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

Status of This Memo

Copyright Notice

IESG Note

   The L2VPN Working Group produced two separate documents, RFC 4761 and
   this document, that perform similar functions using different
   signaling protocols. Be aware that each method is commonly referred
   to as "VPLS" even though they are distinct and incompatible with one
   another.

Abstract

   This document describes a Virtual Private LAN Service (VPLS) solution
   using pseudowires, a service previously implemented over other
   tunneling technologies and known as Transparent LAN Services (TLS).
   A VPLS creates an emulated LAN segment for a given set of users;
   i.e., it creates a Layer 2 broadcast domain that is fully capable of
   learning and forwarding on Ethernet MAC addresses and that is closed
   to a given set of users. Multiple VPLS services can be supported
   from a single Provider Edge (PE) node.

   This document describes the control plane functions of signaling
   pseudowire labels using Label Distribution Protocol (LDP), extending
   RFC 4447. It is agnostic to discovery protocols. The data plane
   functions of forwarding are also described, focusing in particular on
   the learning of MAC addresses. The encapsulation of VPLS packets is
   described by RFC 4448.

Table of Contents

1. Introduction

   Ethernet has become the predominant technology for Local Area Network
   (LAN) connectivity and is gaining acceptance as an access technology,
   specifically in Metropolitan and Wide Area Networks (MAN and WAN,
   respectively). The primary motivation behind Virtual Private LAN
   Services (VPLS) is to provide connectivity between geographically
   dispersed customer sites across MANs and WANs, as if they were
   connected using a LAN. The intended application for the end-user can
   be divided into the following two categories:

   - Connectivity between customer routers: LAN routing application

   - Connectivity between customer Ethernet switches: LAN switching
      application

   Broadcast and multicast services are available over traditional LANs.
   Sites that belong to the same broadcast domain and that are connected
   via an MPLS network expect broadcast, multicast, and unicast traffic
   to be forwarded to the proper location(s). This requires MAC address
   learning/aging on a per-pseudowire basis, and packet replication
   across pseudowires for multicast/broadcast traffic and for flooding
   of unknown unicast destination traffic.

   [RFC4448] defines how to carry Layer 2 (L2) frames over point-to-
   point pseudowires (PW). This document describes extensions to
   [RFC4447] for transporting Ethernet/802.3 and VLAN [802.1Q] traffic
   across multiple sites that belong to the same L2 broadcast domain or
   VPLS. Note that the same model can be applied to other 802.1
   technologies. It describes a simple and scalable way to offer
   Virtual LAN services, including the appropriate flooding of
   broadcast, multicast, and unknown unicast destination traffic over
   MPLS, without the need for address resolution servers or other
   external servers, as discussed in [L2VPN-REQ].

   The following discussion applies to devices that are VPLS capable and
   have a means of tunneling labeled packets amongst each other. The
   resulting set of interconnected devices forms a private MPLS VPN.

2. Terminology

   Q-in-Q   802.1ad Provider Bridge extensions also known
                       as stackable VLANs or Q-in-Q.

   Qualified learning Learning mode in which each customer VLAN is
                       mapped to its own VPLS instance.

Service delimiter   Information used to identify a specific customer
                      service instance. This is typically encoded in
                      the encapsulation header of customer frames
                      (e.g., VLAN Id).

Tagged frame   Frame with an 802.1Q VLAN identifier.

Unqualified learning Learning mode where all the VLANs of a single
                      customer are mapped to a single VPLS.

Untagged frame   Frame without an 802.1Q VLAN identifier.

[2.1]. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this
document are to be interpreted as described in [RFC 2119] [[RFC2119]].

[3]. Acronyms

AC   Attachment Circuit

BPDU   Bridge Protocol Data Unit

CE   Customer Edge device

FEC   Forwarding Equivalence Class

FIB   Forwarding Information Base

GRE   Generic Routing Encapsulation

IPsec   IP security

L2TP   Layer Two Tunneling Protocol

LAN   Local Area Network

LDP   Label Distribution Protocol

MTU-s   Multi-Tenant Unit switch

PE   Provider Edge device

PW   Pseudowire

STP   Spanning Tree Protocol

VLAN   Virtual LAN

VLAN tag   VLAN Identifier

4. Topological Model for VPLS

An interface participating in a VPLS must be able to flood, forward,
and filter Ethernet frames. Figure 1, below, shows the topological
model of a VPLS. The set of PE devices interconnected via PWs
appears as a single emulated LAN to customer X. Each PE will form
remote MAC address to PW associations and associate directly attached
MAC addresses to local customer facing ports. This is modeled on
standard IEEE 802.1 MAC address learning.

```
   +-----+                                           +-----+
   | CE1 +---+      ...........................    +---| CE2 |
   +-----+   |        .                    .       |   +-----+
    Site 1   |   +----+                +----+    |    Site 2
          +---| PE |       Cloud       | PE |---+
             +----+                +----+
              .                    .
              .         +----+        .
              ..........| PE |...........
                        +----+         ^
                         |             |
                         |             +-- Emulated LAN
                        +-----+
                        | CE3 |
                        +-----+
                        Site 3
```
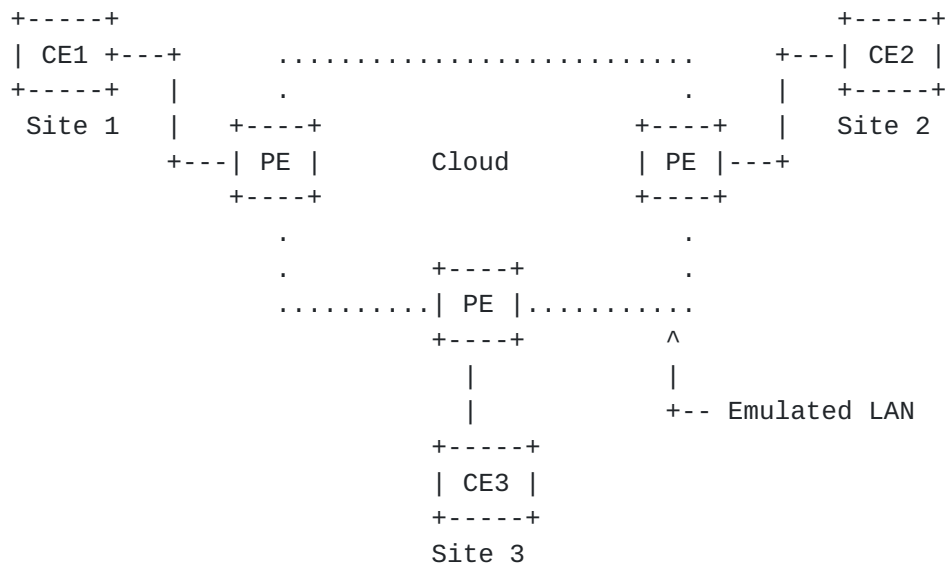
Figure 1: Topological Model of a VPLS for
Customer X with three sites

We note here again that while this document shows specific examples
using MPLS transport tunnels, other tunnels that can be used by PWs
(as mentioned in [RFC4447]) -- e.g., GRE, L2TP, IPsec -- can also be
used, as long as the originating PE can be identified, since this is
used in the MAC learning process.

The scope of the VPLS lies within the PEs in the service provider
network, highlighting the fact that apart from customer service
delineation, the form of access to a customer site is not relevant to
the VPLS [L2VPN-REQ]. In other words, the attachment circuit (AC)
connected to the customer could be a physical Ethernet port, a
logical (tagged) Ethernet port, an ATM PVC carrying Ethernet frames,
etc., or even an Ethernet PW.

The PE is typically an edge router capable of running the LDP
signaling protocol and/or routing protocols to set up PWs. In
addition, it is capable of setting up transport tunnels to other PEs
and delivering traffic over PWs.

4.1. Flooding and Forwarding

One of attributes of an Ethernet service is that frames sent to
broadcast addresses and to unknown destination MAC addresses are
flooded to all ports. To achieve flooding within the service
provider network, all unknown unicast, broadcast and multicast frames
are flooded over the corresponding PWs to all PE nodes participating
in the VPLS, as well as to all ACs.

Note that multicast frames are a special case and do not necessarily
have to be sent to all VPN members. For simplicity, the default
approach of broadcasting multicast frames is used.

To forward a frame, a PE MUST be able to associate a destination MAC
address with a PW. It is unreasonable and perhaps impossible to
require that PEs statically configure an association of every
possible destination MAC address with a PW. Therefore, VPLS-capable
PEs SHOULD have the capability to dynamically learn MAC addresses on
both ACs and PWs and to forward and replicate packets across both ACs
and PWs.

4.2. Address Learning

Unlike BGP VPNs [RFC4364], reachability information is not advertised
and distributed via a control plane. Reachability is obtained by
standard learning bridge functions in the data plane.

When a packet arrives on a PW, if the source MAC address is unknown,
it needs to be associated with the PW, so that outbound packets to
that MAC address can be delivered over the associated PW. Likewise,
when a packet arrives on an AC, if the source MAC address is unknown,
it needs to be associated with the AC, so that outbound packets to
that MAC address can be delivered over the associated AC.

Standard learning, filtering, and forwarding actions, as defined in
[802.1D-ORIG], [802.1D-REV], and [802.1Q], are required when a PW or
AC state changes.

[4.3](#). Tunnel Topology

PE routers are assumed to have the capability to establish transport
tunnels. Tunnels are set up between PEs to aggregate traffic. PWs
are signaled to demultiplex encapsulated Ethernet frames from
multiple VPLS instances that traverse the transport tunnels.

In an Ethernet L2VPN, it becomes the responsibility of the service
provider to create the loop-free topology. For the sake of
simplicity, we define that the topology of a VPLS is a full mesh of
PWs.

[4.4](#). Loop free VPLS

If the topology of the VPLS is not restricted to a full mesh, then it
may be that for two PEs not directly connected via PWs, they would
have to use an intermediary PE to relay packets. This topology would
require the use of some loop-breaking protocol, like a spanning tree
protocol.

Instead, a full mesh of PWs is established between PEs. Since every
PE is now directly connected to every other PE in the VPLS via a PW,
there is no longer any need to relay packets, and we can instantiate
a simpler loop-breaking rule: the "split horizon" rule, whereby a PE
MUST NOT forward traffic from one PW to another in the same VPLS
mesh.

Note that customers are allowed to run a Spanning Tree Protocol (STP)
(e.g., as defined in [[802.1D-REV](#)]), such as when a customer has "back
door" links used to provide redundancy in the case of a failure
within the VPLS. In such a case, STP Bridge PDUs (BPDUs) are simply
tunneled through the provider cloud.

[5](#). Discovery

The capability to manually configure the addresses of the remote PEs
is REQUIRED. However, the use of manual configuration is not
necessary if an auto-discovery procedure is used. A number of auto-
discovery procedures are compatible with this document
([[RADIUS-DISC](#)], [[BGP-DISC](#)]).

[6](#). Control Plane

This document describes the control plane functions of signaling of
PW labels. Some foundational work in the area of support for multi-
homing is laid. The extensions to provide multi-homing support
should work independently of the basic VPLS operation, and they are
not described here.

6.1. **LDP-Based Signaling of Demultiplexers**

   A full mesh of LDP sessions is used to establish the mesh of PWs.
   The requirement for a full mesh of PWs may result in a large number
   of targeted LDP sessions. Section 10 discusses the option of setting
   up hierarchical topologies in order to minimize the size of the VPLS
   full mesh.

   Once an LDP session has been formed between two PEs, all PWs between
   these two PEs are signaled over this session.

   In [RFC4447], two types of FECs are described: the PWid FEC Element
   (FEC type 128) and the Generalized PWid FEC Element (FEC type 129).
   The original FEC element used for VPLS was compatible with the PWid
   FEC Element. The text for signaling using the PWid FEC Element has
   been moved to Appendix A. What we describe below replaces that with
   a more generalized L2VPN descriptor, the Generalized PWid FEC
   Element.

6.1.1. **Using the Generalized PWid FEC Element**

   [RFC4447] describes a generalized FEC structure that is be used for
   VPLS signaling in the following manner. We describe the assignment
   of the Generalized PWid FEC Element fields in the context of VPLS
   signaling.

   Control bit (C): This bit is used to signal the use of the control
   word as specified in [RFC4447].

   PW type: The allowed PW types are Ethernet (0x0005) and Ethernet
   tagged mode (0x004), as specified in [RFC4446].

   PW info length: As specified in [RFC4447].

   Attachment Group Identifier (AGI), Length, Value: The unique name of
   this VPLS. The AGI identifies a type of name, and Length denotes the
   length of Value, which is the name of the VPLS. We use the term AGI
   interchangeably with VPLS identifier.

   Target Attachment Individual Identifier (TAII), Source Attachment
   Individual Identifier (SAII): These are null because the mesh of PWs
   in a VPLS terminates on MAC learning tables, rather than on
   individual attachment circuits. The use of non-null TAII and SAII is
   reserved for future enhancements.

Interface Parameters: The relevant interface parameters are:

- MTU: The MTU (Maximum Transmission Unit) of the VPLS MUST be the
  same across all the PWs in the mesh.

- Optional Description String: Same as [RFC4447].

- Requested VLAN ID: If the PW type is Ethernet tagged mode, this
  parameter may be used to signal the insertion of the appropriate
  VLAN ID, as defined in [RFC4448].

6.2. MAC Address Withdrawal

It MAY be desirable to remove or unlearn MAC addresses that have been
dynamically learned for faster convergence. This is accomplished by
sending an LDP Address Withdraw Message with the list of MAC
addresses to be removed to all other PEs over the corresponding LDP
sessions.

We introduce an optional MAC List TLV in LDP to specify a list of MAC
addresses that can be removed or unlearned using the LDP Address
Withdraw Message.

The Address Withdraw message with MAC List TLVs MAY be supported in
order to expedite removal of MAC addresses as the result of a
topology change (e.g., failure of the primary link for a dual-homed
VPLS-capable switch).

In order to minimize the impact on LDP convergence time, when the MAC
list TLV contains a large number of MAC addresses, it may be
preferable to send a MAC address withdrawal message with an empty
list.

6.2.1. MAC List TLV

MAC addresses to be unlearned can be signaled using an LDP Address
Withdraw Message that contains a new TLV, the MAC List TLV. Its
format is described below. The encoding of a MAC List TLV address is
the 6-octet MAC address specified by IEEE 802 documents [802.1D-ORIG]
[802.1D-REV].

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|F|       Type            |              Length               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       MAC address #1                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        MAC address #1          |        MAC Address #2         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       MAC address #2                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                          ...                                  ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       MAC address #n                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        MAC address #n          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

U bit: Unknown bit. This bit MUST be set to 1. If the MAC address
format is not understood, then the TLV is not understood and MUST be
ignored.

F bit: Forward bit. This bit MUST be set to 0. Since the LDP
mechanism used here is targeted, the TLV MUST NOT be forwarded.

Type: Type field. This field MUST be set to 0x0404. This identifies
the TLV type as MAC List TLV.

Length: Length field. This field specifies the total length in
octets of the MAC addresses in the TLV. The length MUST be a
multiple of 6.

MAC Address: The MAC address(es) being removed.

The MAC Address Withdraw Message contains a FEC TLV (to identify the
VPLS affected), a MAC Address TLV, and optional parameters. No
optional parameters have been defined for the MAC Address Withdraw
signaling. Note that if a PE receives a MAC Address Withdraw Message
and does not understand it, it MUST ignore the message. In this
case, instead of flushing its MAC address table, it will continue to
use stale information, unless:

- it receives a packet with a known MAC address association, but
   from a different PW, in which case it replaces the old
   association; or

- it ages out the old association.

The MAC Address Withdraw message only helps speed up convergence, so
PEs that do not understand the message can continue to participate in
the VPLS.

6.2.2. **Address Withdraw Message Containing MAC List TLV**

The processing for MAC List TLV received in an Address Withdraw
Message is:

For each MAC address in the TLV:

- Remove the association between the MAC address and the AC or PW
   over which this message is received.

For a MAC Address Withdraw message with empty list:

- Remove all the MAC addresses associated with the VPLS instance
   (specified by the FEC TLV) except the MAC addresses learned over
   the PW associated with this signaling session over which the
   message was received.

The scope of a MAC List TLV is the VPLS specified in the FEC TLV in
the MAC Address Withdraw Message. The number of MAC addresses can be
deduced from the length field in the TLV.

7. **Data Forwarding on an Ethernet PW**

This section describes the data plane behavior on an Ethernet PW used
in a VPLS. While the encapsulation is similar to that described in
[RFC4448], the functions of stripping the service-delimiting tag and
using a "normalized" Ethernet frame are described.

7.1. **VPLS Encapsulation Actions**

In a VPLS, a customer Ethernet frame without preamble is encapsulated
with a header as defined in [RFC4448]. A customer Ethernet frame is
defined as follows:

- If the frame, as it arrives at the PE, has an encapsulation that
   is used by the local PE as a service delimiter, i.e., to identify
   the customer and/or the particular service of that customer, then
   that encapsulation may be stripped before the frame is sent into
   the VPLS. As the frame exits the VPLS, the frame may have a
   service-delimiting encapsulation inserted.

   - If the frame, as it arrives at the PE, has an encapsulation that
     is not service delimiting, then it is a customer frame whose
     encapsulation should not be modified by the VPLS. This covers,
     for example, a frame that carries customer-specific VLAN tags that
     the service provider neither knows about nor wants to modify.

   As an application of these rules, a customer frame may arrive at a
   customer-facing port with a VLAN tag that identifies the customer's
   VPLS instance. That tag would be stripped before it is encapsulated
   in the VPLS. At egress, the frame may be tagged again, if a
   service-delimiting tag is used, or it may be untagged if none is
   used.

   Likewise, if a customer frame arrives at a customer-facing port over
   an ATM or Frame Relay VC that identifies the customer's VPLS
   instance, then the ATM or FR encapsulation is removed before the
   frame is passed into the VPLS.

   Contrariwise, if a customer frame arrives at a customer-facing port
   with a VLAN tag that identifies a VLAN domain in the customer L2
   network, then the tag is not modified or stripped, as it belongs with
   the rest of the customer frame.

   By following the above rules, the Ethernet frame that traverses a
   VPLS is always a customer Ethernet frame. Note that the two actions,
   at ingress and egress, of dealing with service delimiters are local
   actions that neither PE has to signal to the other. They allow, for
   example, a mix-and-match of VLAN tagged and untagged services at
   either end, and they do not carry across a VPLS a VLAN tag that has
   local significance only. The service delimiter may be an MPLS label
   also, whereby an Ethernet PW given by [RFC4448] can serve as the
   access side connection into a PE. An RFC1483 Bridged PVC
   encapsulation could also serve as a service delimiter. By limiting
   the scope of locally significant encapsulations to the edge,
   hierarchical VPLS models can be developed that provide the capability
   to network-engineer scalable VPLS deployments, as described below.

7.2. VPLS Learning Actions

   Learning is done based on the customer Ethernet frame as defined
   above. The Forwarding Information Base (FIB) keeps track of the
   mapping of customer Ethernet frame addressing and the appropriate PW
   to use. We define two modes of learning: qualified and unqualified
   learning. Qualified learning is the default mode and MUST be
   supported. Support of unqualified learning is OPTIONAL.

In unqualified learning, all the VLANs of a single customer are
handled by a single VPLS, which means they all share a single
broadcast domain and a single MAC address space. This means that MAC
addresses need to be unique and non-overlapping among customer VLANs,
or else they cannot be differentiated within the VPLS instance, and
this can result in loss of customer frames. An application of
unqualified learning is port-based VPLS service for a given customer
(e.g., customer with non-multiplexed AC where all the traffic on a
physical port, which may include multiple customer VLANs, is mapped
to a single VPLS instance).

In qualified learning, each customer VLAN is assigned to its own VPLS
instance, which means each customer VLAN has its own broadcast domain
and MAC address space. Therefore, in qualified learning, MAC
addresses among customer VLANs may overlap with each other, but they
will be handled correctly since each customer VLAN has its own FIB;
i.e., each customer VLAN has its own MAC address space. Since VPLS
broadcasts multicast frames by default, qualified learning offers the
advantage of limiting the broadcast scope to a given customer VLAN.
Qualified learning can result in large FIB table sizes, because the
logical MAC address is now a VLAN tag + MAC address.

For STP to work in qualified learning mode, a VPLS PE must be able to
forward STP BPDUs over the proper VPLS instance. In a hierarchical
VPLS case (see details in Section 10), service delimiting tags
(Q-in-Q or [RFC4448]) can be added such that PEs can unambiguously
identify all customer traffic, including STP BPDUs. In a basic VPLS
case, upstream switches must insert such service delimiting tags.
When an access port is shared among multiple customers, a reserved
VLAN per customer domain must be used to carry STP traffic. The STP
frames are encapsulated with a unique provider tag per customer (as
the regular customer traffic), and a PEs looks up the provider tag to
send such frames across the proper VPLS instance.

8. Data Forwarding on an Ethernet VLAN PW

This section describes the data plane behavior on an Ethernet VLAN PW
in a VPLS. While the encapsulation is similar to that described in
[RFC4448], the functions of imposing tags and using a "normalized"
Ethernet frame are described. The learning behavior is the same as
for Ethernet PWs.

8.1. VPLS Encapsulation Actions

In a VPLS, a customer Ethernet frame without preamble is encapsulated
with a header as defined in [RFC4448]. A customer Ethernet frame is
defined as follows:

   - If the frame, as it arrives at the PE, has an encapsulation that
      is part of the customer frame and is also used by the local PE as
      a service delimiter, i.e., to identify the customer and/or the
      particular service of that customer, then that encapsulation is
      preserved as the frame is sent into the VPLS, unless the Requested
      VLAN ID optional parameter was signaled. In that case, the VLAN
      tag is overwritten before the frame is sent out on the PW.

   - If the frame, as it arrives at the PE, has an encapsulation that
      does not have the required VLAN tag, a null tag is imposed if the
      Requested VLAN ID optional parameter was not signaled.

   As an application of these rules, a customer frame may arrive at a
   customer-facing port with a VLAN tag that identifies the customer's
   VPLS instance and also identifies a customer VLAN. That tag would be
   preserved as it is encapsulated in the VPLS.

   The Ethernet VLAN PW provides a simple way to preserve customer
   802.1p bits.

   A VPLS MAY have both Ethernet and Ethernet VLAN PWs. However, if a
   PE is not able to support both PWs simultaneously, it SHOULD send a
   Label Release on the PW messages that it cannot support with a status
   code "Unknown FEC" as given in [RFC3036].

9. Operation of a VPLS

   We show here, in Figure 2, below, an example of how a VPLS works.
   The following discussion uses the figure below, where a VPLS has been
   set up between PE1, PE2, and PE3. The VPLS connects a customer with
   4 sites labeled A1, A2, A3, and A4 through CE1, CE2, CE3, and CE4,
   respectively.

   Initially, the VPLS is set up so that PE1, PE2, and PE3 have a full
   mesh of Ethernet PWs. The VPLS instance is assigned an identifier
   (AGI). For the above example, say PE1 signals PW label 102 to PE2
   and 103 to PE3, and PE2 signals PW label 201 to PE1 and 203 to PE3.

```
                                               -----
                                             /  A1  \
        ----                             ----CE1     |
       /    \      --------      ------- /     |     |
      | A2 CE2-     /      \     /      PE1     \     /
       \    /  \   /         \---/       \       -----
        ----     ---PE2                  |
                 | Service Provider Network |
                    \         /   \         /
           -----  PE3       /      \       /
           |Agg|_/  --------       -------
          -|   |
     ----  / -----   ----
    /    \/    \   /    \           CE = Customer Edge Router
   | A3 CE3    -CE4 A4 |            PE = Provider Edge Router
    \    /        \    /            Agg = Layer 2 Aggregation
     ----          ----
```
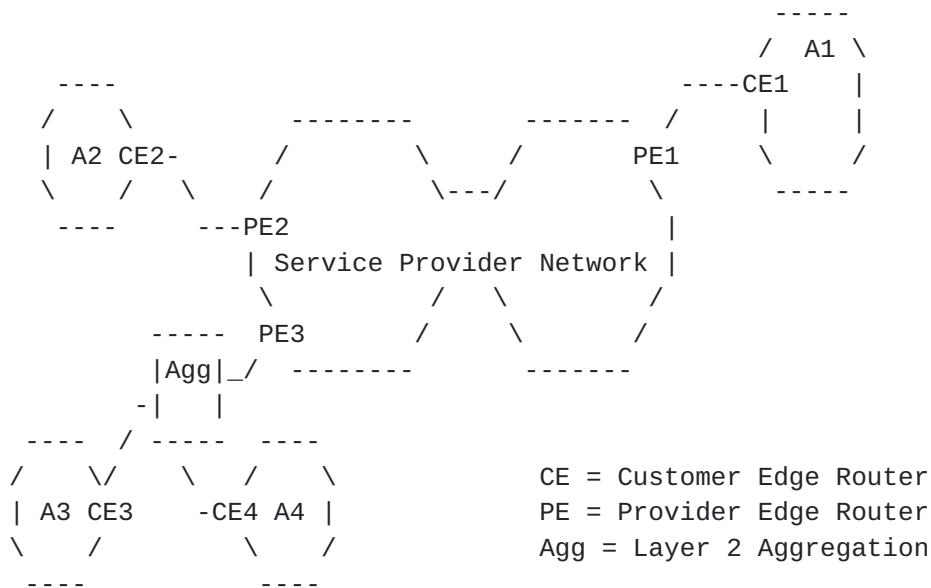
                     Figure 2: Example of a VPLS

   Assume a packet from A1 is bound for A2. When it leaves CE1, say it
   has a source MAC address of M1 and a destination MAC of M2. If PE1
   does not know where M2 is, it will flood the packet; i.e., send it to
   PE2 and PE3. When PE2 receives the packet, it will have a PW label
   of 201. PE2 can conclude that the source MAC address M1 is behind
   PE1, since it distributed the label 201 to PE1. It can therefore
   associate MAC address M1 with PW label 102.

9.1. MAC Address Aging

   PEs that learn remote MAC addresses SHOULD have an aging mechanism to
   remove unused entries associated with a PW label. This is important
   both for conservation of memory and for administrative purposes. For
   example, if a customer site A, is shut down, eventually the other PEs
   should unlearn A's MAC address.

   The aging timer for MAC address M SHOULD be reset when a packet with
   source MAC address M is received.

10. **A Hierarchical VPLS Model**

   The solution described above requires a full mesh of tunnel LSPs
   between all the PE routers that participate in the VPLS service. For
   each VPLS service, n*(n-1)/2 PWs must be set up between the PE
   routers. While this creates signaling overhead, the real detriment
   to large scale deployment is the packet replication requirements for
   each provisioned PWs on a PE router. Hierarchical connectivity,
   described in this document, reduces signaling and replication
   overhead to allow large-scale deployment.

   In many cases, service providers place smaller edge devices in
   multi-tenant buildings and aggregate them into a PE in a large
   Central Office (CO) facility. In some instances, standard IEEE
   802.1q (Dot 1Q) tagging techniques may be used to facilitate mapping
   CE interfaces to VPLS access circuits at a PE.

   It is often beneficial to extend the VPLS service tunneling
   techniques into the access switch domain. This can be accomplished
   by treating the access device as a PE and provisioning PWs between it
   and every other edge, as a basic VPLS. An alternative is to utilize
   [RFC4448] PWs or Q-in-Q logical interfaces between the access device
   and selected VPLS enabled PE routers. Q-in-Q encapsulation is
   another form of L2 tunneling technique, which can be used in
   conjunction with MPLS signaling, as will be described later. The
   following two sections focus on this alternative approach. The VPLS
   core PWs (hub) are augmented with access PWs (spoke) to form a two-
   tier hierarchical VPLS (H-VPLS).

   Spoke PWs may be implemented using any L2 tunneling mechanism, and by
   expanding the scope of the first tier to include non-bridging VPLS PE
   routers. The non-bridging PE router would extend a spoke PW from a
   Layer-2 switch that connects to it, through the service core network,
   to a bridging VPLS PE router supporting hub PWs. We also describe
   how VPLS-challenged nodes and low-end CEs without MPLS capabilities
   may participate in a hierarchical VPLS.

   For rest of this discussion we refer to a bridging capable access
   device as MTU-s and a non-bridging capable PE as PE-r. We refer to a
   routing and bridging capable device as PE-rs.

10.1. **Hierarchical Connectivity**

   This section describes the hub and spoke connectivity model and
   describes the requirements of the bridging capable and non-bridging
   MTU-s devices for supporting the spoke connections.

10.1.1. **Spoke Connectivity for Bridging-Capable Devices**

   In Figure 3, below, three customer sites are connected to an MTU-s
   through CE-1, CE-2, and CE-3. The MTU-s has a single connection
   (PW-1) to PE1-rs. The PE-rs devices are connected in a basic VPLS
   full mesh. For each VPLS service, a single spoke PW is set up
   between the MTU-s and the PE-rs based on [RFC4447]. Unlike
   traditional PWs that terminate on a physical (or a VLAN-tagged
   logical) port, a spoke PW terminates on a virtual switch instance
   (VSI; see [L2FRAME]) on the MTU-s and the PE-rs devices.

```
                                                    PE2-rs
                                                  +--------+
                                                  |        |
                                                  |   --   |
                                                  |  / \   |
          CE-1                                    |  \S /  |
            \                                     |   --   |
             \                                    +--------+
              \    MTU-s                  PE1-rs      /  |
               +--------+               +--------+   /   |
               |        |               |        |  /    |
               |   --   |    PW-1        |   --   |-/     |
               |  / \--|- - - - - - - - - -| / \   |      |
               |  \S / |               |  \S /  |       |
               |   --   |               |   --   |---\   |
               +--------+               +--------+    \  |
                 /                                     \ |
               ----                                 +--------+
              |Agg |                                |        |
               ----                                 |   --   |
              /     \                               |  / \   |
          CE-2   CE-3                               |  \S /  |
                                                    |   --   |
                                                    +--------+
                                                      PE3-rs

     Agg = Layer-2 Aggregation
      --
     /  \
     \S / = Virtual Switch Instance
      --
```
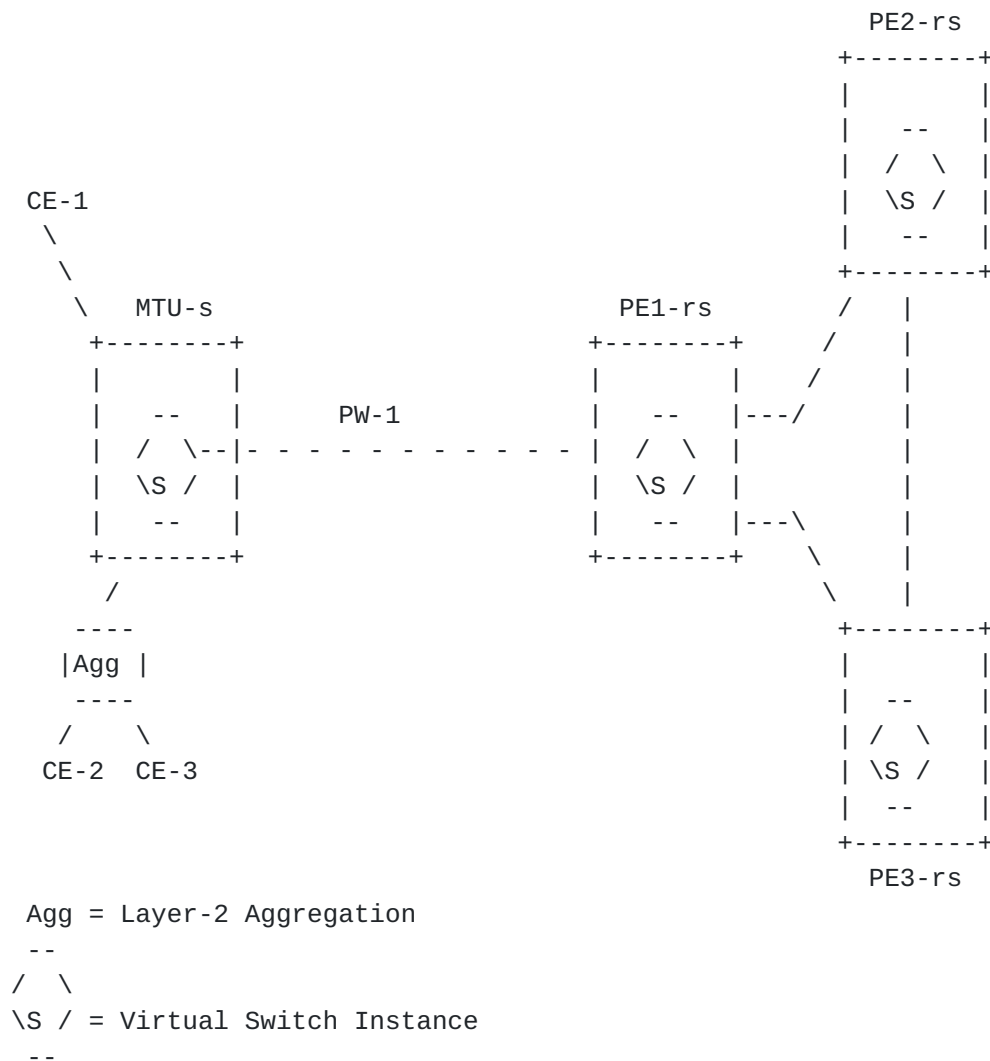
           Figure 3: An example of a hierarchical VPLS model

   The MTU-s and the PE-rs treat each spoke connection like an AC of the
   VPLS service. The PW label is used to associate the traffic from the
   spoke to a VPLS instance.

10.1.1.1. MTU-s Operation

   An MTU-s is defined as a device that supports layer-2 switching
   functionality and does all the normal bridging functions of learning
   and replication on all its ports, including the spoke, which is
   treated as a virtual port. Packets to unknown destinations are
   replicated to all ports in the service including the spoke. Once the
   MAC address is learned, traffic between CE1 and CE2 will be switched
   locally by the MTU-s, saving the capacity of the spoke to the PE-rs.
   Similarly traffic between CE1 or CE2 and any remote destination is
   switched directly onto the spoke and sent to the PE-rs over the
   point-to-point PW.

   Since the MTU-s is bridging capable, only a single PW is required per
   VPLS instance for any number of access connections in the same VPLS
   service. This further reduces the signaling overhead between the
   MTU-s and PE-rs.

   If the MTU-s is directly connected to the PE-rs, other encapsulation
   techniques, such as Q-in-Q, can be used for the spoke.

10.1.1.2. PE-rs Operation

   A PE-rs is a device that supports all the bridging functions for VPLS
   service and supports the routing and MPLS encapsulation; i.e., it
   supports all the functions described for a basic VPLS, as described
   above.

   The operation of PE-rs is independent of the type of device at the
   other end of the spoke. Thus, the spoke from the MTU-s is treated as
   a virtual port, and the PE-rs will switch traffic between the spoke
   PW, hub PWs, and ACs once it has learned the MAC addresses.

10.1.2. Advantages of Spoke Connectivity

   Spoke connectivity offers several scaling and operational advantages
   for creating large-scale VPLS implementations, while retaining the
   ability to offer all the functionality of the VPLS service.

   - Eliminates the need for a full mesh of tunnels and full mesh of
      PWs per service between all devices participating in the VPLS
      service.

   - Minimizes signaling overhead, since fewer PWs are required for the
      VPLS service.

   - Segments VPLS nodal discovery. MTU-s needs to be aware of only
     the PE-rs node, although it is participating in the VPLS service
     that spans multiple devices. On the other hand, every VPLS PE-rs
     must be aware of every other VPLS PE-rs and all of its locally
     connected MTU-s and PE-r devices.

   - Addition of other sites requires configuration of the new MTU-s
     but does not require any provisioning of the existing MTU-s
     devices on that service.

   - Hierarchical connections can be used to create VPLS service that
     spans multiple service provider domains. This is explained in a
     later section.

   Note that as more devices participate in the VPLS, there are more
   devices that require the capability for learning and replication.

10.1.3. Spoke Connectivity for Non-Bridging Devices

   In some cases, a bridging PE-rs may not be deployed, or a PE-r might
   already have been deployed. In this section, we explain how a PE-r
   that does not support any of the VPLS bridging functionality can
   participate in the VPLS service.

   In Figure 4, three customer sites are connected through CE-1, CE-2,
   and CE-3 to the VPLS through PE-r. For every attachment circuit that
   participates in the VPLS service, PE-r creates a point-to-point PW
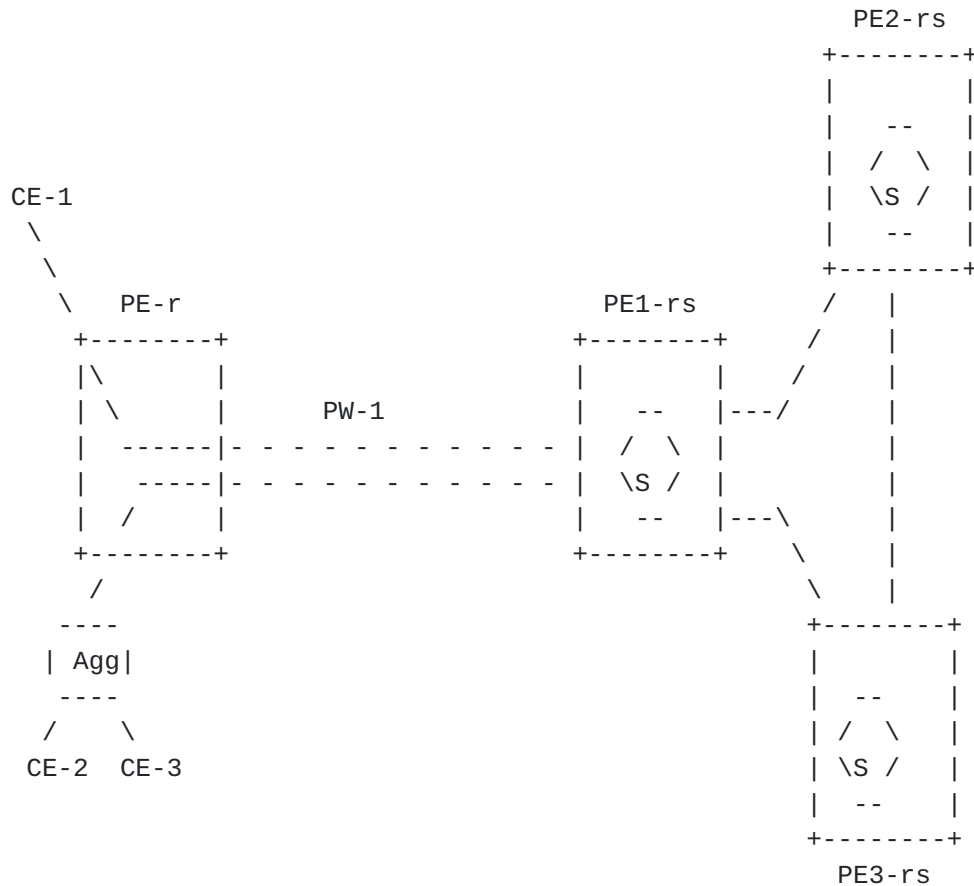   that terminates on the VSI of PE1-rs.

```
                                                   PE2-rs
                                                  +--------+
                                                  |        |
                                                  |   --   |
                                                  |  / \   |
                CE-1                              |  \S /  |
                  \                               |   --   |
                   \                              +--------+
                    \   PE-r              PE1-rs        /   |
                     +--------+          +--------+    /    |
                     |\       |          |        |   /     |
                     | \      |   PW-1   |   --   |--/       |
                     |  ------|- - - - - - - - - -| / \  |          |
                     |   -----|- - - - - - - - - -| \S /  |          |
                     |  /     |          |   --   |---\      |
                     +--------+          +--------+    \     |
                      /                                 \    |
                    ----                              +--------+
                   | Agg|                             |        |
                    ----                              |   --   |
                   /     \                            |  / \   |
                 CE-2   CE-3                          |  \S /  |
                                                      |   --   |
                                                      +--------+
                                                        PE3-rs
```

           Figure 4: An example of a hierarchical VPLS
                    with non-bridging spokes

   The PE-r is defined as a device that supports routing but does not
   support any bridging functions. However, it is capable of setting up
   PWs between itself and the PE-rs. For every port that is supported
   in the VPLS service, a PW is set up from the PE-r to the PE-rs. Once
   the PWs are set up, there is no learning or replication function
   required on the part of the PE-r. All traffic received on any of the
   ACs is transmitted on the PW. Similarly, all traffic received on a
   PW is transmitted to the AC where the PW terminates. Thus, traffic
   from CE1 destined for CE2 is switched at PE1-rs and not at PE-r.

   Note that in the case where PE-r devices use Provider VLANs (P-VLAN)
   as demultiplexers instead of PWs, PE1-rs can treat them as such and
   map these "circuits" into a VPLS domain to provide bridging support
   between them.

This approach adds more overhead than the bridging-capable (MTU-s)
spoke approach, since a PW is required for every AC that participates
in the service versus a single PW required per service (regardless of
ACs) when an MTU-s is used. However, this approach offers the
advantage of offering a VPLS service in conjunction with a routed
internet service without requiring the addition of new MTU-s.

10.2. Redundant Spoke Connections

An obvious weakness of the hub and spoke approach described thus far
is that the MTU-s has a single connection to the PE-rs. In case of
failure of the connection or the PE-rs, the MTU-s suffers total loss
of connectivity.

In this section, we describe how the redundant connections can be
provided to avoid total loss of connectivity from the MTU-s. The
mechanism described is identical for both, MTU-s and PE-r devices.

10.2.1. Dual-Homed MTU-s

To protect from connection failure of the PW or the failure of the
PE-rs, the MTU-s or the PE-r is dual-homed into two PE-rs devices.
The PE-rs devices must be part of the same VPLS service instance.

In Figure 5, two customer sites are connected through CE-1 and CE-2
to an MTU-s. The MTU-s sets up two PWs (one each to PE1-rs and
PE3-rs) for each VPLS instance. One of the two PWs is designated as
primary and is the one that is actively used under normal conditions,
whereas the second PW is designated as secondary and is held in a
standby state. The MTU-s negotiates the PW labels for both the
primary and secondary PWs, but does not use the secondary PW unless
the primary PW fails. How a spoke is designated primary or secondary
is outside the scope of this document. For example, a spanning tree
instance running between only the MTU-s and the two PE-rs nodes is
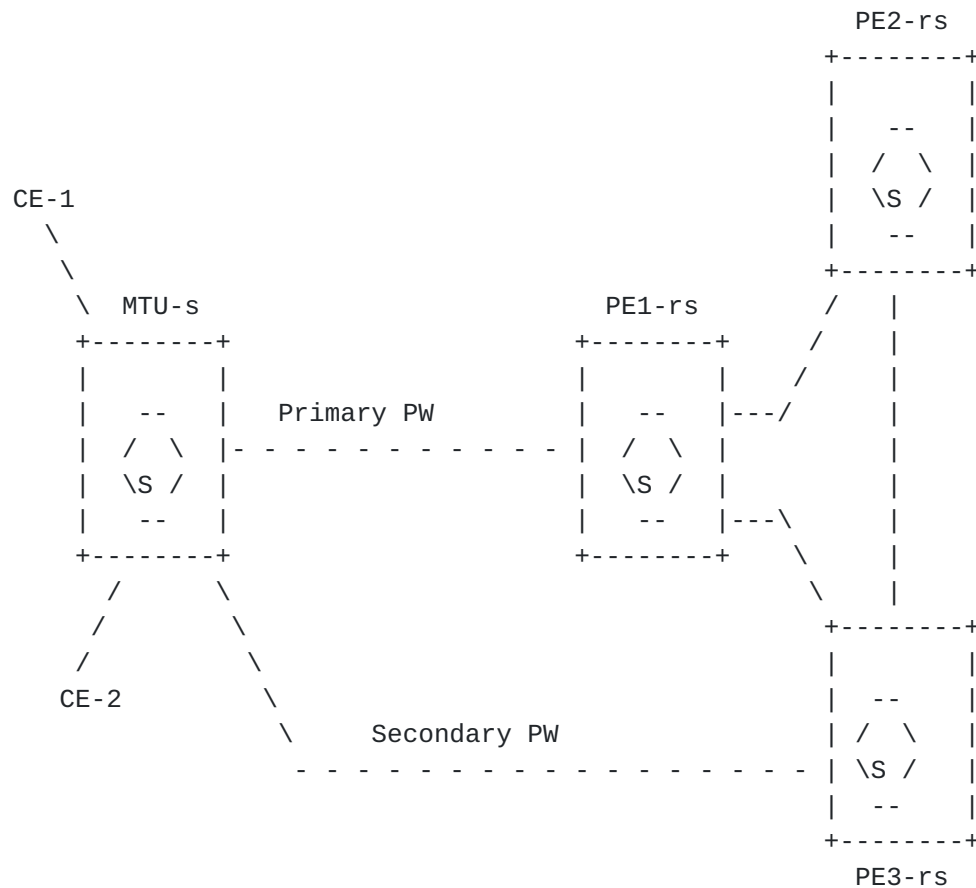one possible method. Another method could be configuration.

```
                                             PE2-rs
                                        +--------+
                                        |        |
                                        |   --   |
                                        |  / \   |
     CE-1                               |  \S /  |
       \                                |   --   |
        \                               +--------+
         \   MTU-s                  PE1-rs      /   |
          +--------+           +--------+     /     |
          |        |           |        |   /       |
          |   --   |  Primary PW    |   --   |---/      |
          |  / \   |- - - - - - - - - - - - |  / \   |          |
          |  \S /  |           |  \S /  |          |
          |   --   |           |   --   |---\      |
          +--------+           +--------+    \     |
            /      \                          \    |
           /        \                          \   |
          /          \                  +--------+
       CE-2           \                 |        |
                       \                |   --   |
                        \   Secondary PW     | / \   |
                         - - - - - - - - - - - - - - - - - -|  \S /  |
                                                |   --   |
                                                +--------+
                                                 PE3-rs
```

               Figure 5: An example of a dual-homed MTU-s

[10.2.2](#). **Failure Detection and Recovery**

   The MTU-s should control the usage of the spokes to the PE-rs
   devices. If the spokes are PWs, then LDP signaling is used to
   negotiate the PW labels, and the hello messages used for the LDP
   session could be used to detect failure of the primary PW. The use
   of other mechanisms that could provide faster detection failures is
   outside the scope of this document.

   Upon failure of the primary PW, MTU-s immediately switches to the
   secondary PW. At this point, the PE3-rs that terminates the
   secondary PW starts learning MAC addresses on the spoke PW. All
   other PE-rs nodes in the network think that CE-1 and CE-2 are behind
   PE1-rs and may continue to send traffic to PE1-rs until they learn
   that the devices are now behind PE3-rs. The unlearning process can
   take a long time and may adversely affect the connectivity of
   higher-level protocols from CE1 and CE2. To enable faster
   convergence, the PE3-rs where the secondary PW got activated may send
   out a flush message (as explained in [Section 6.2](#)), using the MAC List

TLV, as defined in Section 6, to all PE-rs nodes. Upon receiving the
message, PE-rs nodes flush the MAC addresses associated with that
VPLS instance.

10.3. Multi-domain VPLS Service

Hierarchy can also be used to create a large-scale VPLS service
within a single domain or a service that spans multiple domains
without requiring full mesh connectivity between all VPLS-capable
devices. Two fully meshed VPLS networks are connected together using
a single LSP tunnel between the VPLS "border" devices. A single
spoke PW per VPLS service is set up to connect the two domains
together.

When more than two domains need to be connected, a full mesh of
inter-domain spokes is created between border PEs. Forwarding rules
over this mesh are identical to the rules defined in Section 4.

This creates a three-tier hierarchical model that consists of a hub-
and-spoke topology between MTU-s and PE-rs devices, a full-mesh
topology between PE-rs, and a full mesh of inter-domain spokes
between border PE-rs devices.

This document does not specify how redundant border PEs per domain
per VPLS instance can be supported.

11. Hierarchical VPLS Model Using Ethernet Access Network

In this section, the hierarchical model is expanded to include an
Ethernet access network. This model retains the hierarchical
architecture discussed previously in that it leverages the full-mesh
topology among PE-rs devices; however, no restriction is imposed on
the topology of the Ethernet access network (e.g., the topology
between MTU-s and PE-rs devices is not restricted to hub and spoke).

The motivation for an Ethernet access network is that Ethernet-based
networks are currently deployed by some service providers to offer
VPLS services to their customers. Therefore, it is important to
provide a mechanism that allows these networks to integrate with an
IP or MPLS core to provide scalable VPLS services.

One approach of tunneling a customer's Ethernet traffic via an
Ethernet access network is to add an additional VLAN tag to the
customer's data (which may be either tagged or untagged). The
additional tag is referred to as Provider's VLAN (P-VLAN). Inside
the provider's network each P-VLAN designates a customer or more
specifically a VPLS instance for that customer. Therefore, there is
a one-to-one correspondence between a P-VLAN and a VPLS instance. In

this model, the MTU-s needs to have the capability of adding the
additional P-VLAN tag to non-multiplexed ACs where customer VLANs are
not used as service delimiters. This functionality is described in
[802.1ad].

If customer VLANs need to be treated as service delimiters (e.g., the
AC is a multiplexed port), then the MTU-s needs to have the
additional capability of translating a customer VLAN (C-VLAN) to a
P-VLAN, or to push an additional P-VLAN tag, in order to resolve
overlapping VLAN tags used by different customers. Therefore, the
MTU-s in this model can be considered a typical bridge with this
additional capability. This functionality is described in [802.1ad].

The PE-rs needs to be able to perform bridging functionality over the
standard Ethernet ports toward the access network, as well as over
the PWs toward the network core. In this model, the PE-rs may need
to run STP towards the access network, in addition to split-horizon
over the MPLS core. The PE-rs needs to map a P-VLAN to a VPLS-
instance and its associated PWs, and vice versa.

The details regarding bridge operation for MTU-s and PE-rs (e.g.,
encapsulation format for Q-in-Q messages, customer's Ethernet control
protocol handling, etc.) are outside the scope of this document and
are covered in [802.1ad]. However, the relevant part is the
interaction between the bridge module and the MPLS/IP PWs in the
PE-rs, which behaves just as in a regular VPLS.

11.1. Scalability

Since each P-VLAN corresponds to a VPLS instance, the total number of
VPLS instances supported is limited to 4K. The P-VLAN serves as a
local service delimiter within the provider's network that is
stripped as it gets mapped to a PW in a VPLS instance. Therefore,
the 4K limit applies only within an Ethernet access network (Ethernet
island) and not to the entire network. The SP network consists of a
core MPLS/IP network that connects many Ethernet islands. Therefore,
the number of VPLS instances can scale accordingly with the number of
Ethernet islands (a metro region can be represented by one or more
islands).

11.2. Dual Homing and Failure Recovery

In this model, an MTU-s can be dual homed to different devices
(aggregators and/or PE-rs devices). The failure protection for
access network nodes and links can be provided through running STP in
each island. The STP of each island is independent of other islands
and do not interact with others. If an island has more than one
PE-rs, then a dedicated full-mesh of PWs is used among these PE-rs

devices for carrying the SP BPDU packets for that island. On a
per-P-VLAN basis, STP will designate a single PE-rs to be used for
carrying the traffic across the core. The loop-free protection
through the core is performed using split-horizon, and the failure
protection in the core is performed through standard IP/MPLS re-
routing.

[12](#). Contributors

    Loa Andersson, TLA
    Ron Haberman, Alcatel-Lucent
    Juha Heinanen, Independent
    Giles Heron, Tellabs
    Sunil Khandekar, Alcatel-Lucent
    Luca Martini, Cisco
    Pascal Menezes, Independent
    Rob Nath, Alcatel-Lucent
    Eric Puetz, AT&T
    Vasile Radoaca, Independent
    Ali Sajassi, Cisco
    Yetik Serbest, AT&T
    Nick Slabakov, Juniper
    Andrew Smith, Consultant
    Tom Soon, AT&T
    Nick Tingle, Alcatel-Lucent

[13](#). Acknowledgments

    We wish to thank Joe Regan, Kireeti Kompella, Anoop Ghanwani, Joel
    Halpern, Bill Hong, Rick Wilder, Jim Guichard, Steve Phillips, Norm
    Finn, Matt Squire, Muneyoshi Suzuki, Waldemar Augustyn, Eric Rosen,
    Yakov Rekhter, Sasha Vainshtein, and Du Wenhua for their valuable
    feedback.

    We would also like to thank Rajiv Papneja (ISOCORE), Winston Liu
    (Ixia), and Charlie Hundall for identifying issues with the draft in
    the course of the interoperability tests.

    We would also like to thank Ina Minei, Bob Thomas, Eric Gray and
    Dimitri Papadimitriou for their thorough technical review of the
    document.

14. Security Considerations

A more comprehensive description of the security issues involved in
L2VPNs is covered in [RFC4111]. An unguarded VPLS service is
vulnerable to some security issues that pose risks to the customer
and provider networks. Most of the security issues can be avoided
through implementation of appropriate guards. A couple of them can
be prevented through existing protocols.

- Data plane aspects

    - Traffic isolation between VPLS domains is guaranteed by the
       use of per VPLS L2 FIB table and the use of per VPLS PWs.

    - The customer traffic, which consists of Ethernet frames, is
       carried unchanged over VPLS. If security is required, the
       customer traffic SHOULD be encrypted and/or authenticated
       before entering the service provider network.

    - Preventing broadcast storms can be achieved by using routers
       as CPE devices or by rate policing the amount of broadcast
       traffic that customers can send.

- Control plane aspects

    - LDP security (authentication) methods as described in
       [RFC3036] SHOULD be applied. This would prevent
       unauthenticated messages from disrupting a PE in a VPLS.

- Denial of service attacks

    - Some means to limit the number of MAC addresses (per site per
       VPLS) that a PE can learn SHOULD be implemented.

15. IANA Considerations

The type field in the MAC List TLV is defined as 0x404 in Section
6.2.1.

16. References

16.1. Normative References

[RFC4447]   Martini, L., Rosen, E., El-Aawar, N., Smith, T., and
               G. Heron, "Pseudowire Setup and Maintenance Using the
               Label Distribution Protocol (LDP)", RFC 4447, April
               2006.

[RFC4448]   Martini, L., Rosen, E., El-Aawar, N., and G. Heron,
               "Encapsulation Methods for Transport of Ethernet over
               MPLS Networks", RFC 4448, April 2006.

[802.1D-ORIG] Original 802.1D - ISO/IEC 10038, ANSI/IEEE Std
               802.1D-1993 "MAC Bridges".

[802.1D-REV]   802.1D - "Information technology - Telecommunications
               and information exchange between systems - Local and
               metropolitan area networks - Common specifications -
               Part 3: Media Access Control (MAC) Bridges: Revision.
               This is a revision of ISO/IEC 10038: 1993, 802.1j-1992
               and 802.6k-1992. It incorporates P802.11c, P802.1p
               and P802.12e." ISO/IEC 15802-3: 1998.

[802.1Q]   802.1Q - ANSI/IEEE Draft Standard P802.1Q/D11, "IEEE
               Standards for Local and Metropolitan Area Networks:
               Virtual Bridged Local Area Networks", July 1998.

[RFC3036]   Andersson, L., Doolan, P., Feldman, N., Fredette, A.,
               and B. Thomas, "LDP Specification", RFC 3036, January
               2001.

[RFC4446]   Martini, L., "IANA Allocations for Pseudowire Edge to
               Edge Emulation (PWE3)", BCP 116, RFC 4446, April 2006.

[RFC2119]   Bradner, S., "Key words for use in RFCs to Indicate
               Requirement Levels", BCP 14, RFC 2119, March 1997.

16.2. Informative References

[RFC4364]   Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private
               Networks (VPNs)", RFC 4364, February 2006.

[RADIUS-DISC] Heinanen, J., Weber, G., Ed., Townsley, W., Booth, S.,
               and W. Luo, "Using Radius for PE-Based VPN Discovery",
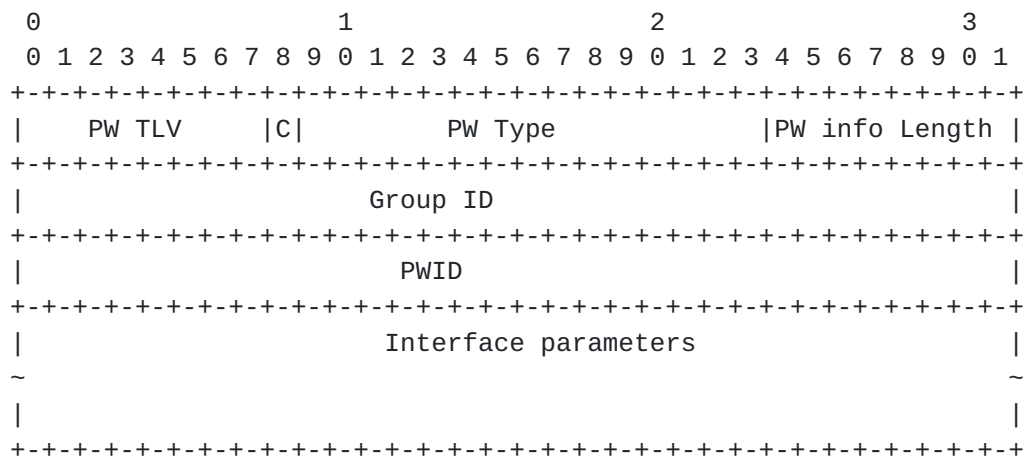   Work in Progress, October 2005.

   [BGP-DISC]   Ould-Brahim, H., Ed., Rosen, E., Ed., and Y. Rekhter,
                  Ed., "Using BGP as an Auto-Discovery Mechanism for
                  Network-based VPNs", Work in Progress, September 2006.

   [L2FRAME]   Andersson, L. and E. Rosen, "Framework for Layer 2
                  Virtual Private Networks (L2VPNs)", RFC 4664,
                  September 2006.

   [L2VPN-REQ]   Augustyn, W. and Y. Serbest, "Service Requirements for
                  Layer 2 Provider-Provisioned Virtual Private
                  Networks", RFC 4665, September 2006.

   [RFC4111]   Fang, L., "Security Framework for Provider-Provisioned
                  Virtual Private Networks (PPVPNs)", RFC 4111, July
                  2005.

   [802.1ad]   "IEEE standard for Provider Bridges", Work in
                  Progress, December 2002.

Appendix A. VPLS Signaling using the PWid FEC Element

   This section is being retained because live deployments use this
   version of the signaling for VPLS.

   The VPLS signaling information is carried in a Label Mapping message
   sent in downstream unsolicited mode, which contains the following
   PWid FEC TLV.

   PW, C, PW Info Length, Group ID, and Interface parameters are as
   defined in [RFC4447].

```
0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    PW TLV     |C|         PW Type         |PW info Length |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Group ID                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          PWID                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Interface parameters                  |
~                                                           ~
|                                                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   We use the Ethernet PW type to identify PWs that carry Ethernet
   traffic for multipoint connectivity.

   In a VPLS, we use a VCID (which, when using the PWid FEC, has been
   substituted with a more general identifier (AGI), to address
   extending the scope of a VPLS) to identify an emulated LAN segment.
   Note that the VCID as specified in [RFC4447] is a service identifier,
   identifying a service emulating a point-to-point virtual circuit. In
   a VPLS, the VCID is a single service identifier, so it has global
   significance across all PEs involved in the VPLS instance.

Authors' Addresses

   Marc Lasserre
   Alcatel-Lucent
   EMail: mlasserre@alcatel-lucent.com

   Vach Kompella
   Alcatel-Lucent
   EMail: vach.kompella@alcatel-lucent.com

Full Copyright Statement

Intellectual Property

Acknowledgement