Virtual Private LAN Service: The new challenge for LAN/WAN connectivity

Marcelo Yannuzzi, Eva Marín Tordera, Xavier Masip-Bruin,
Jordi Domingo-Pascual, Sergi Sánchez-López, Josep Solé-Pareta
Departament d'Arquitectura de Computadors, Universitat Politècnica de Catalunya
Avgda. Víctor Balaguer, s/n, 08800 Vilanova i la Geltrú, Barcelona, Catalunya, Spain
{yannuzzi, eva, xmasip, jordid, sergio, pareta} @ac.upc.es

***Abstract* - This paper presents a technical overview of VPLS (Virtual Private LAN Service), its current state of the art, its strengths and weaknesses, and most important of all, several open issues and future research work.**

I. INTRODUCTION

In the past several years significant technological improvements have taken place in the field of local area networks, including bandwidth jumps from a 10Mbps shared segment up to switched Gbps segments, as well as numerous enhancements in its availability and flexibility just to name a few.

Throughout these years Ethernet has become the most widely deployed and ubiquitous local area network technology, not only due to those incremental speed advances, but also because of its cost advantages and simplicity. From an end user standpoint, Ethernet needs nearly no administration, whereas it also provides high availability and great bandwidth. The sum of all this features turned it into in the most cost-effective solution for local area networks. Nevertheless, some limitations that are inherent to Ethernet switching protocols preclude its use in building L2 VPN (Layer 2 Virtual Private Network) services that scale within, or even beyond, a MAN (Metropolitan Area Network) domain.

These limitations drove researchers and vendors to extend Ethernet's physical reach in order to provide customers with Ethernet services either within the same metro area, or even spanning across several geographically dispersed metropolitan areas. In other words, the main goal lying beneath these techniques was to provide customers with transparent Ethernet services throughout a WAN (Wide Area Network) cloud. From the service provider standpoint, to offer transparent Ethernet services to end customers meant to provide any-to-any (multipoint-to-multipoint), full mesh services. Thus, network architectures such as ATM LAN Emulation emerged as the ATM Forum proposal to provide customers with transparent LAN services. Emulated local area networks became Virtual Private Networks (VPN) in the LANE framework, and soon providers started transparently multiplexing Ethernet frames over their ATM backbones. However, several drawbacks arose from the LANE architecture, mainly due to its complexity, overhead and costs.

From this perspective, new proposals appeared with the aim of providing simpler L2 VPN solutions. Among these, VPLS (Virtual Private LAN Service) has arisen as a strong candidate to meet these needs. VPLS, which is often referred to as TLS (Transparent LAN Service) or VPSN (Virtual Private Switched Network), delivers highly scalable multipoint-to-multipoint Ethernet services that can span several metro areas providing connectivity to multiple sites just as if they were attached to the same Ethernet segment. VPLS is a proposed IETF standard, which provides a MPLS (MultiProtocol Label Switching) based L2 VPN solution, since it uses the IP/MPLS service provider infrastructure.

This paper presents a technical overview of VPLS, which surveys its current state of the art, its strengths and weaknesses, and highlights several open issues and future research work. The remaining of this paper is organized as follows. Section II presents the main concepts of VPLS, and surveys the VPLS control and forwarding planes. Section III discusses different architectures and mechanisms with the aim of providing highly scalable VPLS services, including proposals for an Inter-Domain VPLS service. In addition, this section surveys promising deployment scenarios and briefly describes a few techniques that have been proposed in order to provide VPLS with QoS (Quality of Service). In Section IV several open issues are presented. This section shows that even though some of these issues are part of the ongoing research efforts, others will certainly need to be addressed as future research work. Finally, Section V concludes the paper.

II. OVERVIEW OF VPLS

The main goal of VPLS is to provide L2 connectivity among geographically dispersed customer sites through a WAN or MAN cloud just as if they were attached using

legacy bridged Ethernet ports. In other words, VPLS is capable of constructing numerous private Ethernet services over a service provider's shared network infrastructure which may span several metro areas, or even span across AS (Autonomous System) boundaries. In contrast to many commonly used L2 VPN solutions, which are inherently point-to-point, VPLS offers a complete multipoint-to-multipoint L2 VPN service. Figure 1 depicts the topology paradigm of VPLS. In the figure two different L2 VPNs are shown, VPN1 and VPN2. Even though VPN1 and VPN2 remote sites are attached through a complex IP/MPLS service provider backbone, they seem to each other as directly connected through an Ethernet bus. VPLS presents an Ethernet interface to end customers, which not only simplifies the LAN/WAN connectivity between service providers and customers, but also enables quick and flexible service provisioning, and support for SLA (Service Level Agreement) on a per VPN basis. From the Customer Premises Equipment (CE) point of view, the WAN/MAN infrastructure is not visible. CE appears to each other as connected via fully meshed bridged ports and hence is completely unaware about the VPLS service or the IP/MPLS core network. Moreover, the P routers composing the IP/MPLS service provider core network are also totally unaware of the existence of VPLS, since VPLS pushes complexity to the edge of the provider's network. Each Provider's Edge router (PE router) at the edge of the IP/MPLS service provider's network is enhanced with special VPLS features. A VPLS instance runs on each PE router for each VPLS domain connected to it. If a PE router lacks of VPN customers, then no VPLS instances run on it. Those VPLS instances allow PE routers to learn MAC addresses, bridge Ethernet frames, age MAC addresses, and flood broadcast, multicast and unknown unicast frames on a per-VPLS basis, just as a legacy Ethernet switch would do. From this perspective, each PE router acts as several bridges with several ports; one bridge for each VPN attached to it, and as many ports as VPN sites attached to it that belong to the same VPN domain.

An underlying assumption regarding PE routers within a VPLS domain is that they are assumed to be logically full meshed with MPLS LSP (Label Switched Path) tunnels, enabling any-to-any connectivity. Then, encapsulating and forwarding frames belonging to a VPLS service over this logical full mesh becomes possible. Furthermore, several VPLS services can be carried within each of these LSP tunnels. This means that these tunnels are not only capable of carrying traffic of different VPLS instances belonging to different customers, but also are capable of carrying different VPLS instances belonging to different VLANs (Virtual LANs) but from the same customer.

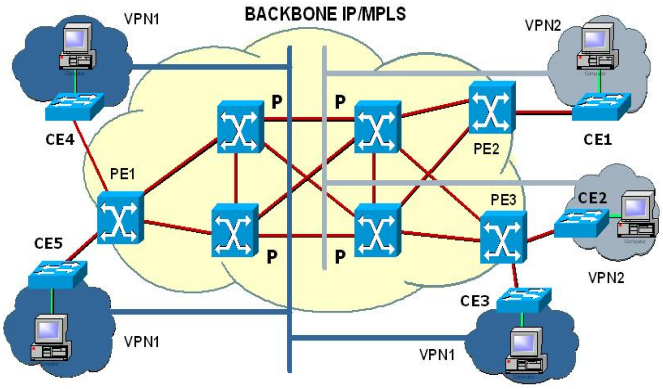

Fig. 1. Topology of VPLS

These tunnels are established by means of a signalling protocol such as RSVP-TE (Resource reservation Protocol-Traffic Engineering) or LDP (Label Distribution Protocol), and they are completely independent of the services offered over them. In particular, VPLS is just one of the many services that they are able to offer. Therefore, the details such as the signalling and the establishment of these tunnels is out of the scope of VPLS and hence of this work.

In [1] the authors define encapsulation methods for transporting Ethernet frames over point-to-point MPLS LSPs, called Pseudo-Wires (PWs). Whereas [1] only deals with Ethernet frames, [2] describes how to transport L2 frames over an IP/MPLS network. VPLS among other things provides extensions to [2] in order to transport 802.3 and VLAN [802.1Q] traffic in a multipoint-to-multipoint fashion from customer sites belonging to the same L2 broadcast domain. As depicted in Figure 2, in VPLS a full mesh of PWs is established among PEs running peering VPLS instances, also known as VSIs (Virtual Switch Instances).

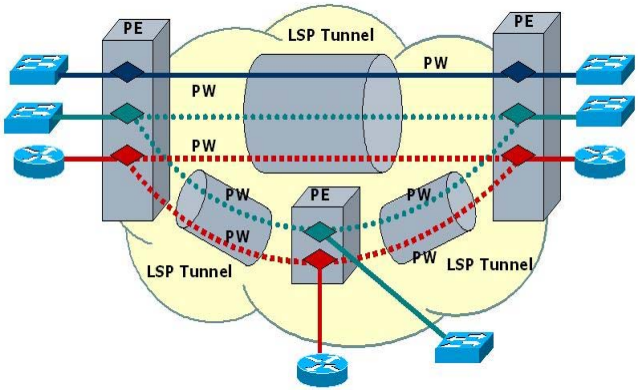

Fig. 2. VPLS VSIs, LSPs and PWs

Those instances are represented as diamonds on each PE. The PWs are used for demultiplexing the L2 encapsulated frames that traverse the LSP tunnels. It is worth mentioning once again that the LSP tunnels among PEs are assumed to exist, and are not a VPLS concern. VPLS provides an overlay structure of PWs over those pre-established LSP tunnels. How these PWs are setup and tear down, and how a set of PE devices interconnected via PWs appears for an end customer as an Ethernet bus, those are definitively VPLS concerns. From this standpoint, each PE acts as a legacy bridge on its customer facing ports, and as a MPLS switch on its core network facing ports. In that sense, PE routers need to learn remote MAC addresses both, from its customer facing ports, and from the PWs, and then to establish associations between PWs and customer ports on a per-VPLS instance basis.

VPLS relies on the one hand on a control plane for the tasks of auto discovery of new VPLS members, and the setup and tear down of PWs on any given VPLS domain. On the other hand, a VPLS data plane defines how VPLS encapsulates and forwards data. Nowadays several proposals concerning the details and functionality of the control and data planes are available. However, it is important to highlight that although a few manufacturers have started presenting some VPLS capable devices, and even though we can find plenty of ongoing research in the area of L2 VPNs, there is no standard for VPLS yet. Two leading IETF (Internet Engineering Task Force) working groups provide different proposals for a VPLS standard, which mainly vary their levels of automation and operational efficiency. The automation refers to the VPN auto discovery process, while the operational efficiency refers to the VPN signalling process or how PWs among VPLS instances of a VPLS domain are setup and teardown. Consequently, it is mainly the VPLS control plane the subject under discussion at the IETF. On the one hand, the IETF draft led by Kireeti Kompella [3] proposes MP-BGP (Multi-Protocol Border Gateway Protocol) as the protocol for both, discovery and signalling. On the other hand, the IETF draft led by Marc Lasserre and Ali Sajassi [4] is agnostic to discovery protocols, and proposes LDP as the signalling protocol. It is worth mentioning that other proposals also exist, such as the one from Juha Heinanen, et al [5], which describes a simple mechanism to implement Provider Provisioned Virtual Private LAN Service (PPVPLS) using Radius for PE discovery and L2TP (Layer 2 Tunneling Protocol) as the control and data plane protocol.

Needless to say, right now it is really uncertain when the current efforts being done on VPLS research will overcome the remaining challenges to generate a standard. The details about how a VPLS service is setup and how the control and data planes operate are the subject of the next subsections.

A. Control Plane

As it was mentioned before, the primary functions of the VPLS control plane are the auto discovery and the setup and tear down of PWs that constitute the VPLS service. The latter set of functionalities is also known as signalling.

1) Auto Discovery

The auto discovery refers to the process of finding all the PEs participating in a given VPLS domain. Since all PEs participating in that domain need of fully meshed PWs, a non-auto discovery scheme would require intense configuration methods from the VPLS service provider point of view. Furthermore, topology changes such as adding or removing a new PE to the network or even adding or removing a new site for an existing VPN would involve several configuration tasks. Through the auto discovery process each PE is able to discover the other PE routers that are part of a VPLS domain. This must be done by some protocol, so several IETF proposals such as MP-BGP, Radius, or DNS (Domain Name Service) extensions exist in order to develop this protocol. In this paper we will briefly analyze the MP-BGP approach mainly for four reasons. Firstly, if BGP is utilized for auto discovery as well as for signalling, then the service provider infrastructure is completely reutilized. At the present, many service providers are offering IP VPN services, often called RFC2547 VPNs, which are indeed based in MP-BGP. In this sense the service provider's network is shared among L2 and L3 VPNs, since the PE routers needed to provide one service or the other are essentially the same. Secondly, BGP can easily reduce the $O(N^2)$ PW mesh to O(N) by means of a BGP Route Reflector or a cluster of Route Reflectors if redundancy and fault tolerance is needed. Thirdly, provisioning Inter-Domain VPLS services becomes clearer using BGP. Finally, it is likely to believe that a combination of MP-BGP discovery and LDP signalling may result in an IETF standard.

The MP-BGP approach uses BGP extended communities to identify members of a given VPLS. The extended community used is the Route Target, whose format is described in [6]. In this scenario a PE announces via IBGP (Internal BGP) to its peering PEs that it belongs or no longer belongs to VPLS X. The next figure shows a typical BGP/VPLS scenario and introduces a L2PE (Layer 2 Aggregation PE), which corresponds to a new device owned by the service provider that decouples several functionalities from the PE routers.

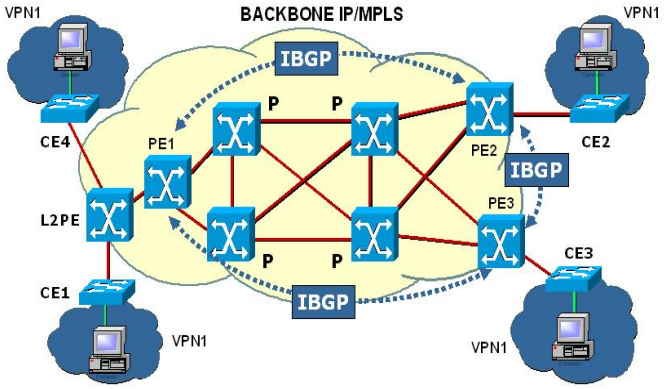

Fig. 3. Full IBGP mesh for MP-BGP Auto Discovery
L2PE decouples some functionality from the PE routers.

L2PE devices allow a more flexible VPLS deployment, and also allow scaling the flat network architecture presented in Figure 1. The details concerning scaling VPLS will be presented in Section III. In this framework, both the L2PE and the PE devices are VPLS aware, however a L2PE device does not need the same level of detail in the VPLS control plane than a PE does. In fact, while the PE deals with the processes of auto discovering other PEs running peering VPLS instances and signalling PWs to those PEs, the L2PE deals among other things with the MAC address learning process for the VPLS domains it knows. In this way the L2PE unloads the PE from one of the biggest burdens in a VPLS framework, which is to handle one MAC address FIB (Forwarding Information Base) for each VPLS instance running on it. Nevertheless, a mixed scenario y also achievable where PE1 in the above figure could handle several FIB from potential VPLS customers directly connected to it, as well as relies on L2PE for handling the aggregated MAC FIB from the set of VPLS customers attached to the L2PE.

2) *Signalling*

Once the auto discovery process is done, the next step is that each PE in a VPLS domain must be able to setup and tear down PWs to each other. The most compelling proposals are MP-BGP and LDP. In the case of MP-BGP signalling, an encapsulation Type and Control Flags are encoded in an extended BGP community attribute [3]. In the case of LDP, which seems the path followed by many manufacturers, once a full mesh of LDP sessions is established, a full mesh of PWs is set up. After a LDP session has been formed between PEs, all PWs among them are signalled over this session. Therefore, the specific L2 VPN information is carried in the LDP messages sent among PEs, and this is done by means of label mappings [4]. The main disadvantages of the LDP proposal are, on the one hand that the requirement of a full mesh of PWs implies $O(N^2)$ of PWs, which can only be reduced rearranging the network topology. The initiative to reduce the number of PWs is again to decouple PE functionalities into a pair of switches, a n-PE (network Provider Edge) switch and an u-PE (User Facing Provider Edge) switch. The role of this new u-PE is essentially the same of the L2PE switch in the BGP framework. The u-PE actually aggregates several VPLS domains and communicates with its corresponding n-PE. The n-PE is once more in charge of auto discovery (not addressed in the LDP approach) and signalling, however the full mesh of LDP sessions and hence the full mesh of PWs is carried only among n-PEs. This approach substantially reduces the number of LDP tunnels and PWs. The MP-BGP provides a better approach to the issue since no topological changes are needed as far a RR (Route Reflector) is present in the service provider network. Nonetheless, this argument is weak since a full mesh of LSP tunnels is still needed in order to provide a non-meshed PW structure based on IBGP and the presence of a RR. The LDP approach establishes those fully meshed LSP tunnels using LDP and relies on a non-flat VPLS topology in order to reduce the number of PWs needed. However, a flat VPLS topology will suffer from several scaling problems, and hence it seems wise to deploy hierarchical VPLS scenarios where the LDP $O(N^2)$ of PWs problem seems to fade.

The other significant disadvantage of the LDP proposal is that a service provider deploying simultaneously L2 and L3 VPN services with TE (Traffic Engineering) needs to manage two different services in nature, whereas the MP-BGP approach offers a common framework for both of them.

Furthermore, the LDP approach seems to diverge from the next generation network trends; however this issue will not be treated here since we will return to it in section IV.

3) *Loop Avoidance*

Unlike Frame Relay or ATM where the termination point becomes the CE node, Ethernet switches have to inspect the L2 fields of the frames to make a switching decision. In the case that the frame is targeted to an unknown destination, or is a broadcast or multicast frame, the frame must be flooded. Consequently, if the PE routers logical topology is not a full mesh, the PE devices may need to forward these frames to other PEs. This would require the use of STP (Spanning Tree Protocol) to avoid loops through the core network topology. However, the use of a STP instance per VPLS domain within the service provider network may have characteristics that are undesirable to the provider. Instead, the joint use of a full

mesh and split-horizon obviates the need for STP. Under this proposal, each PE needs to support a split-horizon scheme in order to prevent loops, and the expected PE behaviour is that it must never forward traffic from one PW to another in the same VPLS. This is due to the fact that each PE has a direct logical connection to all other PEs within the same VPLS. In the case that the customer's network topology presents itself loops, the customers are allowed to run their own STP instances. In such cases the STP from the customers is transparently tunnelled through the service provider's WAN/MAN cloud.

B. Forwarding Plane

Even though services like broadcast and multicast are available in traditional LANs, MPLS does not support them yet. Different customer sites belonging to the same broadcast domain which are connected via an MPLS network expect broadcast, multicast and unicast traffic to be forwarded to the appropriate destinations. This requires MAC address learning and aging on a per-VPLS instance basis, and packet replication across the MPLS LSPs tunnels for broadcast, multicast and unknown unicast destination traffic. Moreover, MAC addresses removal and fast relearning techniques are needed. This is due to the fact that in case a network topology change occurs, the VPLS instances need to withdraw and relearn the affected MAC addresses as soon as possible in order to avoid disrupting the service. In many cases customers may desire to have their CE dually connected to different PEs for fault tolerant reasons. In these cases when the active PE goes down, the standby PE needs to react fast and instruct the other peering PEs to withdraw their entries for a given set of MAC addresses and relearn that those are now accessible through the standby PE.

1) MAC Address Learning

Once the PWs are established, PEs could start learning MAC addresses and sending data frames. Learning is essentially the process of mapping source MAC addresses from received frames with the ports on which they arrive. This mapping information is stored in the FIB, which is then used for forwarding frames. VPLS offers two different learning strategies, called qualified and unqualified learning modes. In qualified learning, the learning decisions are based on the customer MAC address and the VLAN tag, if one is used. In case no VLAN tag exists, the default VLAN is assumed. In this way, for each customer VPLS domain, the VPLS aware devices need to maintain multiple logical FIBs, one for each VLAN tag identified in a customer frame. Conversely, in unqualified learning mode, learning is only based on the customer's MAC addresses, thus only one FIB exists per VPLS domain.

In the case of VPLS, a demultiplexor is used not only to identify the VPLS domain to which a data frame belongs to, but also to identify the ingress PE. While the former information is used for forwarding the frame, the latter is used for learning MAC addresses. In our case the core network is an IP/MPLS network, thus the demultiplexor is a MPLS label. In other words, VPLS uses demultiplexors to discriminate among several different streams of traffic carried over a LSP tunnel. The role of demultiplexors should become clearer from the next example. Let's assume that in Figure 3 the VPLS domain of VPN1 is set up. Therefore, a full mesh of PWs exists among PE1, PE2 and PE3. These PWs could have been created by means of MP-BGP or LDP signalling. Let's also assume that for this VPLS domain PE2 signals the label numbers 21 and 23 for PE1 and PE3 respectively. Likewise, PE3 signals label numbers 31 and 32 for PE1 and PE2 respectively. If a frame from CE2 is bound for CE3 with source MAC address M2 and destination MAC address M3 the following actions take place. Firstly, if PE2 does not know where M3 is, in other words M3 is not in the MAC FIB for VPN1, the frame needs to be broadcasted towards PE1 and PE3 as well as any other bridged port if any, but in this example this is not the case. When PE3 receives the frame, the inner label will be 32, and in this way PE3 is able to conclude that M2 is behind PE2, since it distributed the label number 32 to PE2. At this moment, PE3 has learned M2 MAC address since it is able to associate M2 with label number 23. Please notice that an inner label was mentioned in the previous example and this is because the P switches are completely unaware of the VPLS service. As it was formerly mentioned, all VPLS communications among a pair of peering PE routers is done through a LSP tunnel between them. Therefore, a stack of MPLS labels is used for encapsulating the Ethernet frames towards the tunnel. The outer label is used for identifying the next-hop P switch while traversing the tunnel, whereas the inner label is targeted for the receiving PE.

2) MAC Address Replication

One of the inherent features of an Ethernet service is that all broadcast and unknown MAC addresses traffic is flooded to all participating ports on a given VLAN domain. To provide this flooding, the service provider needs to flood all address unknown unicast and broadcast traffic through the corresponding PWs to all relevant PE routers, as well as flood the frames through its legacy bridged ports participating in the VPLS domain.

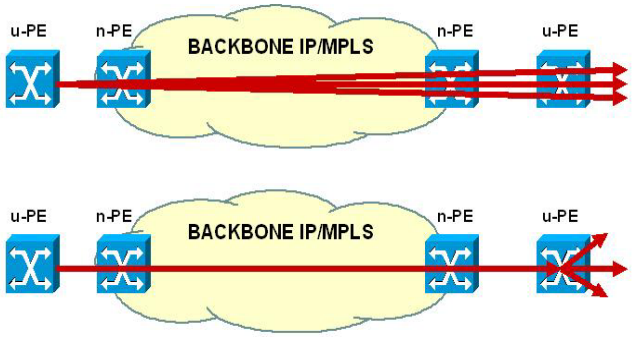

Fig. 4. Replication process: the scheme contrasts how replicating next to the source for three different sites belonging to the same VPLS domain, and attached to the same u-PE wastes bandwidth within the MPLS backbone.

Conversely, multicast frames do not necessarily have to be sent to all VPN members. VPLS allows either to broadcast multicast frames or to rely on IGMP (Internet Group Management Protocol) and PIM (Protocol Independent Multicast) snooping techniques to improve multicast traffic flow efficiency.

It is widely known that this flooding process is one of the most important limitations in order to scale Ethernet LANs. Similarly, this flooding process limits how far a VPLS service is able to scale. Figure 4 contrasts two different approaches to the flooding issue. On the upper scheme, the u-PE device replicates and floods the broadcast frame towards all the relevant PWs. Particularly, if a peering u-PE attaches three different sites from that specific VPLS domain, the source u-PE floods three different frames, one for each remote destination. This process unnecessary wastes bandwidth from the MPLS core network, thus an alternative approach is to replicate as close to the target as possible. Further than this trivial example, it is quite obvious that in a service provider VPLS scale, this flooding issue needs to be addressed so as to limit the number of replicas traversing the MPLS backbone. Surprisingly, this replication issue is barely addressed in both [3] and [4], so it remains as an open concern.

3) *MAC Address Aging*

In the same way as in a legacy Ethernet service, a PE device needs to be endowed with an aging mechanism so that unused MAC addresses could be removed from its FIB. This process not only saves memory, but also allows a consistent state of the PE devices. As soon as a frame with source MAC address X arrives to a PE device, the MAC X is remembered and mapped to its incoming port in the FIB, and an aging timer associated with this new entry starts its countdown. This aging timer should be refreshed each time a new frame arrives with the same source MAC X. In case the timer expires, the entry for MAC X should be removed from the FIB.

4) *MAC Address Withdrawal*

Customers often choose multi-homed access topologies in order to endow with high availability and resilience to their primary Ethernet access links. In case the primary access link fails, it is desirable that the PEs providing the access service could dynamically trigger the processes of removal and relearn of the affected MAC addresses in order to achieve faster convergence. In [4] an interesting approach is taken, since it proposes to use an Address Withdraw message, which is utilized to specify a list of MAC addresses that must be removed or relearned. When such a message arrives to a PE router with a specific list of MAC addresses to be relearned, the PE updates its FIB for that VPLS instance, and forwards the message to the relevant PEs over their corresponding LDP sessions. Please notice that this approach applies indeed in presence of a non-flat VPLS topology. In this scenario, the source of the message is likely to be a L2PE or a u-PE device, so when the corresponding PE receives the Withdraw message, it first processes and then forwards the message to other PEs. If the VPLS topology is flat, then there is no need that PE devices forward to each other the Withdraw messages. The message also offers the possibility of sending an empty list, which instructs the receiving PE to remove all the MAC entries learned for that specific VPLS instance, except of course those learned from the sending PE.

5) *Encapsulation*

Among the foundations of VPLS services are, the setup of overlaying point-to-point PWs between peering PE devices over a fully meshed infrastructure of LSP tunnels, and also the encapsulation methods to transport Ethernet frames over those PWs. These ideas are basically borrowed from [1]. What changes in VPLS from [1] is that instead of providing point-to-point Ethernet services over a shared IP/MPLS backbone, it provides a full multipoint-to-multipoint Ethernet service by means of fully meshed PWs. Therefore, the encapsulation method find in both, the LDP and the MP-BGP VPLS proposals, is in essence the Martini encapsulation technique [1].

In both proposals, when the customer's Ethernet frames need to traverse the service provider's backbone, the Ethernet frames that a PE device receives from any directly attached CE, are encapsulated for transmission over the IP/MPLS network connecting the PEs.

In the case of the MP-BGP VPLS proposal, the encapsulation is as in [1] with some minor changes, which among other things, allows the peering PEs to strip the outermost VLAN tag of an Ethernet frame received from a CE before encapsulating it, and likewise to push a VLAN tag onto a de-capsulated Ethernet frame before sending it to the corresponding CE. These capabilities are provided by three new control flags defined for VPLS, which are encoded in the BGP extended community attribute.

In the LDP VPLS approach, untagged customer Ethernet frames arriving from a directly connected CE are encapsulated as defined in [1], whereas tagged customer Ethernet frames are encapsulated as follows. A tagged Ethernet frame, whose tag is locally used by the ingress PE device as a service delimiter in order to distinguish the different customers, or to distinguish a specific service of a given customer, are stripped before sending them into the IP/MPLS network. Similarly, at the other edge of the IP/MPLS backbone, the egress PE device may insert its local service delimiting tag in case it is needed. Please notice that a hybrid scenario is not only possible, but also desirable. While a PE device may differentiate its attached services on a per-port basis, and hence no tagging is needed, a peering PE may aggregate several customer services on a single port, and then a way to distinguish different services on that port is required. Encapsulations such as 802.1Q are frequently used in such ports, where a VLAN tag allows the PE device to discriminate frames from the different VPLS services. Furthermore, different customer services could be connected to its corresponding PE router through different ATM VCs (Virtual Circuits) where the customer Ethernet frames are transported over an ATM access network. In this case, the ATM VC number is the tag that identifies a particular customer's VPLS instance, and hence this ATM encapsulation must be removed by the ingress PE device before switching the frame towards the VPLS network. Alternatively, when a tagged Ethernet frame whose tag is not a service delimiter arrives to a PE, it should be encapsulated and forwarded towards the IP/MPLS with no modification at all. In other words, this tag is owned by the customer and it is likely to be used to distinguish among the several VLANs within its L2 network. Therefore, VPLS should transparently transport those customers tagged Ethernet frames.

Subsequently, this set of rules establishes that once inside the VPLS infrastructure, the payload of the PDU (Protocol Data Unit) traversing the network is always a customer Ethernet frame. This means that the tagging and stripping actions are locally managed by each PE, and then no tagged frame ever traverses the VPLS network. Moreover, tags may overlap since they are never signalled across the VPLS domain to other PEs.

III. SCALING VPLS

The flat VPLS architecture described in Figure 1 presents several problems in terms of scalability since each VPLS service requires a full mesh of PWs among the participating PE devices. In this framework, the only VPLS aware devices are the PE routers. As the scale grows, not only more PE routers will be needed, which will in turn increase the stress on the mesh of PWs, but also each PE will aggregate a large number of customers. In this scenario, each PE will need to accomplish several demanding tasks simultaneously, like, discovering new VPLS members, signalling of PWs, managing one FIB per-VPLS service, encapsulating/de-encapsulating customer data frames, replicating broadcast and unknown unicast addresses and managing multicast in order to avoid broadcasting multicast frames. From these, frame replication and managing an unbounded number of customer MAC addresses per-FIB are the most demanding tasks. Unlike IP addresses, MAC addresses cannot be aggregated into a summary address block, which means that each FIB in the VPLS network could grow to a large number of individual MAC addresses. Furthermore, each client may demand to manage several FIBs, one per-VPLS instance.

In summary, a hierarchical VPLS architecture is proposed, which reduces the stress on PE devices and at the same time allows large scale deployment.

1) Hierarchical VPLS

H-VPLS (Hierarchical VPLS) is often called distributed VPLS. The key to develop distributed VPLS services is to decouple PE from some of the most demanding tasks. In Section II we introduced such decoupling technique by means of the L2PE and u-PE devices. For the rest of the paper we will call u-PE to this decoupling device. Particularly, the purpose is to decouple MAC address learning, STP and the processes of replication and flooding, from the control plane tasks of discovery and signalling. Please recall that in a flat VPLS topology there was no need to run STP, as long as the PWs were fully meshed and split-horizon was active on PE routers. Moreover, customer STP instances were transparently transported over the VPLS network. However, STP is certainly needed to select active ports when, for network resilience reasons, a u-PE device is connected to multiple n-PE devices, which is often called multi-homing.

An interesting approach to distribute VPLS is to manage QinQ logical interfaces (802.1Q encapsulation) between the u-PE and its corresponding n-PE devices. The Q-in-Q encapsulation is basically another L2 tunnelling scheme, and

as it was mentioned earlier, it can be used by the n-PE as a tagging method in order to distinguish among different VPLS services within the same incoming physical port. This is only possible if the u-PE is a bridging capable device, since it needs to be able to discriminate different virtual services within the same physical port. In case that the u-PE is a non-bridging capable device, the only way to decouple PE, and hence to delegate different VPLS services to the u-PE is by means of managing several physical ports, one port per-VPLS service.

The next figure depicts a hierarchical VPLS framework. In this scenario the mesh of PWs is now bounded only to a full mesh of PWs among n-PEs. This topology allows for hybrid deployment of services were a n-PE may have several u-PE devices attached to it, but it may have also some directly connected CEs. A hybrid scheme may become interesting in intermediate scaling cases, were some sort of hierarchy is needed in order to release some stress from the PE devices, but a pure hierarchical model may result unjustifiable or too expensive.

In order to discriminate VPLS services sharing the same access port, the u-PE devices may use 802.1Q tags, or any other tagging method, with its attached CE. These tags allow a u-PE to distinguish different VPLS services belonging to a single customer. Then, each of those tags needs to be mapped to a QinQ tag, so that the corresponding n-PE is able to discriminate them. In turn, a mapping between QinQ tags and a MPLS label stack is accomplished by the n-PE device. Among the information exchanged between the u-PE and its corresponding n-PEs are, the VPLS domain ids, the customer sites ids, as well as their corresponding tags on the u-PE to n-PE link.

Each n-PE advertises to its peering n-PEs a set of labels, where each label represents a VPLS domain id, a unique u-PE attached to it, a, and a unique customer site id within this u-PE. Once more, a MPLS label stack is needed since a full mesh of LSP tunnels between all n-PEs is assumed to exist.

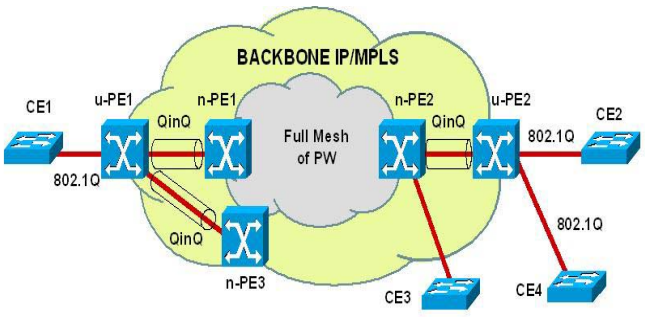

Fig. 5. Hierarchical VPLS topology

The outer label is used while traversing the tunnel, whereas the inner one is targeted to the egress n-PE device, which allows it to precisely identify the outgoing u-PE and customer VPLS service pairs.

In this model the u-PE devices achieve the task of MAC address learning both, from directly attached customer ports, and from the u-PE to n-PE link. Moreover, u-PE devices handle flooding in order to release n-PE devices from the frame replication process. We will not discuss here the details of replication since it is still an open issue. However, it is worth mentioning that when the replication takes place at the u-PE, then this device generates several replicas of the same frame but with different QinQ tags targeted to its n-PE.

2) Multi-Homing

H-VPLS introduces an architecture that avoids several limitations that a flat VPLS topology presents. In most cases, the easiest way to deploy H-VPLS services is by means of spoke connectivity between u-PE devices and their correspondent n-PE device. Nevertheless, this proposal presents a single point of failure, and hence many customers would suffer from total loss of connectivity with their peering sites at the other edge of the VPLS cloud in case the link u-PE to n-PE fails. Therefore, an enhanced solution is to attach each u-PE with two different n-PE devices. In this case, the u-PE could negotiate PWs with both of them, but only use the alternative PW in case the primary fails. This is an attractive proposal due to the fact that there is no need to run a STP since only one link is active at any given time. When a failure is detected, the u-PE switches from the primary PW to its backup and the corresponding n-PE starts learning MAC addresses. Regardless of how fast the u-PE is able to switch upon a failure, the rest of the n-PEs still believe that the set of MAC addresses it manages are accessible through the former n-PE. This situation will continue until the rest of the n-PE devices learn the new location of the desired MAC addresses. In order to provide faster convergence, the MAC address withdrawal technique must be used. When the rest of the n-PE devices receive a withdrawal message, they remove all MAC address entries related to the VPLS instance announced in the message. In Figure 5 u-PE1 is dually homed to n-PE1 and n-PE3.

Typically, hello messages between u-PE and n-PE are utilized to detect link or device failures. The specific details of these hello messages depend on the implementation. For instance, in the LDP scheme, the LDP hello messages are used for this propose.

An alternative approach to the hub and spoke connections is to deploy VLAN islands connecting a u-PE with multiple

n-PEs. In this case, STP should be used in order to prevent that loops are created.

Multi-homing not only applies to u-PE devices, since customers may also want a resilient connection scheme. Therefore, CEs may be dually attached to the same u-PE, to different u-PE or even mixed situations like attaching to n-PE devices. In any case the concerns are to avoid looping conditions, and that any given entry within a VPLS VSI/FIB should never flap. Whether to use a primary/secondary link or the STP approach depends on the topology and deploying scenario.

3) Inter-Domain VPLS

H-VPLS not only means to deploy scalable VPLS services within a single administrative domain, it also means that a VPLS service may span across different domains, which is often called Inter-Domain VPLS or Multi-AS VPLS. The two leading IETF working groups provide different proposals for Inter-Domain VPLS. The LDP working group barely addresses the issue and proposes to use gateways to connect two VPLS networks using only one LSP tunnel among them. Then, PWs are setup over the LSP tunnel, one PW per-VPLS service. However, the proposal does not provide any additional details about these gateways and is left as future work. On the other hand the BGP working group deepens a little bit more in the issue. They also propose, among other things, to use gateways and EBGP (External BGP) sessions among them. This method requires IBGP sessions between the PEs and the gateway within each AS, and an EBGP session between the gateways. From the signalling point of view, the distribution of labels goes from one PE device, to its corresponding gateway, then to the remote gateway, and then to a PE device within the neighbouring AS. In this framework, the peering gateways need a MPLS based connection in order to exchange label information. Additionally, it is important to notice that both gateways participate in the distribution of VPLS information, so the gateways need to be VPLS aware. Finally, a loop free topology relies on BGP, thus there is no need to run STP for each inter-domain VPLS instance.

4) VPLS Deployment Scenarios

One of the most relevant issues when deploying VPLS services is when to use switches and when to use routers as CE. On the one hand, switches typically cost less than routers, and are easier to manage. On the other hand, routers are more flexible, provide much more management features than switches, and whereas switches cannot act as routers, many routers are able to behave as legacy L2 switching devices. Furthermore, when some sort of QoS is needed, routers are able to provide a broader set of tools to police and shape traffic. Even though many switches are able to police traffic, they do not have the same policing capabilities than routers do. Routers are able to police traffic based on IP ToS, DSCP, TCP port, UDP port, and more, even when acting as L2 switches, which certainly offers more granular policing schemes. Therefore, when deploying VPLS services, a trade-off between the flexibility and capabilities to manage QoS per VPLS-instance, and the cost to do it exists.

IV. OPEN ISSUES AND FUTURE WORK

There are several issues surrounding VPLS implementations which are yet to be determined and improved. We can divide these issues into two categories, those that are under discussion right now, and those that have not been addressed yet.

Among those under discussion at the present we can find first, issues regarding VPLS scalability, which in turn could be divided into two different problems. One related to the amount of MAC address learning needed for large numbers of customer endpoints, and another focusing on the inefficiency in the broadcast replication process that it must be performed whenever transmitting broadcast, multicast or unknown MAC address Ethernet frames. While both leading IETF workgroups suggest H-VPLS as the key to address the former one, surprisingly, the latter is scarcely addressed in either of them. Even though H-VPLS arises as a strong solution to bound the MAC FIB boom, a lot of work still needs to be done, mainly in order to provide resilient H-VPLS topologies. For instance, in multi-homing schemes, how to accomplish fast convergence upon failures, how to implement hello messages for this task, when or when not STP is a feasible solution, just to name a few.

On the other hand, the flooding issue requires a solution that decouples the n-PE devices from the replication process, while trying at the same time not to waste unnecessary backbone bandwidth. In Figure 4 this could be achieved if an additional pair (QinQ tag)/label exists in order to allow broadcasting within each VPLS domain attached to a u-PE device, instead of broadcasting on a per-VPLS customer site basis. This is feasible since once a broadcast frame belonging to a given VPLS domain arrives to a u-PE bridging capable, the switch is able to flood the frame throughout all the necessary access ports as any legacy L2 switch would do.

Another important issue refers to how VPLS domains can spread across autonomous system boundaries (Inter-Domain VPLS). A far more complex situation than the one presented in the previous section occurs when the VPLS

services are required among non-neighbouring AS, in other words, the VPLS peering AS span across several AS hops. In this case one or more transit network operators are needed. The functionalities that these transit operators should provide are presently under discussion. An alternative approach is to implement EBGP multi-hop between peering PE devices from each AS. While this proposal completely releases intermediate gateways from managing VPLS information in both the control and data VPLS planes, they still need to establish LSP tunnels among remote PEs.

An additional point is that there is still no standard for VPLS. As aforementioned, this is mainly due to the ongoing discussion concerning the discovery and signalling processes. Therefore, until a standard is reached VPLS is an open issue itself.

Other issues under discussion are security, and management of VPLS networks.

Among the issues not addressed yet, perhaps the most important is the interaction of VPLS with the future optical networks. At the present, a combination of LDP signalling with maybe, MP-BGP auto discovery appears as a strong candidate to reach a standard in the framework of IP/MPLS based services. However, the trend towards IP and Optical integration shows GMPLS with RSVP-TE and OSPF-TE as the strong candidate to deploy the future optical network services. It seems quite reasonable then to foresee how to migrate VPLS IP/MPLS services to a GMPLS network core. Nowadays IP/MPLS application services such as H-VPLS are able to run on top of GMPLS LSP. However, the real issue here is how to efficiently transport Metro Ethernet VPNs services like VPLS over GMPLS, and what challenges this will bring since the establishment of PWs will probably rely on LDP signalling.

V. CONCLUSIONS

Altogether, VPLS is a novel technology that arises as a very promising source of revenue for service providers. This is mainly because it leverages the reutilization of many service provider core networks, while at the same time offers an attractive access solution to end users since Ethernet has proven to be the most widely accepted and cost-effective local area network technology. Presently, great efforts are being done in order to standardize IP/MPLS VPLS services, however even if a standard arises there are still several issues that need to be improved. Moreover, the present proposals for the VPLS control plane in IP/MPLS based networks may not be aligned with the signalling tendency in the future optical networks.

REFERENCES

[1] Martini, L., et al, “Encapsulation Methods for Transport of Ethernet Frames Over IP/MPLS Networks", Internet draft, work in progress, February 2003.

[2] Martini, L., et al, "Transport of Layer 2 Frames over MPLS", Internet draft, work in progress, February 2003.

[3] Kireeti Kompella, “Virtual Private LAN Service”, Internet draft, work in progress, January 2004.

[4] Marc Lasserre, Ali Sajassi, et al, “Virtual Private LAN Services over MPLS”, Internet draft, work in progress, November 2003.

[5] Juha Heinanen, et al, “Radius/L2TP Based VPLS”, Internet draft, work in progress, January 2004.

[6] Sangli, S., D. Tappan, and Y. Rekhter, "BGP Extended Communities Attribute", Internet draft, work in progress.