BGP MPLS based EVPN
And its implementation and use cases

KANWAR ALAM SINGH

University of Alberta

Supervisor: Juned Noonari

Submitted to: Dr. Mike MacGregor

Abstract

MPLS based VPLS is proven L2VPN technology delivering Ethernet based Services. However Ethernet and its requirements are continually changing due to modernization of Networks, Ethernet Service model such as separation of Data and Control plane, MEF based service delivery etc. In order to meet those requirements EVPN are introduced in the industry under L2VPN IEFT Working Group.

Intent of project is research EVPN, its requirements, its various implementation methods and comparing it with VPLS to justify its relative benefits.

Due to EVPN being a new technology not much documentation could be found from the vendors like Cisco and Alcatel. Juniper only being the vendor that has the command reference for EVPN out. However, I succeeded in implementing a Lab scenario of BGP based VPLS with Multi-homing on Alcatel Lucent 7750 SR-OS routers.

Table of Contents

Table of Figures

1 Introduction

During the last decade the demand for implementing scalable L2 VPN has been increased which led to development in the area of Layer 2 Services. Many MPLS based technologies were brought in, Virtual Private LAN Service being one of them. With VPLS, multipoint Ethernet services over the MPLS infrastructure can provided using a full mesh of pseudowires. It emulates as a virtual bridge to the end customers, and moreover they are free to run any routing protocol of their choice. However VPLS has its own demerits, one of them being MAC learning happens in data plane, other drawbacks include no multihoming with active-active link, no multipoint to multipoint multicast LSPs. To overcome these shortcomings a new solution is proposed called Ethernet Virtual Private Network.

Internet Engineering Task Force (IETF) recently published their 11th internet draft which describes the procedures for BGP MPLS based EVPN. RFC 7209 published by IETF describes the requirements for EVPN and how it fulfills the demands of the new applications such Data Center Interconnect and overcomes the limitations of current L2VPN technology VPLS.

The scope of the project is to study EVPN, specifying its advantages over VPLS and the different implementation scenario possible.

2 Background & Technology Overview

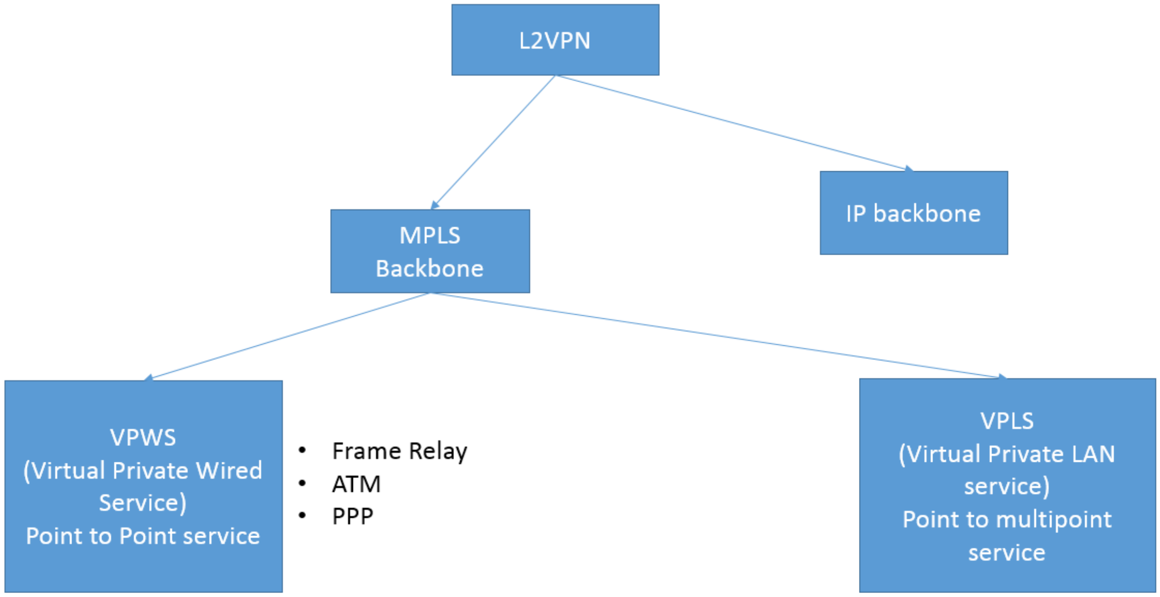
2.1 MPLS:

Multi Protocol Label Switching is a widely used technology since the past decade. At first it was being used for fast switching of packets but due to its scalability and resiliency capabilities soon it became the most popular and vastly used technology across the networks. It simplifies the process of routing the packets through the network by labeling the packets, this leads to fast packet forwarding as the packet itself does not need to be examined and the routing decisions can be made based on the labels itself. The feature of traffic management and the capability to support multiple service models makes it more appealing to the ISPs.

2.2 L2VPN

Layer 2 VPN allows users to connect different LANs together over Layer 3 networks that are using MPLS or IP core. From Customer Prospective it seems that they are connected through a big switch and feel they are part of the same broadcast domain.

Layer 2 VPN provides users the flexibility to scale bandwidth according to their needs. L2 Ethernet are preferred over L3 IP VPN for efficient transport of Non-IP applications including traditional voice and video, along with this, application security is an another L2 consideration as the customers can control their own routing tables.


L2VPN
MPLS Backbone
IP backbone
VPWS (Virtual Private Wired Service) Point to Point service
• Frame Relay
• ATM
• PPP
VPLS (Virtual Private LAN service) Point to multipoint service

Unlike in case of Layer 3 VPN, with Layer 2 VPN service the Customer Edge router and Provider Edge router does not need to peer up, the customer routing tables are not stored on provider network reducing complexity.

2.3 VPLS (Virtual Private LAN Service)

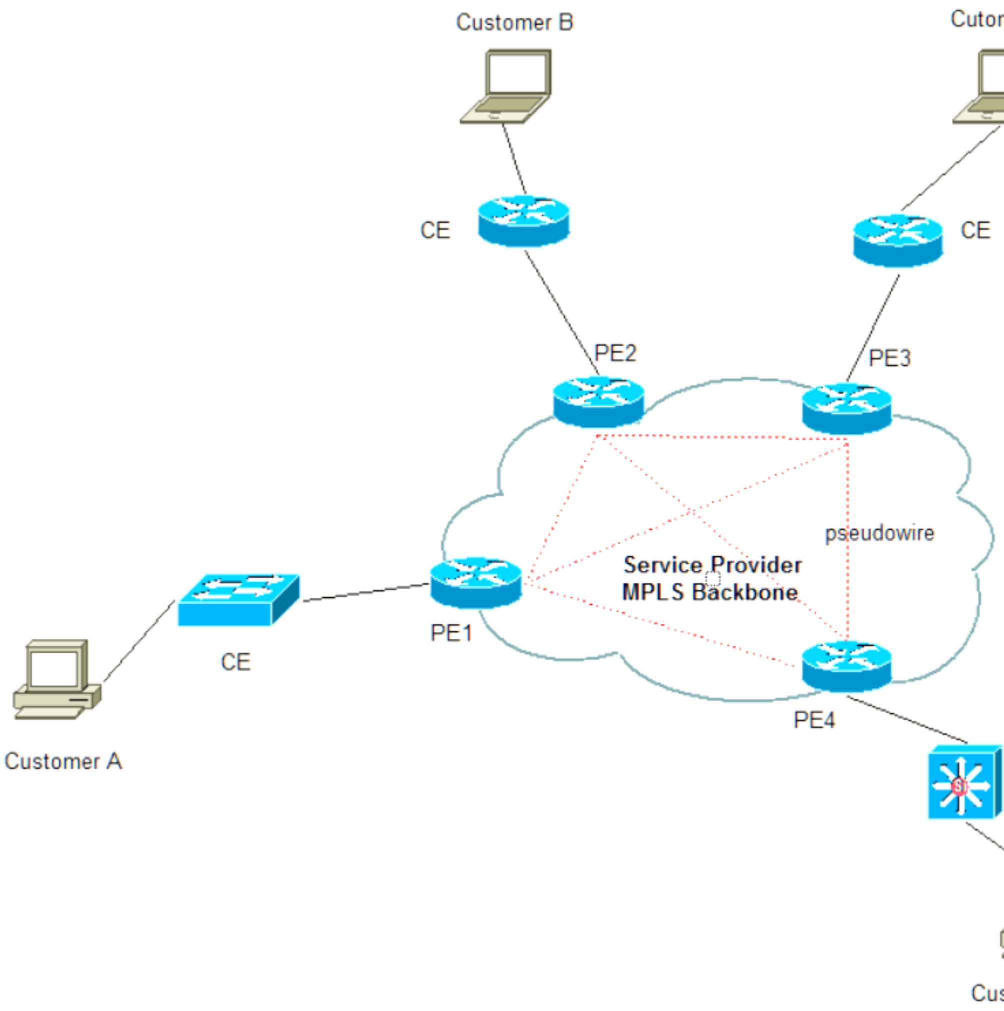
VPLS is a Layer 2 multipoint Ethernet service over the MPLS network, it allows to connect different geographically separated LAN sites together in a single bridged domain over the MPLS backbone and the end users at the sites get an emulation that they are all connected on the same LAN regardless of their actual location. Moreover customers are free to run any Layer 3 protocol between the sites.

VPLS was originally being used to provide transparent LAN service to enterprise customers, but now service providers are also using it as an infrastructure technology with the emergence of Metro-Ethernet networks.

As being a multipoint service unlike other L2VPN services, VPLS requires just one logical interface between CE and PE to be in a full mesh. All the PEs are fully meshed by pseudowires so that the PE receiving the frame can identify for which VPLS instance it belongs to on the basis of pseudowire labels. One advantage of being fully meshed is that there is no need for the PEs to run STP to prevent loops, it uses split horizon to prevent forwarding loops.PE maintains a separate forwarding table for each VPLS customer.

In VPLS there is no advertisement of MAC addresses, they are always learnt by the PEs.MAC aging mechanisms are used, the stale MAC are removed from the forwarding table. Furthermore if the limit reaches the MAC addresses can be removed which have not been used for a while.

Figure 1. A simple VLPS Model



VPLS Control Plane

VPLS control plane has two main functions:

- Auto-Discovery: It refers to the process of finding out which PE router belongs to which VPLS instance.

- Signaling: It Includes the setup, maintain and teardown of Pseudowires.

There are two alternate ways of deploying control planes, one is using LDP and the other one is using BGP. The scaling characteristics of both these control planes differ a lot. BGP VPLS has numerous advantages over LDP VPLS.

While using LDP, in order for the signaling of full mesh pseudowires a full mesh of targeted LDP sessions is required between the PEs. There is no auto discovery in case we are using LDP for signaling so these LDP sessions must be manually configured on each PE router. These sessions carry the VC ID which is used to identify to which VPLS instance the LDP message refers to. While using BGP for signaling, with auto discovery capability a PE gets to know which PEs are the members of a given VPLS instance. The BGP NLRI contains the information that let the automatic setting up the full mesh of pseudowires. On each PE router a Route Target (RT) and a Route Distinguisher (RD) is configured for each VPLS.RT is the same for a particular VPLS across all PEs and is used to identify to which VPLS the BGP message belongs to, while RD is used to distinguish between routes. Each PE has a identifier for each VPLS known as VPLS Edge Identifier (VE ID), it should be unique for each VPLS.VE ID is communicated over to other PEs.VE ID and the BGP NLRI provides the means to the other PEs to calculate value of the pseudo label required to reach the advertising PE. Advantage of using BGP for signaling is that all PEs get to know about each other without any manual configurations.

VPLS Forwarding Plane

Forwarding Plane Mechanisms for the unicast and to some extent for multicast traffic are almost the same for both BGP VPLS and LDP VPLS.A PE VPLS data plane acts as a learning bridge performing the same functions of a standard bridge such as MAC address learning, flooding and aging. For each VPLS instance a separate MAC forwarding table is maintained by each participating PE router.

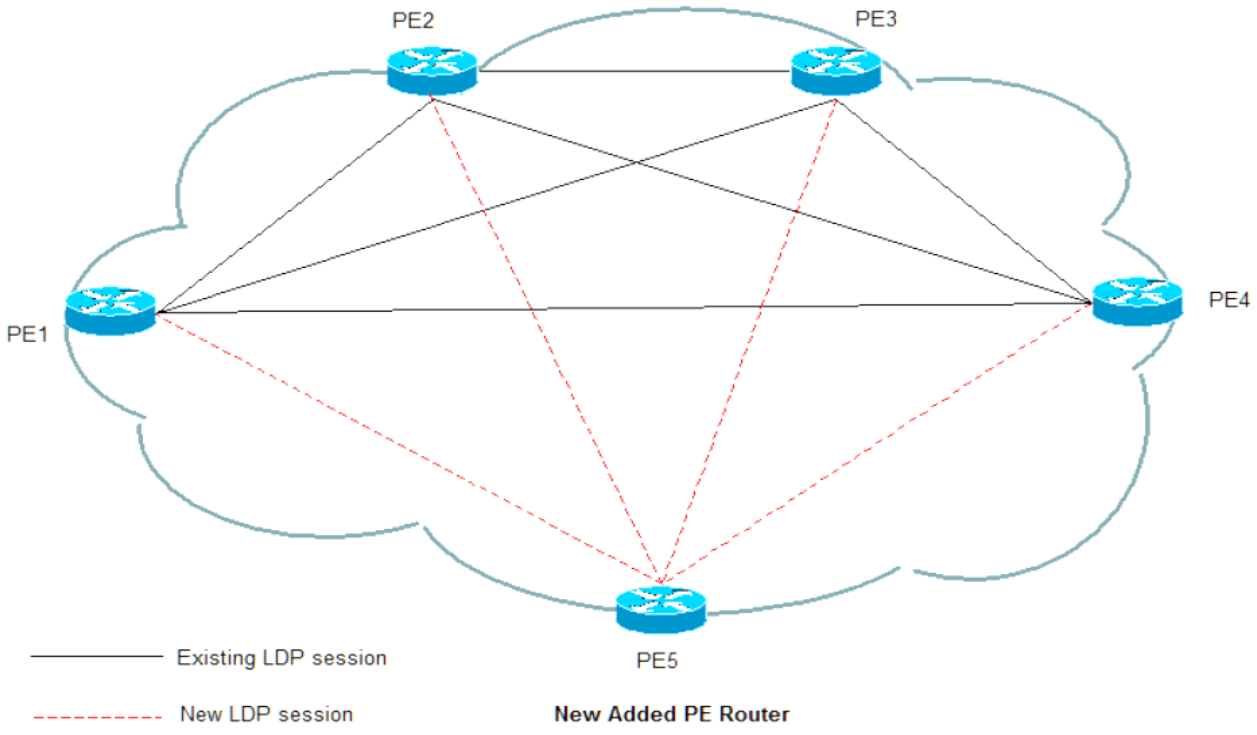
LDP vs BGP for control plane implementation:

Main difference between LDP and BGP schemes is that in case of LDP there should be a full mesh of LDP sessions between PE routers while in case of BGP a full mesh of BGP sessions is not required. This give rise to two issues:

1. A PE router is involved in a large number of targeted LDP sessions that leads to a limit how much a network can grow.

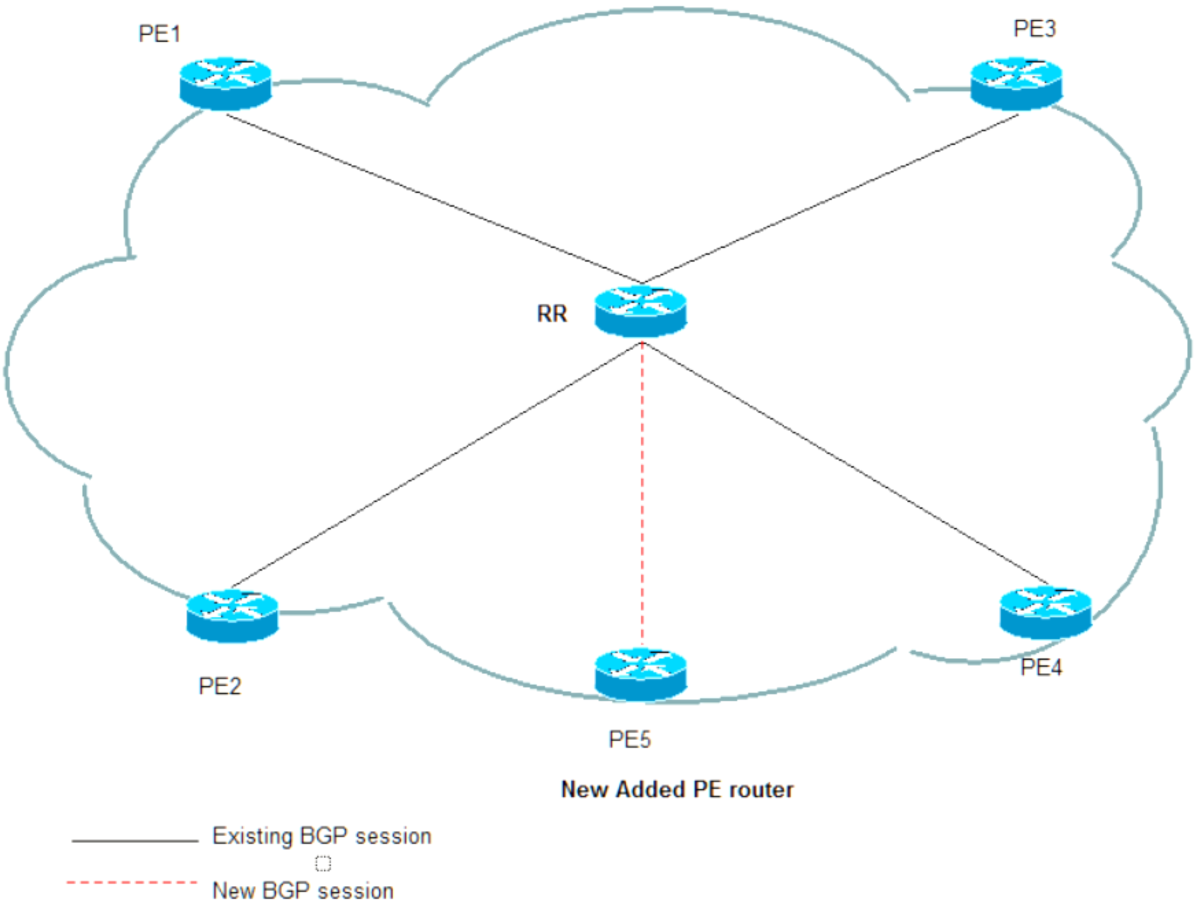
2. Manual configurations of LDP sessions has to be made on each PE router whenever a new PE router is added or an existing PE router is deleted from the network. This is definitely a cumbersome task.

Figure 2. Addition of PE router to an Existing VPLS requires modification on each PE router to accomplish a full mesh of LDP sessions.



The same issues will be faced if a full mesh of BGP sessions is required, to tackle this issue BGP Router Reflector (RR) can be used. It is in use with L3VPN services but can be used for VPLS as well. In that case, only RRs have to be in a full mesh. This makes it simpler when adding a new PE router or deleting an existing one as only the BGP session between the RR and that PE router needs to be modified.

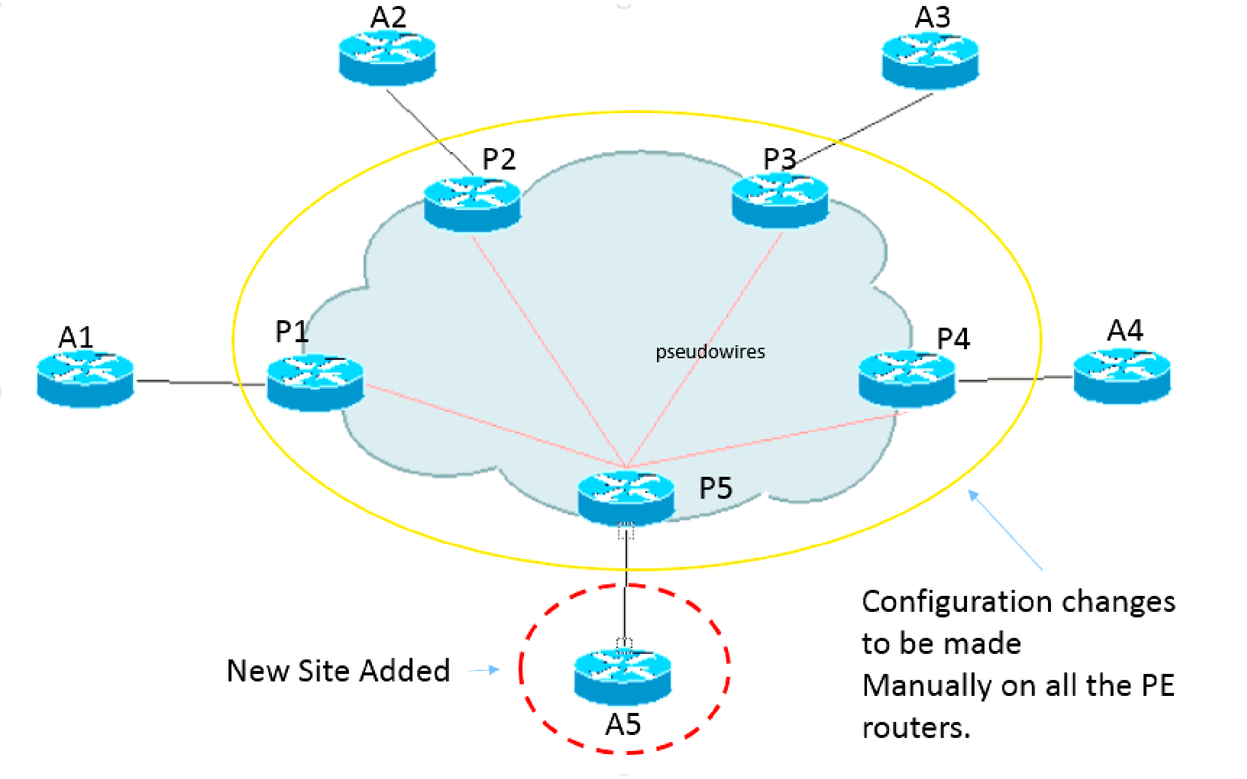
Figure 3. Only BGP session need to set up between the new PE and RR.



In case of LDP VPLS, above scalability issue can be solved by Hierarchical VPLS but at the expense of adding some other issues.

Other Difference includes the auto-discovery of the PEs, in case of LDP there is no such mechanism so if a new PE is added to the network then configurations to be changed on all the existing PE routers to make them in full mesh with the new PE.

Figure 4. In case of LDP VPLS, manual configurations are to be made on all existing PE routers to make them in a full mesh.


A2
A3
P2
P3
A1
P1
pseudowires
P4
A4
P5
Configuration changes to be made Manually on all the PE routers.
New Site Added
A5

An external auto discovery mechanism can be added to the LDP VPLS. There can be three possible combinations:

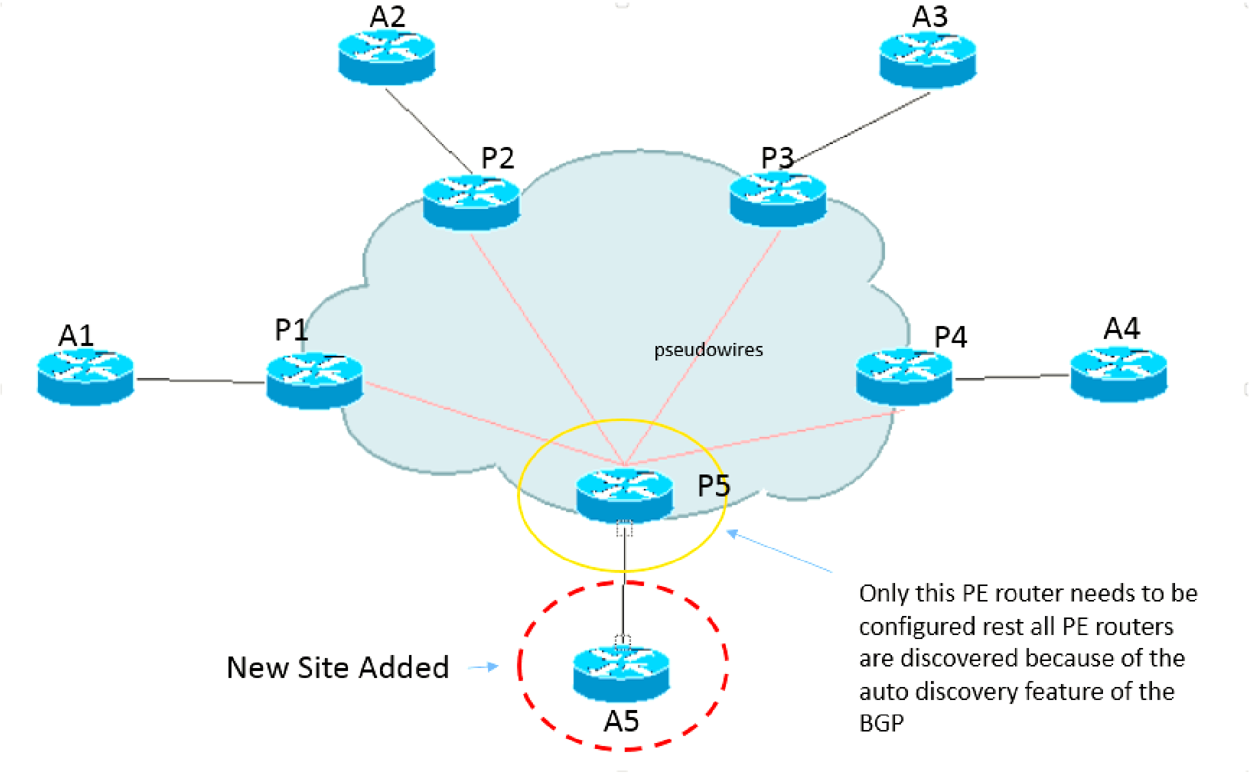
1. LDP+BGP= BGP being used only for auto discovery mechanism while LDP being still used for signaling of pseudowires.

2. RADIUS=RADIUS server being used for auto discovery.PE sends a request specific to VPLS instance and the server replies back with a list of PEs belonging to that VPLS.

3. Extending LDP=No work on this topic by IETF till date.

While in case of BGP there is auto-discovery capability which leads to just configuration changes on the PE to which site is attached.

Figure 5. In case of BGP VPLS, configuration is to be made on only the new PE router rest all existing PE routers are auto discovered.


A2
A3
P2
P3
A1
P1
pseudowires
P4
A4
P5
Only this PE router needs to be configured rest all PE routers are discovered because of the auto discovery feature of the BGP
New Site Added
A5

The main drawbacks of using BGP signaling:

1. Wastage of Bandwidth as Pseudowire Signaling of peer to peer parameters are broadcast to all the PE routers.

2. Label information is broadcast to all PE routers linked with a particular VPLS instance due to a full mesh, this is acceptable for the initial VPLS auto discovery, but subsequent PW discovery is not efficient.

3. BGP signaling use BGP to flush MAC address instead of IEEE Spanning Tree TCNs making is incompatible with IEEE bridges.

(RFC 4761)

Multihoming

For resilience the customer maybe connected to more than one PE, in that case forwarding loops can occur. Like in case of IP forwarding we can use TTL to limit the circulation of a packet, in this case the customer has to use Routers to prevent loops. If the customer chooses to use Ethernet switch as CE then there can be great possibility of loops unless no countermeasure is taken. One of the countermeasure is that customer run STP on all its Switches so there is a loop free topology. But this is not acceptable by the service provider as they have to rely on the customer to implement it correctly, it might affect other customers if not implemented correctly. The other countermeasure is that service provider allow only one port to be active at a time. This can be done in case if we use BGP for signaling, the service provider configure the same VE ID on the PEs connecting to the same CE. By this way each PE receives two advertisement with the same VE ID.BGP applies its selection rules and installs the best NLRI out of them. Suppose if we want the traffic to exit from a specific PE then we can make the local preference higher on that PE than others, this way all PEs choose to install the route advertised by that particular PE. In case of link failure to that PE, other route is installed by PEs. This countermeasure is better than relying on the customer to run STP as in this case bandwidth is saved as broadcast traffic is sent on only one link.

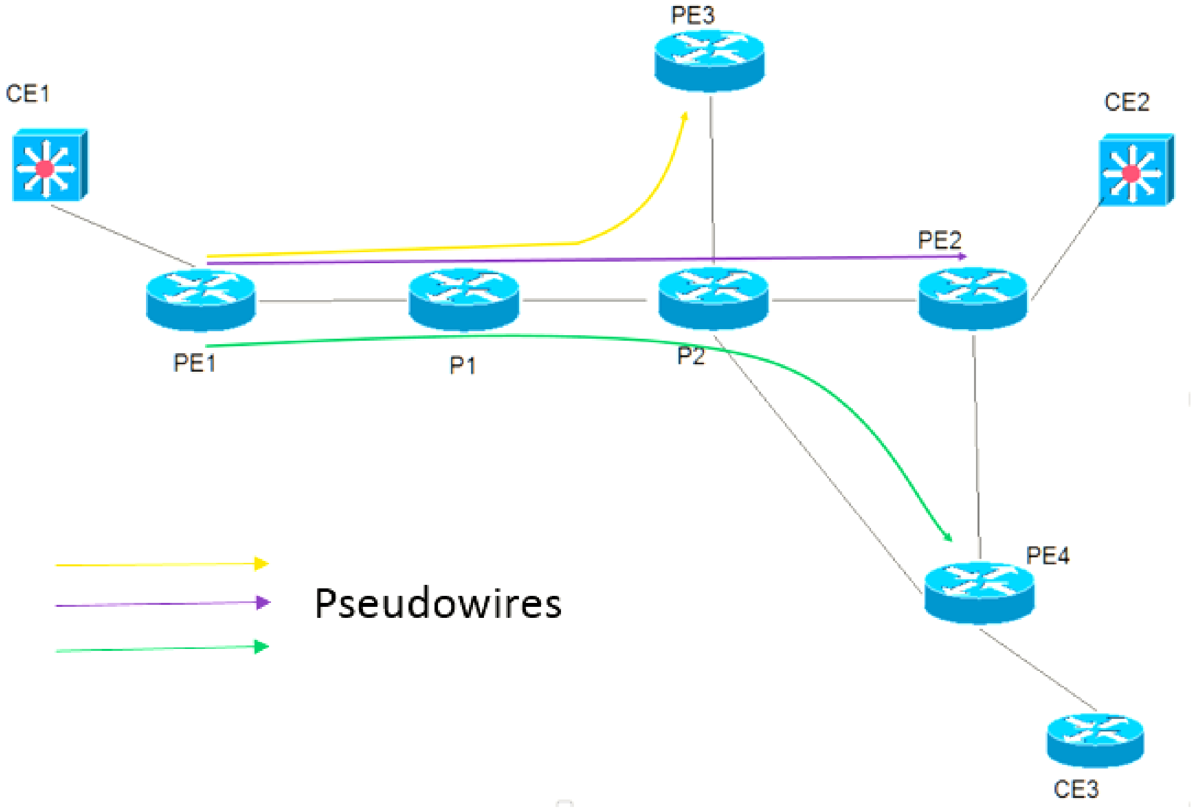
Multihoming with BGP is possible by changing the local preference on the PE but In case of LDP there is no such BGP attribute. The only way of multihoming in LDP case is to implement H-VPLS by deploying a spoke PE on the customer premises, which is homed to multiple-hub PEs.

H-VPLS

Hierarchical version of LDP VPLS can significantly alleviate the pseudowire scalability problem. It is achieved by reducing the number of PE routers that have to been in full mesh, hence improving the control plane scalability. It helps to reduce the signaling and replication overhead to allow large scale deployments.

Taking an example, suppose the physical topology is as shown in the figure below:

Figure 6. A simple VPLS topology.



All PE routers are connected to each other in a full mesh by psuedowires.

Suppose PE1 wants to send 100mb of broadcast data then it will replicate it three times, such that each PE router receives a copy of it. The links PE1-P1 and P1-P2 have to carry 300mb of data which is kind of inefficient usage of the links. With the addition of more PE routers the amount of data which these links carry will linearly increase, hence leading to scalability issues.

Loop problem is solved by using a split horizon rule that means a PE router will only forward traffic to its locally attached hosts not to other PEs, so no need for STP.

To resolve the above discussed problem, H-VPLS uses spoke pseudowire which can carry both mesh and spoke pseudowire traffic so that they can relay traffic between PE routers.

The above VPLS topology can be re-designed using H VPLS in two ways:

Figure 7.1 Logical connection between PE1-PE3, PE1-PE2 and PE2-PE4.

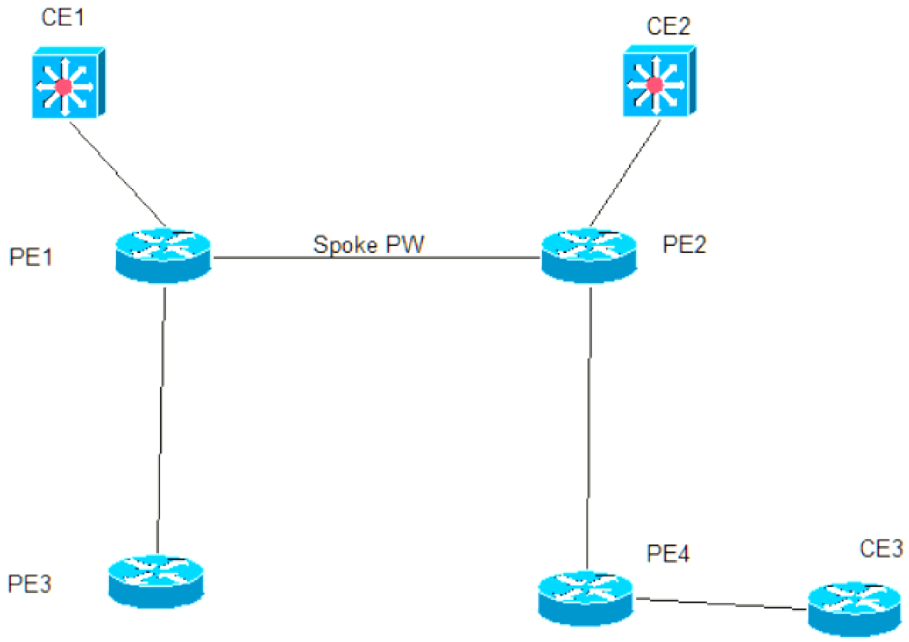
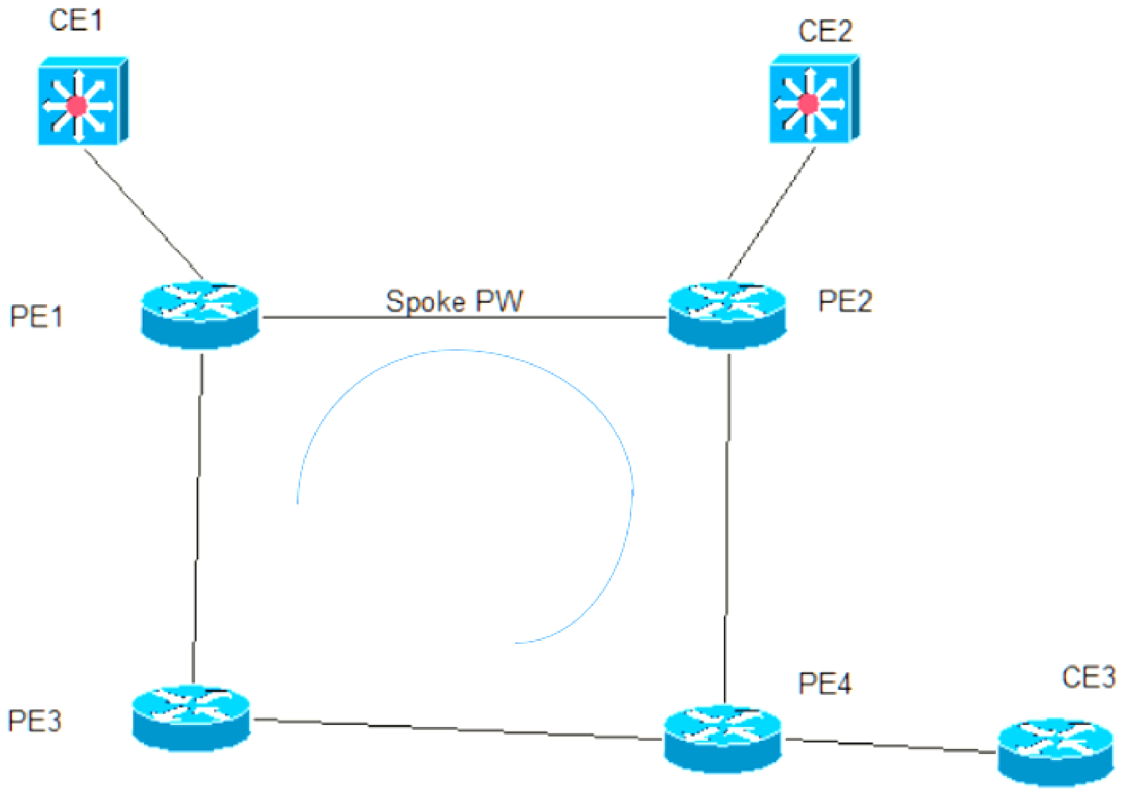


Figure 7.2 Logical connection between PE1-PE3, PE1-PE2, PE2-PE4 and PE3-PE4.



In the Figure 7.1, there is VPLS service defined on PE1 which has two spoke pseudowires connecting to PE2 and PE3 and one link to local circuit. There is no link between PE1 – PE4.

While in Figure 7.2, all the PE routers are connected in a ring by spoke pseudowires.

In H-VPLS, split horizon rule is not followed, any traffic arriving on a spoke pseudowire is forwarded to all other pseudowires (spoke and meshes) and to local circuits. This leads the need to run STP (Spanning Tree Protocol) across the network by Service Provider to avoid loops. This can be considered one of the drawback of Using H-VLPS because the architecture of spokes and meshes has to been planned carefully else there is always chances of loops to occur. Moreover the high convergence time of STP adds on to the problems.

H-VPLS somehow reduces the scalability issue but also adds on some other issues.

Taking into account the topology in the Figure 7.1, Even after using the spoke pseudowires the links PE1-P1, P1-P2 are still carrying two copies of each broadcast traffic, hence not much of bandwidth efficiency improvement. Second issue is the traffic reaching PE4 has to be always relayed by PE2, as there is no direct link between PE1-PE4 leading to congestion on PE2-PE4 link. Moreover in case of failure of PE2, PE4 get disconnected from the other PE routers. Third issue is that PE2 router has to learn all the MAC addresses that fall behind PE1 and PE4 because all traffic from PE1 to PE4 will be terminated at PE2 then it is PE2 router responsible to forward it to PE4 and vice versa.

Taking into account the other H-VPLS architecture i.e. Figure 7.2, all PE routers are connected in a ring so there is a chance of occurrence of loop. STP has to be run to avoid loops. Assuming if STP put the spoke link between PE1-PE2 in block state, now there is no direct link between PE1-PE2 this leads to traffic being relayed through PE4 creating the same issue as discussed above in case of Figure 7.1.

No doubt the advantages of H-VPLS over VPLS are:

- Improves control plane scalability by reducing the number of pseudowires that are to be in a full mesh
- Reduces the burden on core devices by adding a hierarchical aggregation layer.
- Size of MAC table can be reduced if combined with MAC in MAC stacking.

This concludes that H-VPLS provides a solution to improve the pseudowire scalability to some extend but not an efficient way to handle multicast traffic. Additionally H-VPLS brings in other problems along with it.

Operational considerations

The operational issues faced while running VPLS services are:

MAC address scaling-One of the main consideration for service provider is the number of MAC addresses that are to be stored by every PE, along with taking into account a PE might be providing VPLS services to a large number of customers. Service providers have to keep an account of the number of MAC addresses stored for each VPLS customer and impose a limit on it by using VPLS implementations that allow limits on per VPLS basis or per interface basis.

Limiting Broadcast and Multicast Traffic-Another concern for Service providers is to limit the Broadcast and Multicast Traffic as it can be costly to send such traffic especially if there are large number of PE members and there is ingress replication. There are implementations that allow Service providers to keep an account of this type of traffic.

Policing of VPLS traffic-Service providers need to police the amount of traffic being send into the network by the customer on each access port if the VPLS service is being offered over the 100 Mbps or 1 Gbps native Ethernet ports.

3 Discussion

3.1 VPLS Limitations

VPLS has the following limitations that leads to the emergence of a new technology called EVPN:

- Multihoming – Currently VPLS can only support active-standby mode multihoming it does not support active-active redundancy mode multihoming. Multihoming Load balancing across provider edge nodes cannot be provided by VPLS because the MAC address learning happens in data plane.
- Multicast optimization- No solution for leveraging Multipoint to Multipoint LSPs with VPLS, multicast LSPs in conjunction with VPLS is only limited to Point to Multipoint LSPs.
- Provisioning Simplicity- BGP based Auto discovery service can be used to simplify things but still it requires the operator to configure a number of network side parameters on top of access-side Ethernet configuration.
- Flooding of BUM traffic – This is a big issue faced while using VPLS services. Solution required to minimize and localise the amount of flooding of multi-destination frames.
- Convergence Time – Network re-convergence time can be an issue as it is dependent upon the number of MAC addresses in the broadcast domain. Need a solution for having network re-convergence independent of the number MAC addresses.

- Data Center Interconnect – VPLS cannot handle new applications as they require the Layer 2 and Layer 3 services over the same interface. Scalability and control like L3VPN.

In order to meet the above listed challenges, EVPN was introduced in the industry under the L2VPN IEFT Working group.

3.2 EVPN (ETHERNET VPN):

EVPN is the next generation solution that provides L2VPN services over the MPLS network, it inherits a decade of VPLS operational experience in production networks but it differs from VPLS as MAC learning occurs in control plane over the core. MAC information is carried by multiprotocol BGP control plane and provides different choices for Data Plane encapsulation.

EVPN is a significant milestone for the industry as it provides Layer 2 and Layer 3 service in a single VPN with scalability and control of L3VPN.Like VPLS, EVPN allows you to connect different customer sites located at different geographical locations together into a single virtual LAN emulating the end users as if they are connected to the same LAN. Now the service providers are able to meet evolving demands of higher speeds along with sophisticated QoS. Moreover EVPN can support the evolving demands of the new applications which cannot be met by the existing technologies. (RFC 7209)

EVPN BASIC CONCEPTS

EVPN Terminology:

Customer Edge (CE) devices (Switch, Router and Host) connected to Provider Edge devices.

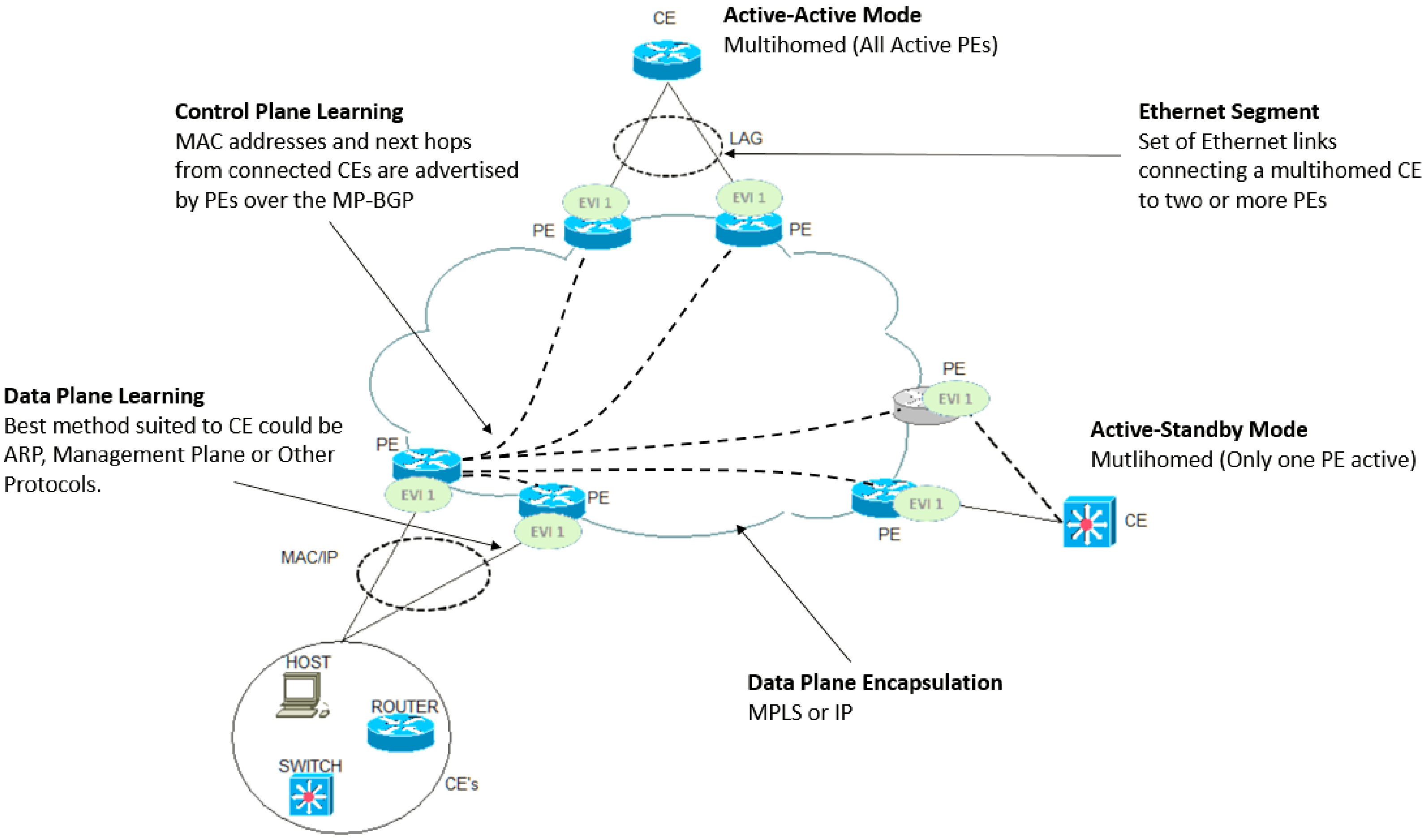
EVI: It identifies a particular EVPN Instance (EVI)

ES: Set of Ethernet links connecting a multihomed device or network to two or more PEs is known as Ethernet Segment (ES)

ESI: Ethernet Segment Identifier (ESI) is a non zero number that identifies a particular Ethernet Segment

Ethernet Tag: It identifies a particular Broadcast or Bridge Domain in the EVI.

Figure 8. EVPN



EVPN Control Plane

MAC learning over Control plane provides greater control of who learns what, provides ability to apply policies and helps to maintain isolation and virtualization of EVPN instances. PEs advertise MAC addresses and IPs for next hop to other PEs over the MP-BGP with an EVPN NLRI. In addition to load balancing over multiple LSPs between the same PEs, control plane learning also provides load balancing of traffic between the CEs that are multihomed to number of PEs. This helps improve convergence time in case of one of the CE-PE link failure.

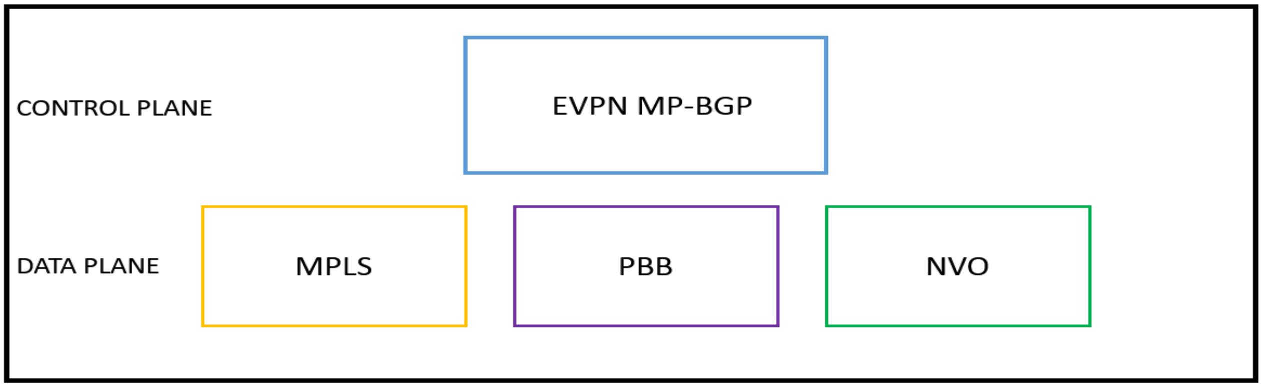
Though MAC address learning over the MPLS core happens in control plane, but still learning between PEs and CEs is done by method preferred by the CE, it could be data plane learning, 802.1 aq, ARP or other protocols.

It can be locally decided whether the PE L2 forwarding table should include all the MAC addresses known to the control plane or it should include only the MAC addresses of the active flows or it can also implement cache based scheme.

EVPN policy attributes are almost similar to those of L3VPN. Each EVI needs a unique RD per PE and one or more globally unique RTs

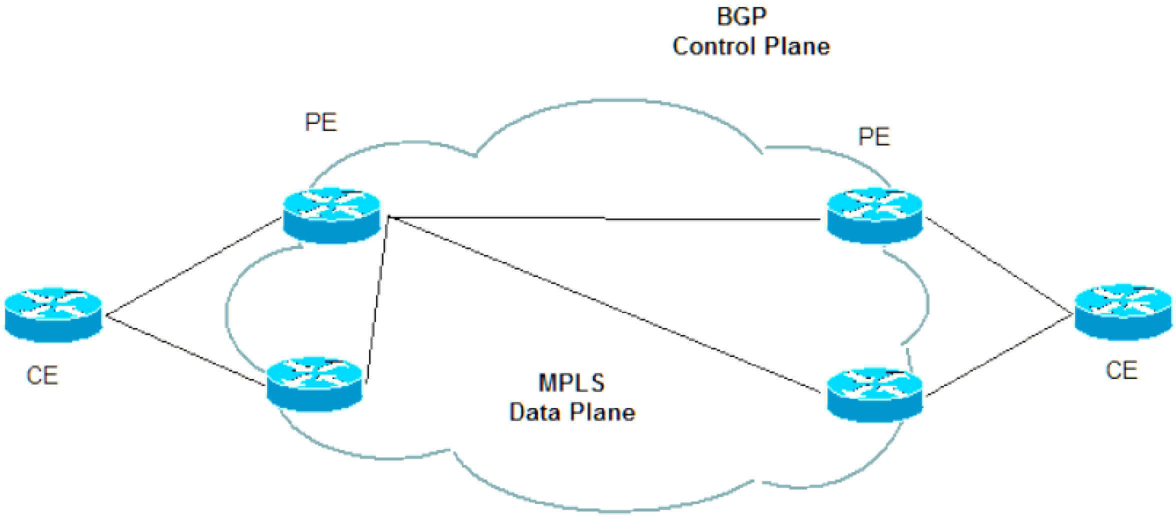
EVPN Data Plane

EVPN abstracts and separates the control plane and the data plane. Multiprotocol BGP as the control plane with different data plane encapsulation choices can be as shown in figure below:



EVPN-MPLS Data Plane

Figure 9. EVPN-MPLS Data Plane

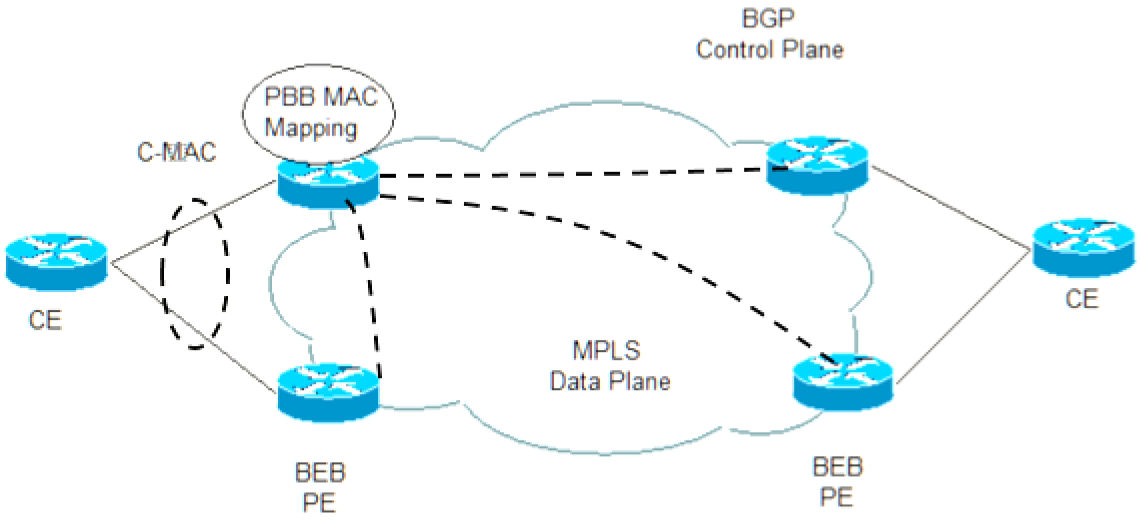


- EVPN-MPLS Data plane is the original EVPN solution in the base specification, it provides a simple method of implementing EVPN over an existing MPLS core.
- MPLS runs in the core networks control plane and data plane providing all the MPLS features.

- Provides all active multihoming for Virtual Private Wire Service (VPWS).
- It requires no pseudowires.
- IGP, RSVP-TE or LDP is required for MPLS and BGP for EVPN.

Provider Backbone Bridge (PBB) – EVPN Data Plane

Figure 10. PBB-EVPN Data Plane

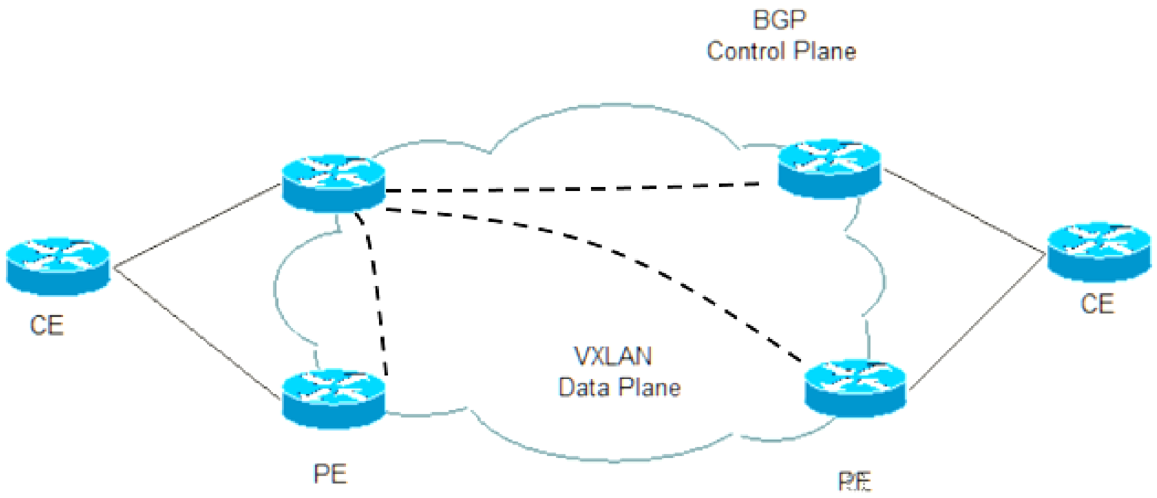


- PBB-EVPN combines IEEE 802.1ah PBB with EVPN to support scaling of very large networks with all active multihoming over MPLS.
- It lowers the number of MAC addresses by aggregrating customer MACs with backbone MACs, same concept as of route aggregation in IP.

- Backbone Edge Bridges (BEB) PEs only advertise backbone MACs with BGP, while customer MAC and backbone MAC mapping is learned in the data plane.
- MPLS runs in the control plane and the data plane.
- This topology or architecture can be useful where the number of MAC addresses are too large as this hides the customer MACs from the backbone elevating to high MAC scalability.

Network Virtualization Overlay

Figure 11. NVO



- EVPN over NVO tunnels (VXLAN,NVGRE,MPLSoGRE) provides L2 and L3 Data Center Interconnect and resilience to simple IP networks.
- EVPN over Virtual Extensible LAN (VXLAN) Data Plane can be used in place of MPLS Data plane when MPLS is not available in the core network.

- EVPN-VXLAN can be used to provide an L2 overlay over an IP network. It is quite flexible, VXLAN can be routable with IP irrespective of the underlying network being used.
- VXLAN Data plane encapsulates VXLAN header and L2 Frame using UDP and can run over IPv4 or IPv6 while EVPN uses BGP Control Plane to advertise MAC routes.
- Possible to provide a VPN to a hypervisor attached to a Virtual Machine, as the VXLAN tunnel endpoints can be on Virtual Machines.
- This architecture can provide EVPN services to DCI and virtual network without requiring MPLS.

Ethernet Segment

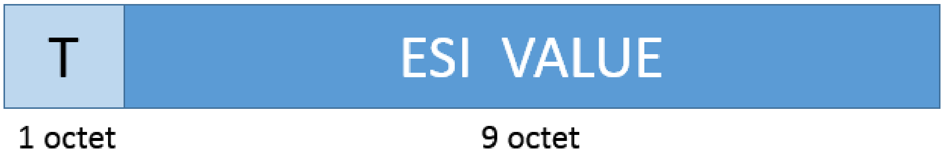
If a Customer Edge (CE) is multihomed to two or more Provider Edge (PEs), the set of Ethernet links that attach CE to PEs is known as Ethernet Segment (ES). To CE an ES seems to be a Link Aggregation Group (LAG). Each Ethernet Segment has a unique non zero identifier called the Ethernet Segment Identifier (ESI). ESI is a ten octet integer in line format sent with the most significant octet first.

The value of the ESI has to be unique and non-reserved across all the EVPN instances on a PE. With the managed CE the ESI uniqueness should be guaranteed by the network operator but if the CE is not managed by the network operator then a network wide unique ESI has to be configured for that ES. The uniqueness helps the auto discovery of the ES and DF election.

The two reserved values of ESI are:

- ESI 0 represents a single homed CE i.e. CE is linked to only one PE.
- Maximum ESI value 0xFF (10 times) is reserved.

ESI has the following format:



T is a 1 octet field (most significant octet) that represents the type of ESI. It specifies the next 9 octets i.e. the ESI value.

There are six types of ESI:

- Type 0 (T=0x00) It indicates an arbitrary ESI value which is configured by the network operator.
- Type 1 (T=0x01) this indicates auto generated ESI value by LACP when IEEE 802.1AX LACP is used between CEs and PEs.
- Type 2 (T=0x02) this is used in case of when there is indirectly connected hosts (bridged LAN). The ESI value is automatically generated by the L2 bridge protocol.
- Type 3 (T=0x03) this is MAC based ESI value that is auto generated or configured by the network operator.
- Type 4 (T=0x04) this is Router-ID based ESI value that is auto generated or configured by the network operator.
- Type 5 (T=0x05) this is AS based ESI value that is auto generated or configured by the network operator.

Ethernet Tag

Ethernet Tag denotes a particular Broadcast domain such as VLAN. It is possible to have one or more broadcast domains in a particular EVPN. Ethernet Tag ID is a 32 bit field that contains 24 bit or 12 bit of the identifier. The 12 bit identifier is also called the VLAN ID (VID). Service providers assign VLANs to a particular EVPN. A particular VLAN may contain more than one VID. In case of multiple VIDs within a VLAN the participating PEs in that VLAN for a particular EVPN are responsible for VID translation to and from the attached CEs. There is no need of VID translation on PEs if there is only one VID in that VLAN. There are deployment scenarios that have unique VID across all EVPN instances and there are scenarios where all points of attachment for a particular EVPN instance use the same VID, no other EVPN instance use that same VID. RTs for each EVPN instance are automatically derived from the corresponding VID.

For an EVPN instance, each PE performs a mapping of Ethernet Tag and broadcast domain identifier (VID).

The following relationships can exist between VLANs, Ethernet Tag IDs and MAC-VRFs:

VLAN based Service Interface:

There is only one broadcast domain (VLAN) associated with the EVPN instance, so there is one to one mapping of VID and EVI. VID translation is allowed, if the VLAN consists of multiple VIDs i.e. different VID per Ethernet Segment per PE, then each PE needs to translate VID for the frames destined to its Ethernet Segments. Overlapping of MAC addresses across different VLANs is possible. Ethernet Tag ID in all EVPN routes is set to zero.

VLAN Bundle Service Interface:

In this type of service interface there are multiple broadcast domains associated with a single EVPN instance, so there is many to 1 mapping of VID and EVI. Multiple VLANs share the same bridge domain. MAC addresses are to be unique across VLANs. VID translation is not allowed. Ethernet Tag ID in all EVPN routes is set to zero.

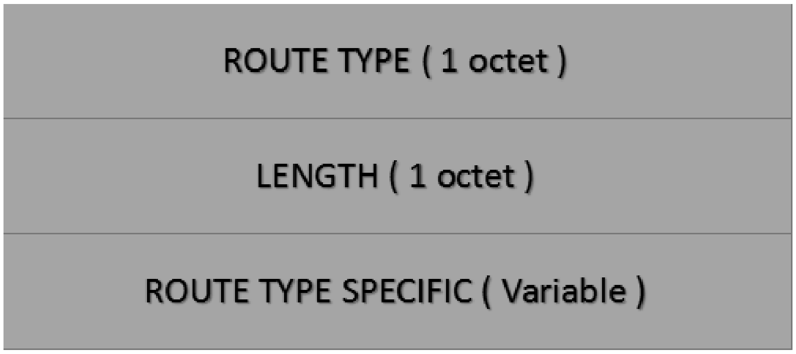
VLAN Aware Bundle Service Interface:

In this type of Service interface there are multiple broadcast domains associated with a single EVPN instance with each VLAN having its own bridge domain. In this case VID translation is allowed and Ethernet Tag ID is assigned by Service Provider.

BGP EVPN routes:

As we know EVPN uses Control Plane for MAC learning so it introduces a new BGP NLRI (Network Layer Reachability Information) called EVPN NLRI.

EVPN NLRI format:



ROUTE TYPE (1 octet)
LENGTH (1 octet)
ROUTE TYPE SPECIFIC (Variable)

The Route type describes the Route Type Specific EVPN NLRI. While the Length field denotes the number of octets of the Route Type Specific field.

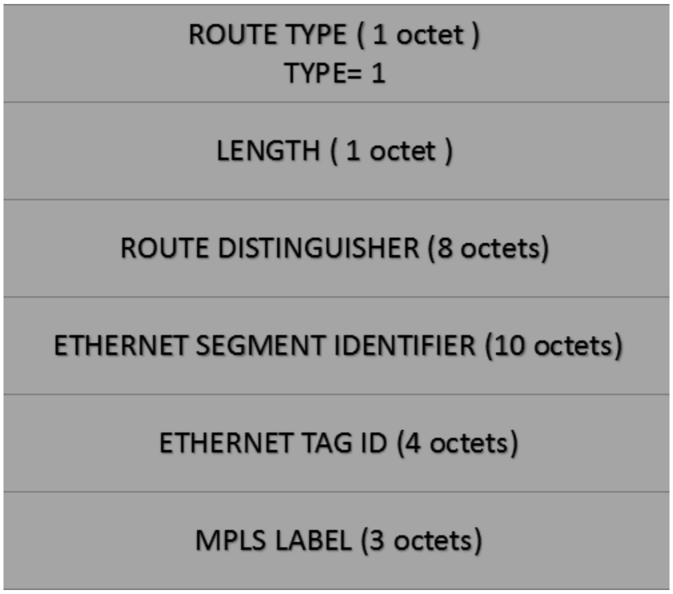
Multiprotocol BGP extensions AFI (Address Family Identifier) of 25 and SAFI (Subsequent Address Family Identifier) of 70 is used to carry the EVPN NLRI. The Two BGP speakers must use the BGP capabilities Advertisement to properly process this EVPN NLRI.

Route types are as follows:

- **Ethernet Auto Discovery(A-D) Route**

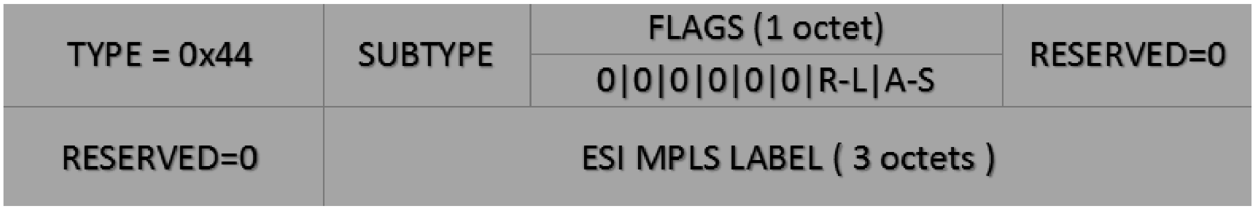
As we already know MAC addresses are learned by BGP control plane over MPLS core in case of EVPN, so in case of failure the convergence time is function of MAC address advertisement routes that a PE router needs to withdraw. This can be an issue for large network as the number of MAC addresses would be high leading to slow convergence. To overcome this issue EVPN introduces a mechanism using new route types to signal the PE routers to update their forwarding tables. This (**FAST CONVERGENCE**) happens by making each PE router advertise the Ethernet Auto Discovery route type per Ethernet Segment to all remote PE routers. When there is a connectivity failure to an attached segment, then the PE withdraws the corresponding Ethernet Auto Discovery route triggering each remote PE router to update their forwarding table for the MAC addresses associated to that Ethernet Segment and if there is any backup PE available for the same Ethernet Segment then all the remote PE router update their next hop for all the MAC addresses associated with that Ethernet Segment if no such backup is there then it simply invalidates those MAC address entries for that Particular ES.

Fig. Ethernet A-D route type



ROUTE TYPE (1 octet) TYPE= 1
LENGTH (1 octet)
ROUTE DISTINGUISHER (8 octets)
ETHERNET SEGMENT IDENTIFIER (10 octets)
ETHERNET TAG ID (4 octets)
MPLS LABEL (3 octets)

Fig. ESI MPLS Extended Community



<table>
<tr><td rowspan="2">TYPE = 0x44</td><td rowspan="2">SUBTYPE</td><td>FLAGS (1 octet)</td><td rowspan="2">RESERVED=0</td></tr>
<tr><td>0|0|0|0|0|0|R-L|A-S</td></tr>
<tr><td>RESERVED=0</td><td colspan="3">ESI MPLS LABEL (3 octets)</td></tr>
</table>

ESI MPLS Extended community is also advertised along with the Ethernet Auto Discovery Route.

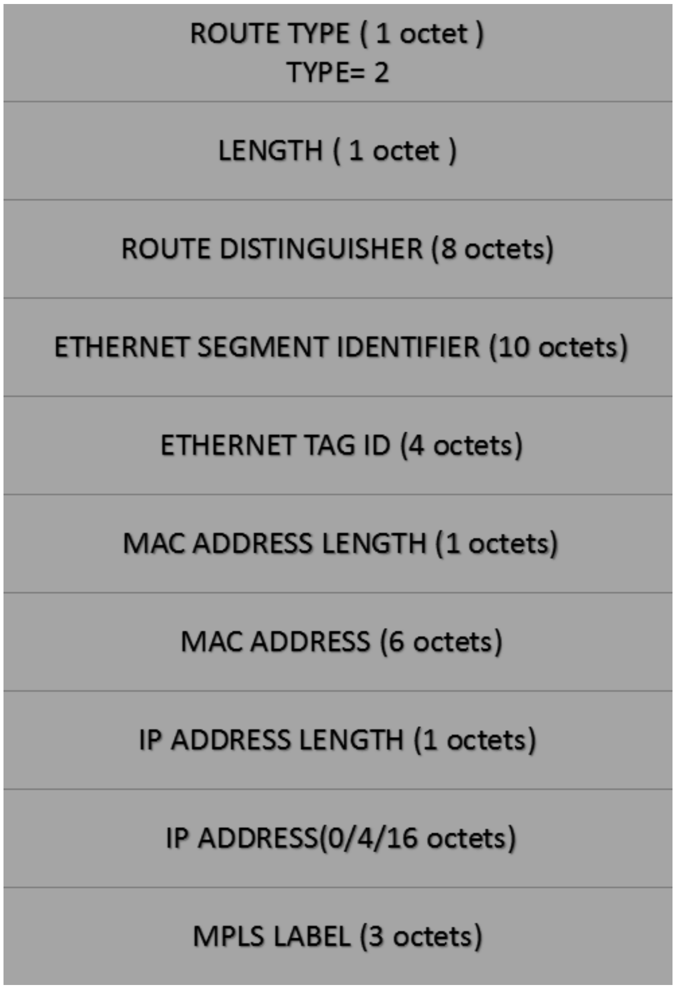
The last two flags are:

R-L for Root/Leaf

A-S for Active/Standby, for single-active mode it is 1 and for all-active mode it is 0.

- **MAC/IP Advertisement Route**

As we know the MAC address learning between CE and PE routers happens in Data Plane. The MAC addresses behind CE are advertised by PE router to other PEs in BGP NLRI using a MAC Advertisement Route Type as shown in Figure below:



ROUTE TYPE (1 octet)
TYPE= 2

LENGTH (1 octet)

ROUTE DISTINGUISHER (8 octets)

ETHERNET SEGMENT IDENTIFIER (10 octets)

ETHERNET TAG ID (4 octets)

MAC ADDRESS LENGTH (1 octets)

MAC ADDRESS (6 octets)

IP ADDRESS LENGTH (1 octets)

IP ADDRESS(0/4/16 octets)

MPLS LABEL (3 octets)

This Route Type is used by the PE router to advertise to remote PEs the MAC addresses that were learned locally. MAC address aggregation can be possible and instead of advertising each MAC address individually, MAC prefix can be used. When MAC prefix is advertised IP ADDRESS LENGTH field is set to 0 and IP ADDRESS field is left empty. But if each MAC address is advertised individually then IP ADDRESS field corresponds to that MAC address. If a PE router receives an ARP request for an IP address from a CE router, the PE router checks out the IP address and if there is an MAC address binding for that IP address it acts as ARP proxy and responds back to the ARP request.

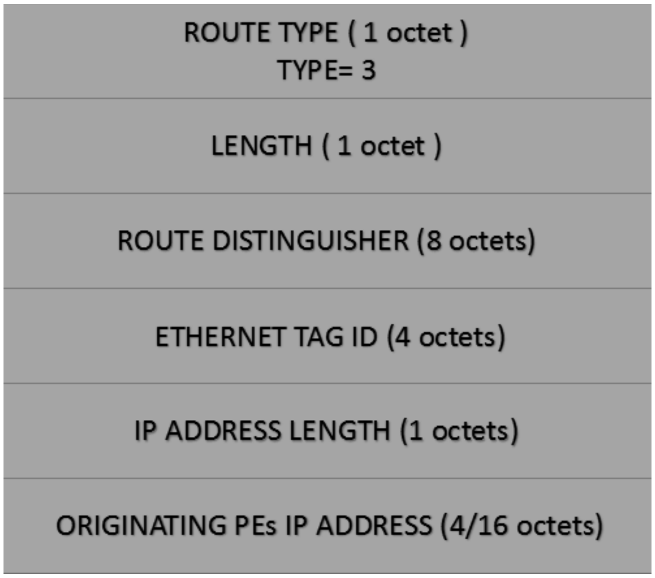
MPLS Label field is of three octets, it depends on the following procedures:

- PE router may advertise the same MPLS Label for all MAC addresses in that particular EVI
- PE router may advertise a unique MPLS label per (ESI, Ethernet Tag) combination
- PE router may advertise a unique MPLS label for each MAC address

- **Inclusive Multicast Ethernet Tag Route**

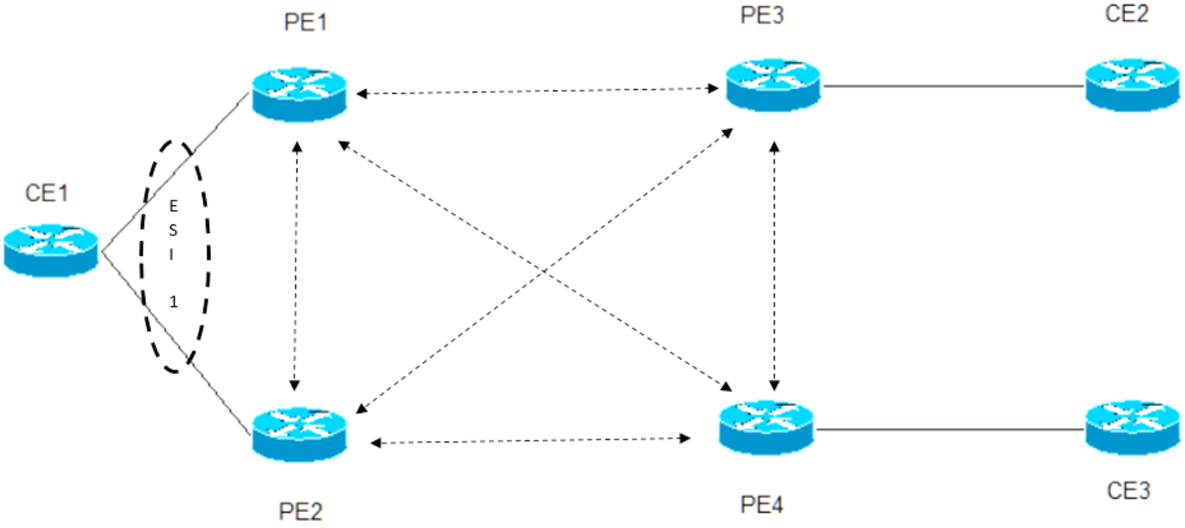
Inclusive Multicast Ethernet Tag Route Type is advertised by a PE router to all remote PE routers, this assists the PE router to forward multi destination traffic to all remote PEs.

The format for Inclusive Multicast Ethernet Tag Route Type is as shown below:



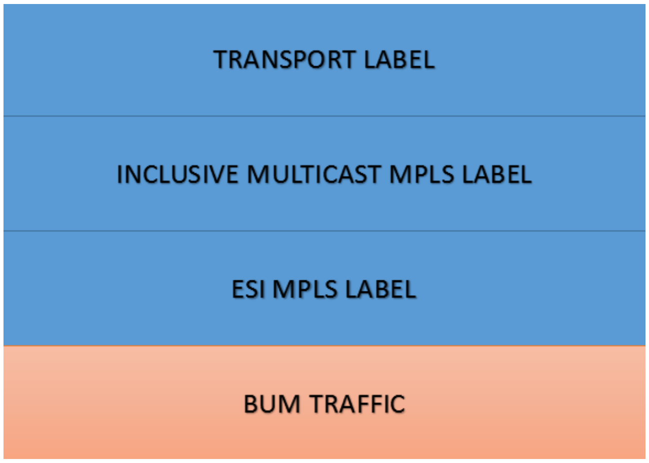
ROUTE TYPE (1 octet) TYPE= 3
LENGTH (1 octet)
ROUTE DISTINGUISHER (8 octets)
ETHERNET TAG ID (4 octets)
IP ADDRESS LENGTH (1 octets)
ORIGINATING PEs IP ADDRESS (4/16 octets)

Suppose for a given scenario as below:



Every PE router advertises Inclusive Multicast Ethernet Tag Route Type for that particular EVI to all other PE routers.

Suppose PE2 gets BUM packet from CE1 then it will encapsulate it as shown:



After encapsulating, it will send it to PE1 only as PE3 and PE4 do not advertise Ethernet Auto Discovery routes. The encapsulation at PE2 occurs with adding of ESI MPLS label received from PE1 on top of the BUM traffic and then on top of ESI MPLS label is added Inclusive Multicast MPLS label and Transport label being finally at the top. This route type helps with split horizon as when PE1 router receives encapsulated traffic, by checking out the ESI MPLS Label is the same as it forwarded then it does not forward it to the ESI.

- **Ethernet Segment Route**

On the basis of the information in the Ethernet Segment Route a Designated Forwarder is elected. Ethernet Segment Route type helps in Service carving. Service carving is the procedure of electing multiple Designated Forwarders per Ethernet Segment (one per EVI) to perform load balancing of multi destination traffic destined for a segment. Election of every PE is made in a way that it is Designated Forwarder for a disjoint set of EVIs.

Ethernet Segment Route format is a below:

ROUTE TYPE (1 octet) TYPE= 4
LENGTH (1 octet)
ROUTE DISTINGUISHER (8 octets)
ETHERNET SEGMENT ID (10 octets)
IP ADDRESS LENGTH (1 octets)
ORIGINATING PEs IP ADDRESS (4/16 octets)

ES Import Extended community format:

TYPE = 0x44	SUB TYPE	ES-IMPORT
ES IMPORT (6 Octets)		

Service Carving mechanism:

- Ethernet Segment Route along with the ES import extended community attribute is advertised by the PE router on the discovery of ESI of the attached Ethernet Segment.
- PE router sets a timer with default value 3 seconds (This value has to be same across all PEs connected to that ES) to receive Ethernet Segment routes from all other PEs connected to the same Ethernet Segment.
- With the expiry of the timer each PE creates an ordered list of IP addresses of the PEs connected to that ES in ascending numeric value. These IPs are obtained from the

ORIGINATING PEs IP ADDRESS field in the Ethernet Segment Route. On the basis of numeric value of IP addresses an ordinal number is assigned to each PE with 0 ordinal being representing the IP address with lowest numeric value. The PE router with the highest ordinal number is elected the Designated Forwarder and the next highest ordinal number is elected as the Backup Designated Forwarder

- The Designated Forwarder elected for that particular EVI unblocks all traffic for the Ethernet Tags of that EVI.

In the scenario of a link/port failure, the affected PE withdraws its Ethernet Segment Route retriggering the service carving procedure on all PEs.

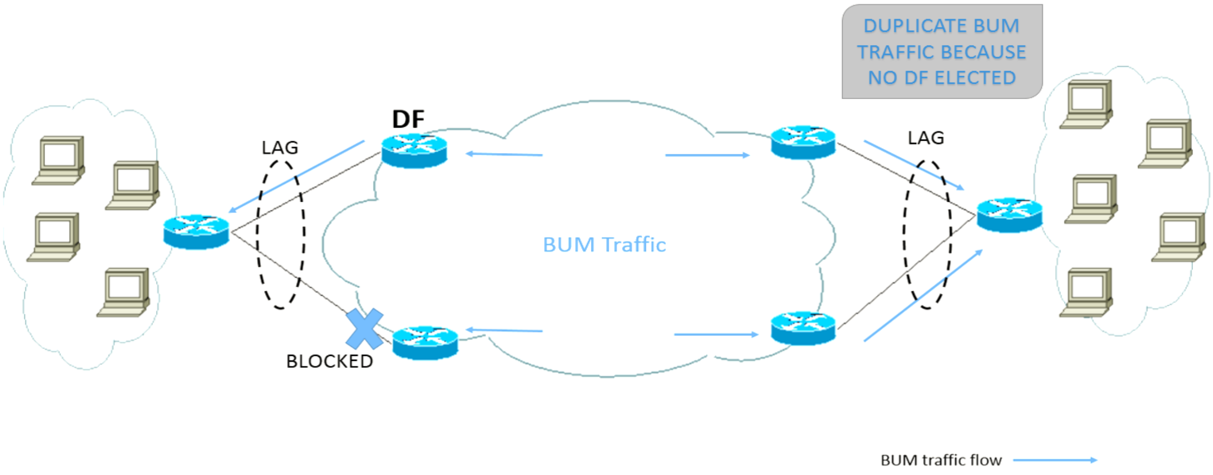
3.3 EVPN KEY FEATURES

Unlike the traditional VPLS which rely only on Data plane learning, EVPN uses control plane for learning which adds on new features and functions overcoming the VPLS limitations.

Following are the control plane key features:

- All Active Multihoming and Designated Forwarder

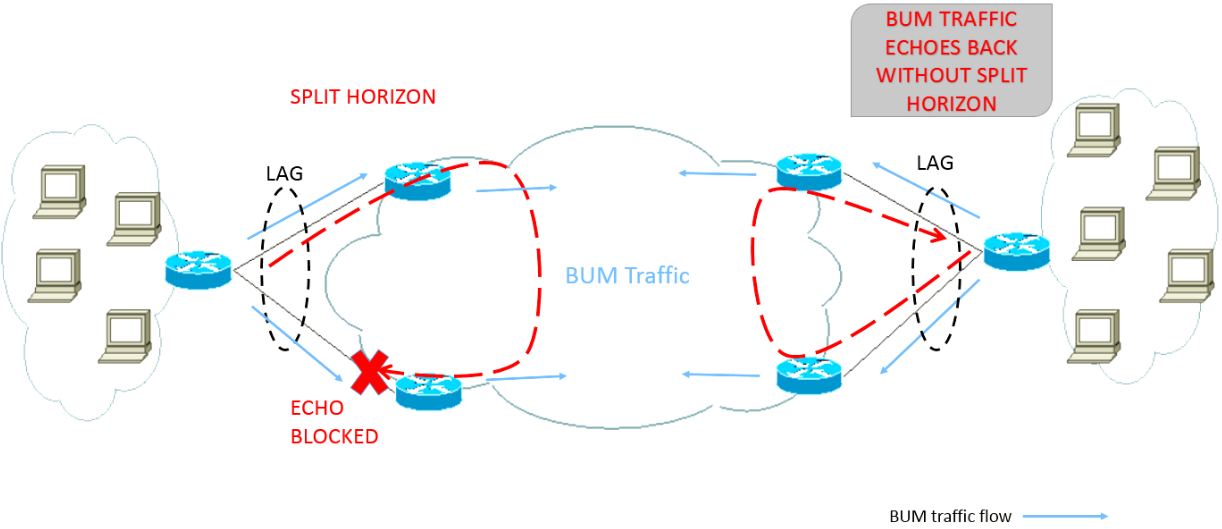
Figure 12.



For all active redundancy mode, the bridged network is connected to two or more PEs using Link Aggregation Group. One of the PE is elected as Designated Forwarder based on the information in the Ethernet Segment Route using the Service Carving procedure. All the BUM (Broadcast, Unknown Unicast, Multicast) traffic is forwarded by only the DF to Ethernet Segment towards the CE as shown in the scenario above, it avoids duplicate flooding of BUM traffic to all active CEs. Other PEs block BUM traffic to CE.

- All Active Multihoming and Split horizon

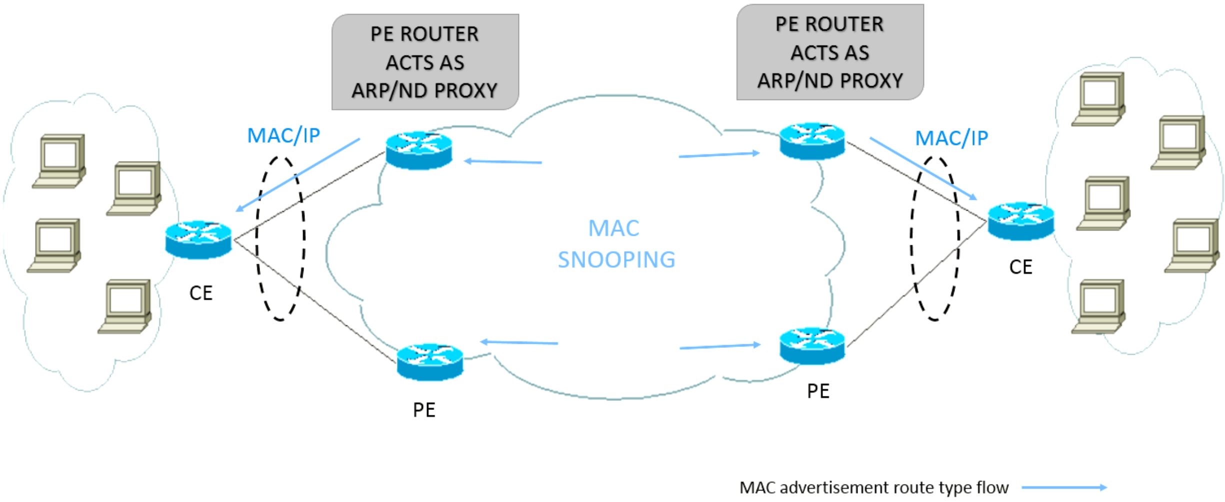
Figure 13.



To achieve the split horizon function EVPN uses the Ethernet Auto Discovery Route Type per Ethernet Segment, it contains the ESI label which helps identify the Ethernet Segment this traffic originated from. This is achieved by encapsulating every BUM traffic originating from a non-DF PE with an MPLS label that identifies the Ethernet Segment where the traffic originated. Egress PE use this label to filter out the BUM traffic that has the same ESI label to where it is destined.

- ARP and ND

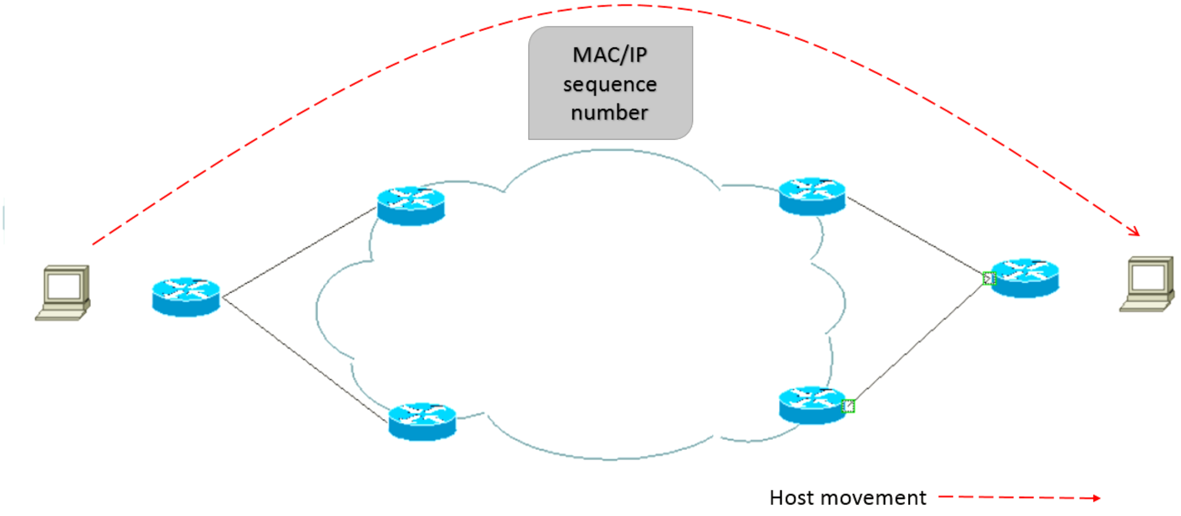
Figure 14



ARP and ND (Neighbour Discovery) flooding over the MPLS network can be controlled by adding the associated IP address to the MAC advertisement route type. When a PE learns the IP address associated with a MAC address of the locally attached CE it can advertise it to other PEs by adding it into the MAC advertisement Route. If there are more than one IP address associated with that MAC address then, the number of MAC advertisement route generated are equal to the number of IP addresses. Whenever there is an ARP request for an IP address, PE checks out it MAC address binding for that IP address, If found it responds back to the ARP request acting as an ARP proxy.

- MAC mobility

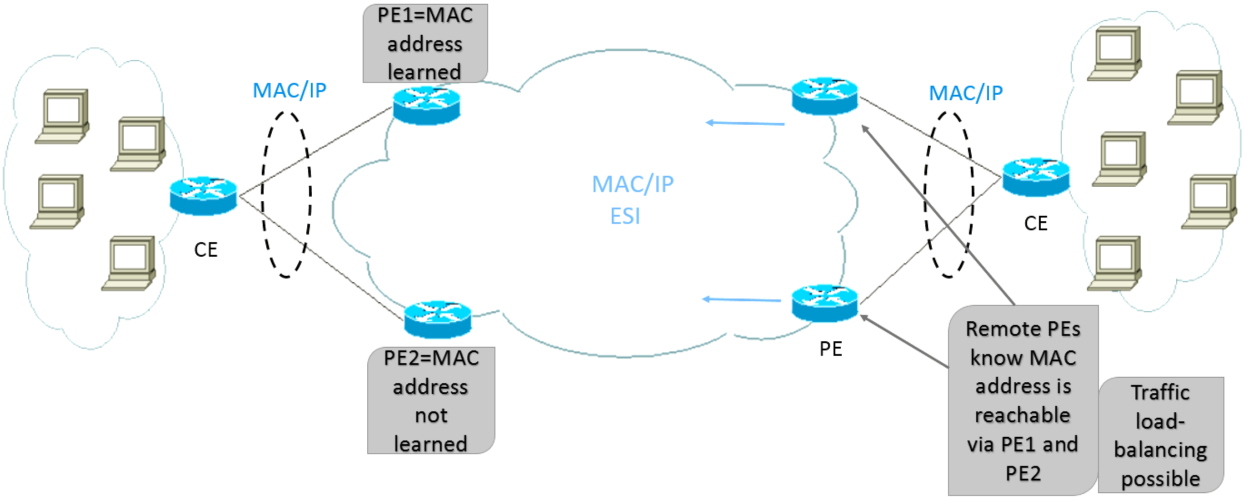
Figure 15.



With EVPN it is possible to move a host from one Ethernet segment to another using MAC advertisement route. There would be two MAC advertisements, one with the previous Ethernet Segment and one with the new Ethernet Segment. Whenever a MAC move happens all the PEs withdraw the MAC advertisement with the previous Ethernet Segment. With local learning in data plane it is not detectable that a MAC move has happened. Each MAC advertisement has a sequence number with it that helps the PE routers to update reachability to that MAC address. MAC advertisement route with highest sequence number is saved and the lower one is withdrawn.

- Aliasing and Backup Path

  Figure 16.


PE1=MAC address learned
MAC/IP
MAC/IP
CE
MAC/IP
ESI
CE
PE
PE2=MAC address not learned
Remote PEs know MAC address is reachable via PE1 and PE2
Traffic load-balancing possible

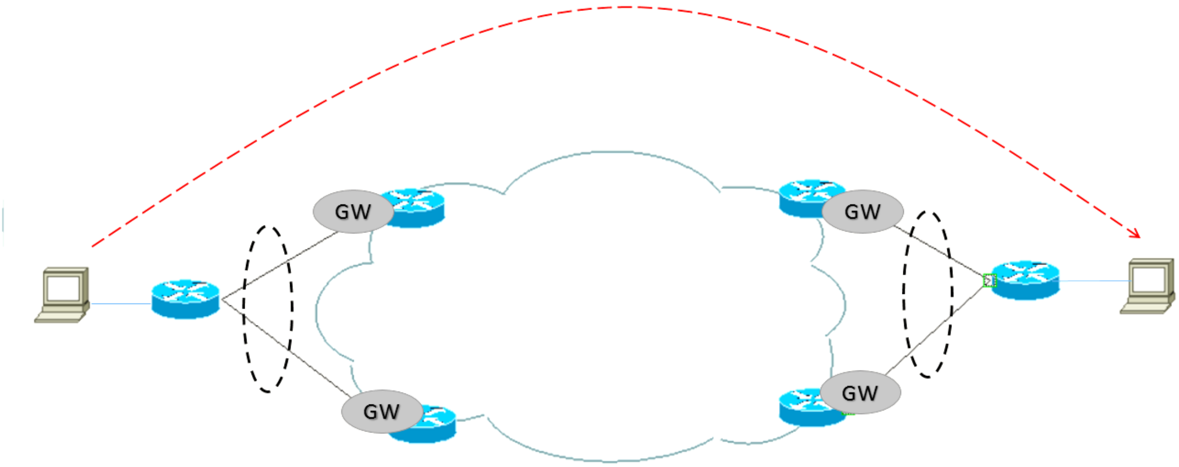
In case of a CE multihomed to two PEs, in all active redundancy mode using LAG, only one of the PE learns the MAC addresses of that CE. This leads to an issue where only one PE advertises the MAC advertisement routes for these MAC addresses to other remote PEs making inefficient for the remote PEs to load balance traffic among the two PEs attached to multhihomed Ethernet Segment.

EVPN comes with a solution “Aliasing”. It allows a PE to advertise reachability to an EVI/ES even when it has learned no MAC addresses from that EVI/ES, making possible for remote PEs to load balance traffic among all PEs advertising it.

In case of Single Active redundancy mode, it can also be used as backup path for that MAC address.

- Default Gateway

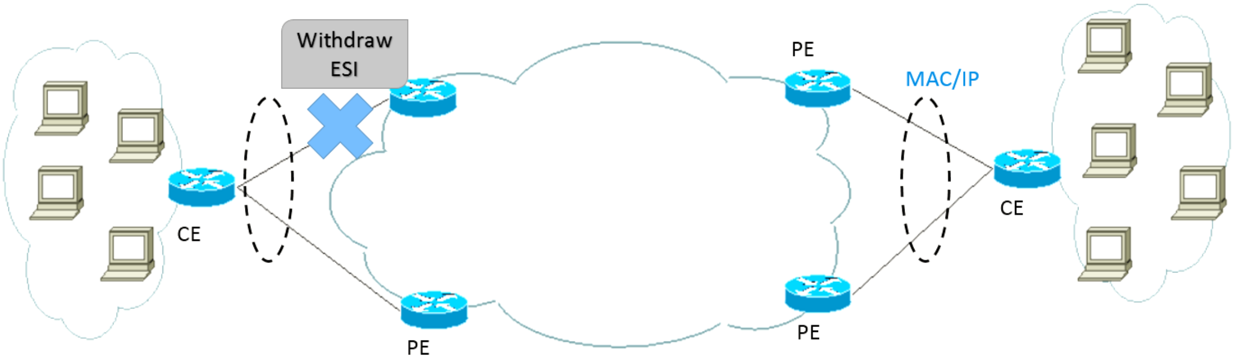
Figure 17.



Inter-subnet forwarding is possible with EVPN. PE router that performs this is called the default gateway for that EVI, it responds to the ARP request for the IP address that is configured as the default gateway address. Every PE that acts as default gateway advertises its MAC address in the EVPN control plane in the MAC advertisement route along with the Default Gateway extended community. The IP address field in the MAC advertisement route is set to Default gateway IP address for that EVI, For an EVI the default gateway IP address is same across all the PE routers in that EVI. Each PE compares the received IP address in the MAC advertisement route with the default gateway IP address it already contains, and if there is any difference PE reports an error message.

- MAC Mass Withdraw

  Figure 18.

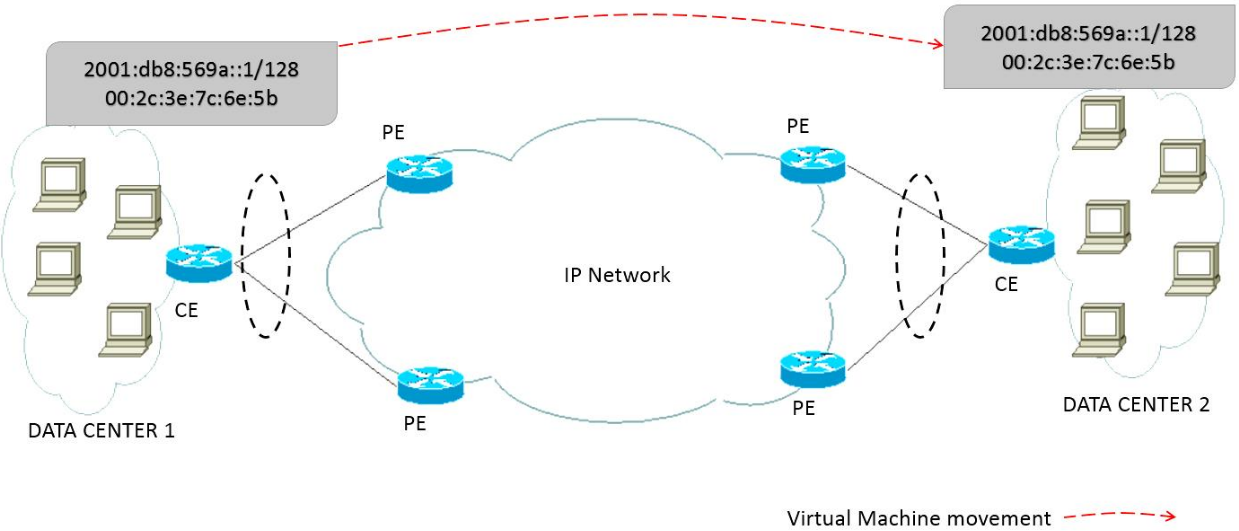


In case of a Link failure, EVPN provides fast convergence even if there are large number of MAC addresses associated with that. Each PE advertises two type of routes i.e. MAC/IP address and its ESI, the other one being connectivity to ESIs. Whenever a Failure occurs instead of withdrawing MAC addresses individually, the PE simply just removes the route for that ESI. Remote PEs remove the failed PE from the path that leads to all the MAC addresses linked to the ESI. This mass MAC removal leads to fast convergence in case of a link failure.

3.4 EVPN Applications

For today EVPN can offer the following services:

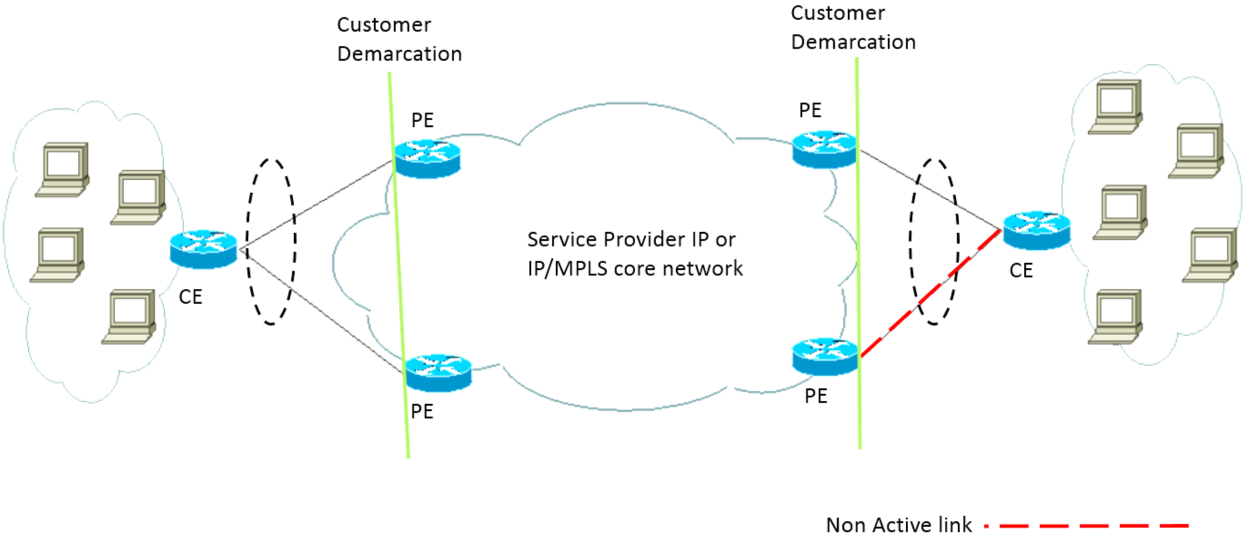
- **Layer 2 or Layer 3 Data Center Interconnect:**
  Figure 19.



  - Provides scalable L2 or L3 DCI services for Virtualized Data Centers.
  - Control Plane signaling of IP/MAC provides mobility to Virtual Machines to move between the data centers.
  - Local IP gateways at each PE router optimize routing.
  - Integrated L2 switching and L3 routing over the same service interface.

- **Layer 2 and Layer 3 services**

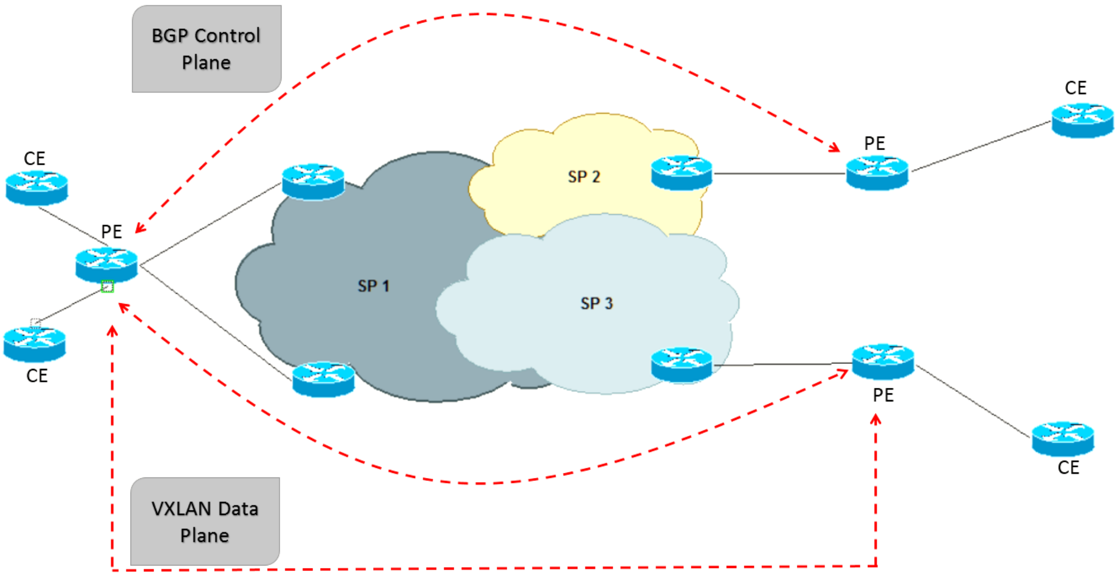
  Figure 20.



  - With EVPN it is possible to provide L2 and L3 services on single interface/VLAN to customers.
  - No need for multiple VPN protocols, it provides both services using only one network topology.
  - Supports active-active or active-standby connections between PE and CE.
  - EVPN service can be run over any core network whether it be MPLS core or IP core.

- **Flexible L2 and L3 Site to Site solution**

  Figure 21.



  - EVPN VXLAN can run over any IP network to provide flexible L2 and L3 VPN services to connect different sites together.
  - No need of MPLS service between the sites, just needs IP connectivity within the sites without any special configuration to be provided by the service provider i.e. EVPN overlay is transparent to service providers and vice-versa.
  - There can be several different service providers between the sites.
  - VPN service at the endpoints can be controlled with BGP and routing policies.
  - Routing and MAC/IP advertisement is managed with IBGP between PE routers.

4 Lab Demo

BGP + T-LDP based VPLS with Multi-homed CE

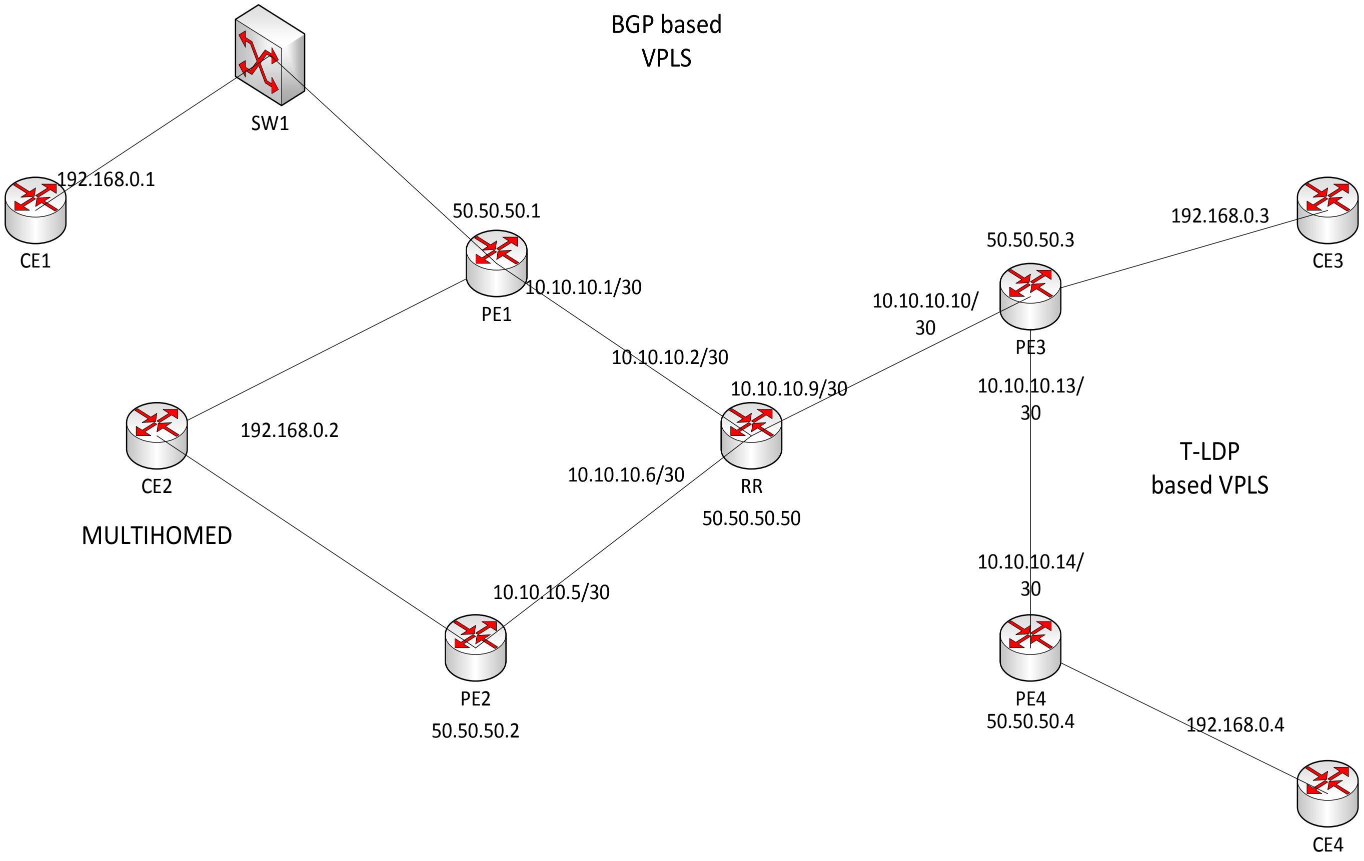
The lab scenario is as follows:

- 3 PE routers; PE1, PE2, PE3 are running BGP based VPLS.
- 1 PE router; PE4 is running T-LDP based VPLS.
- 1 P router; RR is acting as Route Reflector.
- 1 L2 Switch; SW1 Connected to PE1.
- 1 CE router; CE2 Multihomed to PE1, PE2.
- 3 CE routers; CE1, CE3, CE4 are connected to SW1, PE3, PE4 respectively.

Hardware used:

- PE1, PE2, PE3 , RR ; Alcatel Lucent 7750 SR-OS
- PE4 ; Alcatel Lucent 7710 SR-c4
- All CE ; Cisco 2911 router
- SW1 ; Cisco 3560 catalyst Switch

Figure 22: Network Diagram



Configuration steps:

- Configuring IGP: OSPF
  OSPF (IGP) protocol is configured on all the PE routers as well as on RR router for the routers to exchange their routing information. All the routers are kept in the single domain i.e. Area 0.

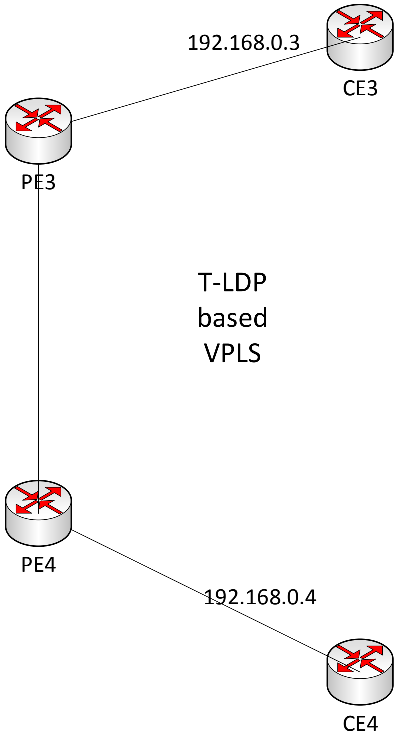
- LDP for transport tunnels
  Label Distribution Protocol is required to exchange the label mapping information. It is configured on all the core routers.

- Creating pseudo wire templates to auto create the SDPs. LDP based pseudo wires can be automatically created using the pseudo wire template.

- Configure BGP with address family L2VPN.
  Configure BGP on PE1, PE2, PE3 and RR router with the address family L2VPN parameter and on RR router add the cluster command to make it the Route Reflector.

- Configure BGP VPLS on PE1, PE2, and PE3. Creating customer and defining the Route Distinguisher and Route Target. Assigning a VE-ID and defining the maximum value for it. The VE-ID and VE-NAME should be unique. STP on PE1, PE2 should be shut down as because of multihoming, it has to be run on the customer edge router to avoid loops.

- PE3 and PE4 are configured to run T-LDP VPLS between them. This way both BGP based VPLS and T-LDP based VPLS are demonstrated.



- Configure Multi Site on PE1, PE2 by giving same site number on both the PE routers. SAP created on the ports that are connected to the CE routers.

- Configure Port channel on CE2 router to make it multihomed to routers PE1, PE2.

5 Conclusion

Current software codes or hardware from major vendors do not support EVPN and its implementations. Due to non-availability of the EVPN supported routers it was not possible to implement any of the EVPN scenario. To conclude this report VPLS with both BGP and T-LDP was implemented in one scenario on Alcatel Lucent 7750 SR-OS demonstrating the limitations of VPLS. One of the biggest drawback of VPLS is that, there is no Active-Active Multi homing support available, one of the PE routers has to be Standby. As we know internet traffic is rising at an enormous rate, to tackle this, load sharing with multihoming is becoming a necessity for not only just big companies. VPLS doesn’t support load sharing. The reason for only one link being active is that the VPLS was developed this way. Other drawbacks include that VPLS can’t support Multipoint to Multipoint LSPs i.e. no optimized delivery of multi destined frames. Network re-convergence time is dependent on the number of MAC addresses learned by the PE router, this is an issue when there are large number of virtualized machines attached at the customer end leading to a large volume of MAC addresses to be taken care of. To demonstrate the above two limitations, Ixia assessment tools are required. The new Data Center Interconnect applications require Layer 2 and Layer 3 services over the same interface added with the scalability and control like L3VPN, this is where VPLS fails.

Future Research:

There are still many inquisitive research queries on how EVPN can will handle the new DCI applications.

According to the authors of RFC7209 EVPN seems to overcome the shortcomings of VPLS, but future research and practical implementation is required to find out whether the theoretical claims of EVPN capabilities are any true and moreover does it introduce any of its own glitches.

One interesting research topic relevant to EVPN that comes to my mind is that how migration from VPLS to EVPN can be done, assuming that VPLS is already implemented and is running. Would it be possible to move to EVPN without turning down the existing VPLS network and can VPLS and EVPN both run together?

6 Bibliography

Alcatel Lucent (February 2014) ETHERNET VPN (EVPN) NEXT-GENERATION VPN FOR ETHERNET SERVICES. Retrieved from

https://conference.apnic.net/data/37/2014-02-24-apricot-evpn-presentation_1393283550.pdf

Alcatel Lucent Scalable IP Networks Self-Study Guide by Kent Hundley (BOOK)
http://www.aldraji.com/download/ScalableIP_NRS1.pdf

Alcatel Lucent 7750 SR OS Routing Protocols Guide (book)
https://infoproducts.alcatel-lucent.com/cgi-bin/dbaccessfilename.cgi/9300740801_V1_7750%20SR%20OS%20ROUTING%20PROTOCO.pdf

Analysis of VPLS Deployment R. Gu, J. Dong, M. Chen, Q. Zeng (Huawei) Z. Liu (China Telecom)

IETF80 L2VPN Mar. 2011 Prague

draft-gu-l2vpn-vpls-analysis-00
https://tools.ietf.org/agenda/80/slides/l2vpn-2.pdf

BGP based VPLS. Retrieved from

http://wiki.mikrotik.com/wiki/Manual:BGP_based_VPLS

BGP MPLS based Ethernet VPN
https://sites.google.com/site/amitsciscozone/home/important-tips/mpls-wiki/bgp-mpls-based-ethernet-vpn

BGP MPLS Based Ethernet VPN draft-ietf-l2vpn-evpn-02
http://tools.ietf.org/pdf/draft-ietf-l2vpn-evpn-02.pdf

BGP MPLS Based Ethernet VPN draft-ietf-l2vpn-evpn-11
http://tools.ietf.org/html/draft-ietf-l2vpn-evpn-11

BGP Multi-Homing for VPLS Networks by Alcatel Lucent
https://infoproducts.alcatel-lucent.com/html/0_add-h-f/93-0267-HTML/7X50_Advanced_Configuration_Guide/BGP_MH.html

Configuring VPLS Routing Instances by juniper
https://www.juniper.net/documentation/en_US/junos14.2/topics/usage-guidelines/vpns-configuring-vpls-routing-instances.html

Configuring a VPLS Service with CLI by Alcatel Lucent
https://infoproducts.alcatel-lucent.com/html/0_add-h-f/93-0076-HTML/7750_SR_OS_Services_Guide/services_vpls.html

EVPN Overview by juniper
http://www.juniper.net/techpubs/en_US/junos13.3/topics/concept/evpns-overview.html

EVPN: Intro to next gen L2VPN
http://packetpushers.net/evpn-introduction-next-generation-l2vpn/

Ethernet VPN Layer 2 Scalability by Shivlu Jain
http://www.slideshare.net/ShivluJain/ethernet-vpn-layer-2-scalability

Ethernet VPN (EVPN) for integrated layer 2-3 services – Alcatel Lucent
http://www2.alcatel-lucent.com/techzine/ethernet-vpn-evpn-integrated-layer-2-3-services/

Ethernet VPN (EVPN) and Provider Backbone Bridging-EVPN: Next Generation Solutions for MPLS-based Ethernet Services by cisco(white paper)
http://www.cisco.com/c/en/us/products/collateral/routers/asr-9000-series-aggregation-services-routers/whitepaper_c11-731864.html

ETHERNET VPN (EVPN) OVERLAY NETWORKS FOR ETHERNET SERVICES by Greg Hankins (Alcatel Lucent)
https://ripe68.ripe.net/presentations/170-ripe-68-evpn.pdf

Hierarchical VPLS. Retrieved from

https://sites.google.com/site/amitsciscozone/home/vpls/hierarchical-vpls

https://routingfreak.wordpress.com/2011/02/21/does-hierarchical-vpls-solve-all-scaling-issues-found-in-vpls/

Implementing IEEE 802.1ah Provider Backbone Bridge by cisco
http://www.cisco.com/c/en/us/td/docs/routers/asr9000/software/asr9k_r4-3/lxvpn/configuration/guide/lesc43xbook/lesc43pbb.html#90421

MPLS and VPLS security by Enno Rey.
https://www.blackhat.com/presentations/bh-europe-06/bh-eu-06-Rey-up.pdf

MPLS-Enabled Applications: Emerging Developments and New Technologies By Ina Minei, Julian Lucek
https://books.google.ca/books?id=2lxbaQ-VN8sC&pg=PA396&lpg=PA396&dq=multihoming+drawback+MPLS&source=bl&ots=L5BUW3DvIU&sig=x5qIErvuqKeCXlnkgM84X2zssQU&hl=en&sa=X&ei=3E7pVLiwFoyWgwSks4OIAQ&ved=0CCwQ6AEwAjgU#v=onepage&q=multihoming%20drawback%20MPLS&f=false

Network Configuration Example Validating a BGP-Based VPLS Multihoming Configuration
http://www.juniper.net/techpubs/en_US/release-independent/nce/information-products/topic-collections/nce/bgp-vpls-multihoming/validating-a-bgp-based-vpls-multihoming-configuration.pdf

(PBB-)EVPN Seamless Integration with (PBB-)VPLS draft-sajassi-bess-evpn-vpls-seamless-integ-00
https://tools.ietf.org/html/draft-sajassi-bess-evpn-vpls-seamless-integ-00

Requirements for Ethernet VPN (EVPN) (RFC 7209)
https://tools.ietf.org/html/rfc7209

Scaling Virtual Private LAN Services. Retreived from

http://blog.ine.com/2010/11/26/scaling-virtual-private-lan-services-vpls/

Scale and Extend VPLS with LDP-BGP VPLS Interworking Retrieved from
http://www.eetimes.com/document.asp?doc_id=1208544&

Scaling BGP Cisco live
http://d2zmdbbm9feqrf.cloudfront.net/2013/usa/pdf/BRKRST-3321.pdf

Technical Brief: Offering Scalable Layer 2 Services with VPLS and VLL
http://www.brocade.com/downloads/documents/technical_briefs/Offering_Scalable_Layer2_Services_with_VPLS_and_VLL.pdf

Virtual Private LAN Service (VPLS) Interoperability with Customer Edge (CE) Bridges (RFC 6246)

Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling(RFC 4761)

Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling(RFC 4762)

Virtual Private LAN Service (Wikipedia)
http://en.wikipedia.org/wiki/Virtual_Private_LAN_Service

Virtual Private LAN Services (VPLS) by cisco
http://www.cisco.com/c/en/us/td/docs/switches/lan/catalyst6500/ios/12-2SY/configuration/guide/sy_swcg/vpls.pdf

VPLS BGP Signaling by cisco
http://www.cisco.com/c/en/us/td/docs/ios-xml/ios/iproute_bgp/configuration/xe-3s/irg-xe-3s-book/irg-vpls-bgp-sig.html

WHITE PAPER - LDP-BGP VPLS Interworking by juniper.
http://www.juniper.net/us/en/local/pdf/whitepapers/2000282-en.pdf