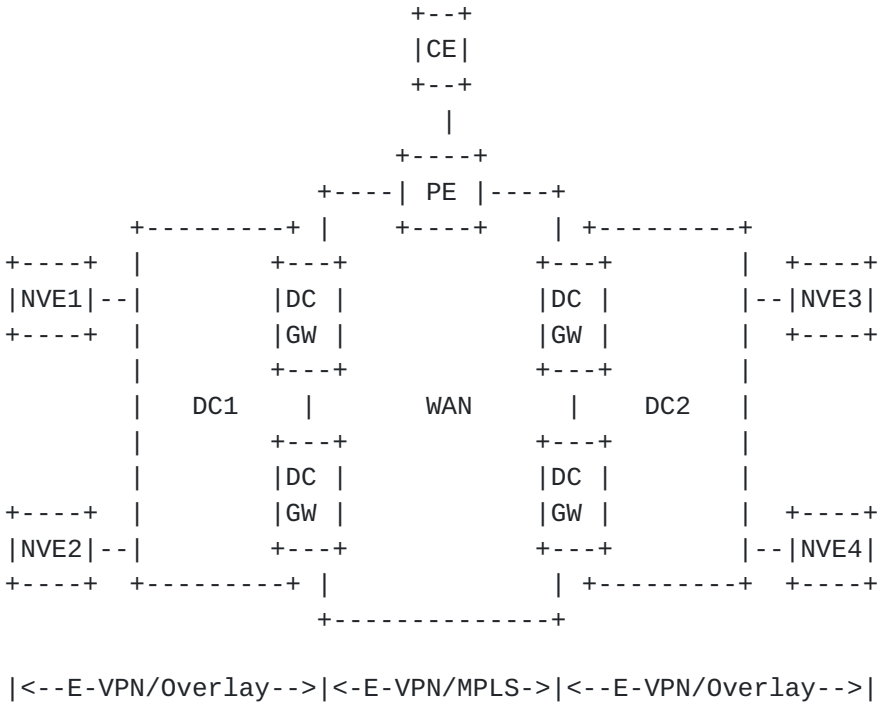


Figure 2 E-VPN DCI model

More information will be added in future versions of this document.

4. PBB-EVPN DCI for E-VPN overlay networks

[PBB-EVPN] is yet another DCI option. It requires the use of DC GWs where the multiplexing of I-components into the B-component is carried out. E-VPN will run in both components.

More information will be added in future versions of this document.

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].

DF: Designated Forwarder

DC GW: Data Center Gateway

DCI: Data Center Interconnect

ES: Ethernet Segment

ESI: Ethernet Segment Identifier

EVI: E-VPN Instance

NVE: Network Virtualization Edge

TOR: Top-Of-Rack switch

VNI/VSID: refers to VXLAN/NVGRE virtual identifiers

6. Security Considerations

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, January 2007.

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

Signaling", RFC 4762, January 2007.

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

8.2. Informative References

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

[E-VPN-OVERLAYS] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using E-VPN", draft-sd-l2vpn-evpn-overlay-01.txt, work in progress, February, 2013

9. Acknowledgments

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

10. Authors' Addresses

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

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

Florin Balus
Nuage Networks
Email: florin@nuagenetworks.net

Senthil Sathappan
Alcatel-Lucent
Email: senthil.sathappan@alcatel-lucent.com

Senad Palislamovic
Alcatel-Lucent

Email: senad.palislamovic@alcatel-lucent.com