Internet Draft Document   Marc Lasserre
Provider Provisioned VPN Working Group   Vach Kompella
draft-ietf-l2vpn-vpls-ldp-03.txt   (Editors)
Expires: October 2004   April 2004

               Virtual Private LAN Services over MPLS
                   draft-ietf-l2vpn-vpls-ldp-03.txt

1.   Status of this Memo

This document is an Internet-Draft and is in full conformance
with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering
Task Force (IETF), its areas, and its working groups. Note that
other groups may also distribute working documents as Internet-
Drafts.

Internet-Drafts are draft documents valid for a maximum of six
months and may be updated, replaced, or obsoleted by other documents
at any time. It is inappropriate to use Internet-Drafts as
reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
     http://www.ietf.org/ietf/1id-abstracts.txt
The list of Internet-Draft Shadow Directories can be accessed at
     http://www.ietf.org/shadow.html.

2.   Abstract

This document describes a virtual private LAN service (VPLS)
solution using pseudo-wires, 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. It delivers a layer 2 broadcast domain that is fully capable
of learning and forwarding on Ethernet MAC addresses that is closed
to a given set of users. Multiple VPLS services can be supported
from a single PE node.

This document describes the control plane functions of signaling
demultiplexor labels, extending [PWE3-CTRL]. 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 [PWE3-
ETHERNET].

3.   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

RELATED DOCUMENTS

www.ietf.org/internet-drafts/draft-ietf-ppvpn-l2vpn-requirements-
01.txt
www.ietf.org/internet-drafts/draft-ietf-ppvpn-l2-framework-03.txt
www.ietf.org/internet-drafts/draft-ietf-pwe3-ethernet-encap-02.txt
www.ietf.org/internet-drafts/draft-ietf-pwe3-control-protocol-01.txt

Table of Contents

4.   Overview

Ethernet has become the predominant technology for Local Area
Networks (LANs) 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 MAN/WAN network(s), 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 LSP basis, packet
replication across LSPs for multicast/broadcast traffic and for
flooding of unknown unicast destination traffic.

[PWE3-ETHERNET] defines how to carry L2 PDUs over point-to-point
MPLS LSPs, called pseudowires (PW). Such PWs can be carried over
MPLS or GRE tunnels. This document describes extensions to [PWE3-
CTRL] 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
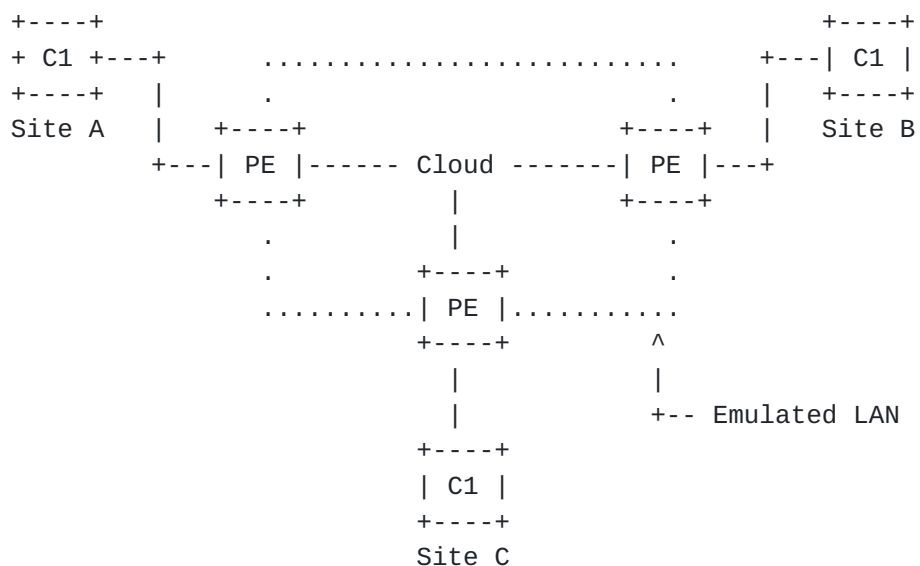external servers, as discussed in [L2VPN-REQ].

The following discussion applies to devices that are VPLS capable
and have a means of tunneling labeled packets amongst each other.
While MPLS LSPs may be used to tunnel these labeled packets, other
technologies may be used as well, e.g., GRE [MPLS-GRE]. The
resulting set of interconnected devices forms a private MPLS VPN.

5.   Topological Model for VPLS

An interface participating in a VPLS must be able to flood, forward,
and filter Ethernet frames.

```
+----+                                          +----+
+ C1 +---+      ...........................      +---| C1 |
+----+   |        .                      .       |   +----+
Site A   |    +----+                    +----+    |   Site B
         +---| PE |------ Cloud -------| PE |---+
             +----+          |          +----+
              .              |            .
              .           +----+          .
              ..........| PE |...........
                         +----+         ^
                          |             |
                          |             +-- Emulated LAN
                         +----+
                         | C1 |
                         +----+
                         Site C
```

The set of PE devices interconnected via pseudowires appears as a
single emulated LAN to customer C1. Each PE device will learn remote
MAC address to pseudowire associations and will also learn directly
attached MAC addresses on customer facing ports.

We note here again that while this document shows specific examples
using MPLS transport tunnels, other tunnels that can be used by
pseudo-wires, e.g., GRE, L2TP, IPSEC, etc., 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].

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

5.1.   Flooding and Forwarding

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

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 can be used. The use of
IGMP snooping and PIM snooping techniques should be used to improve
multicast efficiency.

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

5.2.   Address Learning

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

A pseudowire consists of a pair of uni-directional VC LSPs. The
state of this pseudowire is considered operationally up when both
incoming and outgoing VC LSPs are established. Similarly, it is
considered operationally down when one of these two VC LSPs is torn
down. When a previously unknown MAC address is learned on an
inbound VC LSP, it needs to be associated with the its counterpart
outbound VC LSP in that pseudowire.

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

5.3.   Tunnel Topology

PE routers are assumed to have the capability to establish transport
tunnels. Tunnels are set up between PEs to aggregate traffic.
Pseudowires are signaled to demultiplex the L2 encapsulated packets
that traverse the 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
tunnels and pseudowires.

5.4.   Loop free L2 VPN

For simplicity, a full mesh of pseudowires is established between
PEs. Ethernet bridges, unlike Frame Relay or ATM where the
termination point becomes the CE node, have to examine the layer 2
fields of the packets to make a switching decision. If the frame is
directed to an unknown destination, or is a broadcast or multicast
frame, the frame must be flooded.

Therefore, if the topology isn't a full mesh, the PE devices may
need to forward these frames to other PEs. However, this would
require the use of spanning tree protocol to form a loop free
topology that may have characteristics that are undesirable to the
provider. The use of a full mesh and split-horizon forwarding
obviates the need for a spanning tree protocol.

Each PE MUST create a rooted tree to every other PE router that
serves the same VPLS. Each PE MUST support a "split-horizon" scheme
in order to prevent loops, that is, a PE MUST NOT forward traffic
from one pseudowire to another in the same VPLS mesh (since each PE
has direct connectivity to all other PEs in the same VPLS).

Note that customers are allowed to run STP 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 BPDUs are simply
tunneled through the provider cloud.

6.   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], [LDP-DISC]).

7.   Control Plane

This document describes the control plane functions of Demultiplexor
Exchange (signaling of VC 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 are not described here.

7.1.   LDP Based Signaling of Demultiplexors

In order to establish a full mesh of pseudowires, all PEs in a VPLS
must have a full mesh of LDP sessions.

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

In [PWE3-CTRL], two types of FECs are described, the FEC type 128
PWid FEC Element and the FEC type 129 Generalized PWid FEC Element.
The original FEC element used for VPLS was compatible with the PWid
FEC Element. The text for signaling using PWid FEC Element has been
moved to Appendix 1. What we describe below replaces that with a
more generalized L2VPN descriptor through the Generalized PWid FEC
Element.

7.1.1.   Using the Generalized PWid FEC Element

[PWE3-CTRL] describes a generalized FEC structure that is be used
for VPLS signaling in the following manner. The following describes
the assignment of the Generalized PWid FEC Element fields in the
context of VPLS signaling.

Control bit (C): Depending on whether, on that particular
pseudowire, the control word is desired or not, the control bit may
be specified.

PW type: The allowed PW types in this version are Ethernet and
Ethernet VLAN.

VC info length: Same as in [PWE3-CTRL].

AGI, Length, Value: The unique name of this VPLS. The AGI
identifies a type of name, the length denotes the length of Value,
which is the name of the VPLS. We will use the term AGI
interchangeably with VPLS identifier.

TAII, SAII: These are null because the mesh of PWs in a VPLS
terminate on MAC learning tables, rather than on individual
attachment circuits.

Interface Parameters: The relevant interface parameters are:
     MTU: the MTU of the VPLS MUST be the same across all the PWs in
          the mesh.
     Optional Description String: same as [PWE3-CTRL].
     Requested VLAN ID: If the PW type is Ethernet VLAN, this
          parameter may be used to signal the insertion of the
          appropriate VLAN ID.

7.1.2.   Address Withdraw Message Containing MAC TLV

When MAC addresses are being removed or relearned explicitly, e.g.,
the primary link of a dual-homed MTU-s has failed, an Address
Withdraw Message with the list of MAC addresses to be relearned can

be sent to all other PEs over the corresponding directed LDP
sessions.

The processing for MAC TLVs received in an Address Withdraw Message
is:
   For each MAC address in the TLV:
   - Relearn the association between the MAC address and the
      interface/pseudowire over which this message is received

   For an 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 this link (over the pseudowire associated with the
      signaling link over which the message is received)

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

7.2.   MAC Address Withdrawal

It MAY be desirable to remove or relearn MAC addresses that have
been dynamically learned for faster convergence.

We introduce an optional MAC TLV that is used to specify a list of
MAC addresses that can be removed or relearned using the Address
Withdraw Message.

The Address Withdraw message with MAC 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 MTU-s).
If a notification message is sent on the backup link (blocked link),
which has transitioned into an active state (e.g., similar to
Topology Change Notification message of 802.1w RSTP), with a list of
MAC entries to be relearned, the PE will update the MAC entries in
its FIB for that VPLS instance and send the message to other PEs
over the corresponding directed LDP sessions.

If the notification message contains an empty list, this tells the
receiving PE to remove all the MAC addresses learned for the
specified VPLS instance except the ones it learned from the sending
PE (MAC address removal is required for all VPLS instances that are
affected). Note that the definition of such a notification message
is outside the scope of the document, unless it happens to come from
an MTU connected to the PE as a spoke. In such a scenario, the
message will be just an Address Withdraw message as noted above.

7.2.1.   MAC TLV

MAC addresses to be relearned can be signaled using an LDP Address

Withdraw Message that contains a new TLV, the MAC TLV. Its format

is described below. The encoding of a MAC TLV address is the 6-byte
MAC address specified by IEEE 802 documents [g-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 #n                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

U bit
     Unknown bit. This bit MUST be set to 0. 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 (subject to IANA
approval). This identifies the TLV type as MAC TLV.

Length
     Length field. This field specifies the total length of the MAC
addresses in the TLV.

MAC Address
     The MAC address(es) being removed.

The LDP Address Withdraw Message contains a FEC TLV (to identify the
VPLS in consideration), a MAC Address TLV and optional parameters.
No optional parameters have been defined for the MAC Address
Withdraw signaling.

8.   Data Forwarding on an Ethernet VC Pseudowire

This section describes the dataplane behavior on an Ethernet
pseudowire used in a VPLS. While the encapsulation is similar to
that described in [PWE3-ETHERNET], the NSP functions of stripping
the service-delimiting tag and using a "normalized" Ethernet packet
are described.

8.1.   VPLS Encapsulation actions

In a VPLS, a customer Ethernet packet without preamble is

encapsulated with a header as defined in [PWE3-ETHERNET]. A
customer Ethernet packet is defined as follows:

   - If the packet, 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 is stripped before the packet
     is sent into the VPLS. As the packet exits the VPLS, the
     packet may have a service-delimiting encapsulation inserted.

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

As an application of these rules, a customer packet 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 packet may be tagged again, if a
service-delimiting tag is used, or it may be untagged if none is
used.

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

Contrariwise, if a customer packet 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 packet that traverses a
VPLS is always a customer Ethernet packet. 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 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 pseudowire given by [PWE3-
ETHERNET] can serve as the access side connection into a PE. An
RFC1483 PVC encapsulation could be another 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 VPLS deployments, as described below.

8.1.1.   VPLS Learning actions

Learning is done based on the customer Ethernet packet, as defined

above. The Forwarding Information Base (FIB) keeps track of the
mapping of customer Ethernet packet addressing and the appropriate

pseudowire to use. We define two modes of learning: qualified and
unqualified learning.

In unqualified learning, all the customer VLANs 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 UNI interface 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.

For STP to work in qualified 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
Martini) can be added by MTU-s nodes such that PEs can unambiguously
identify all customer traffic, including STP/MSTP 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/MSTP
traffic. The STP/MSTP 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.

9.   Data Forwarding on an Ethernet VLAN Pseudowire

This section describes the dataplane behavior on an Ethernet VLAN
pseudowire in a VPLS. While the encapsulation is similar to that
described in [PWE3-ETHERNET], the NSP functions of imposing tags,
and using a "normalized" Ethernet packet are described. The
learning behavior is the same as for Ethernet pseudowires.

9.1.   VPLS Encapsulation actions

In a VPLS, a customer Ethernet packet without preamble is

encapsulated with a header as defined in [PWE3-ETHERNET]. A
customer Ethernet packet is defined as follows:

   - If the packet, 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 packet is sent into the VPLS,
     unless the Requested VLAN ID optional parameter was signaled.
     In that case, the VLAN tag is overwritten before the packet is
     sent out on the pseudowire.

   - If the packet, 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 packet 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 pseudowire is a simple way to preserve customer
802.1p bits.

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

10.   Operation of a VPLS

We show here 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.

Initially, the VPLS is set up so that PE1, PE2 and PE3 have a full-
mesh of Ethernet pseudowires. The VPLS instance is assigned a
unique VCID.

For the above example, say PE1 signals VC Label 102 to PE2 and 103
to PE3, and PE2 signals VC Label 201 to PE1 and 203 to PE3.

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 multicast the packet to PE2 and
PE3. When PE2 receives the packet, it will have an inner 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 VC Label 102.

```
                                             -----
                                            /  A1 \
         ----                         ----CE1     |
        /     \       --------      -------  /     |     |
       | A2 CE2-      /        \    /        PE1     \     /
        \    /   \   /          \---/          \      -----
         ----      ---PE2                        |
                    | Service Provider Network |
                    \          /   \           /
            -----  PE3        /     \         /
            |Agg|_/  --------       -------
           -|    |
      ----  /  -----   ----
     /    \/    \   /    \              CE = Customer Edge Router
    | A3 CE3    --C4 A4 |              PE = Provider Edge Router
     \    /         \    /             Agg = Layer 2 Aggregation
      ----            ----
```
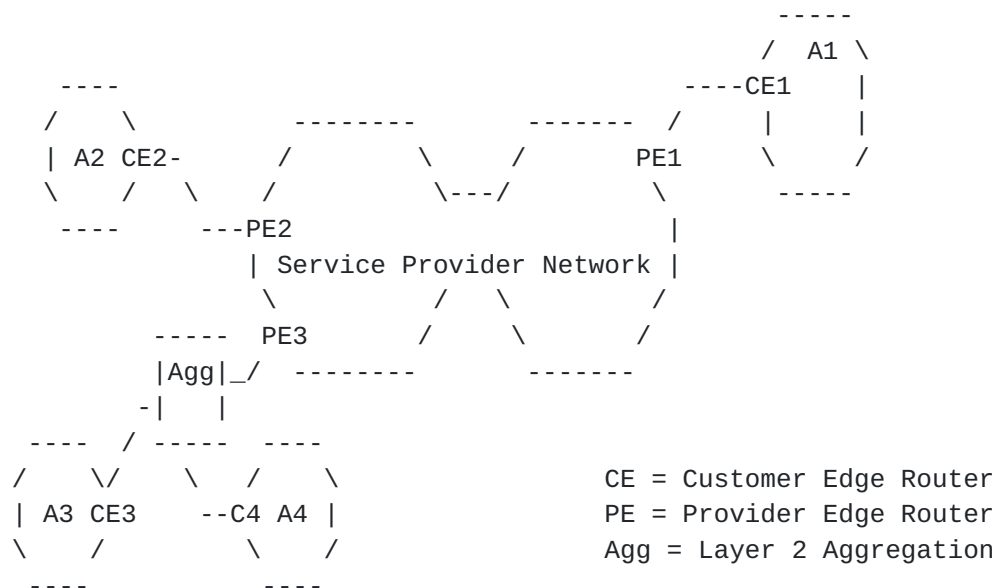
10.1.   MAC Address Aging

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

As packets arrive, MAC addresses are remembered. The aging timer
for MAC address M SHOULD be reset when a packet is received with
source MAC address M.

11.   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 pseudowires must be setup between
the PE routers. While this creates signaling overhead, the real
detriment to large scale deployment is the packet replication
requirements for each provisioned VCs 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 device 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 PE VPLS access points.

It is often beneficial to extend the VPLS service tunneling
techniques into the MTU (multi-tenant unit) domain. This can be
accomplished by treating the MTU device as a PE device and

provisioning pseudowires between it and every other edge, as an
basic VPLS. An alternative is to utilize [PWE3-ETHERNET]
pseudowires or Q-in-Q logical interfaces between the MTU 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
pseudowires (Hub) are augmented with access pseudowires (Spoke) to
form a two-tier hierarchical VPLS (H-VPLS).

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

11.1.   Hierarchical connectivity

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

For rest of this discussion we will refer to a bridging capable MTU
device as MTU-s and a non-bridging capable PE device as PE-r. A
routing and bridging capable device will be referred to as PE-rs.

11.1.1.   Spoke connectivity for bridging-capable devices

As shown in the figure below, consider the case where an MTU-s
device has a single connection to the PE-rs device placed in the CO.
The PE-rs devices are connected in a basic VPLS full mesh. For each
VPLS service, a single spoke pseudowire is set up between the MTU-s
and the PE-rs based on [PWE3-CTRL]. Unlike traditional pseudowires
that terminate on a physical (or a VLAN-tagged logical) port at each
end, the spoke pseudowire terminates on a virtual bridge instance on
the MTU-s and the PE-rs devices.
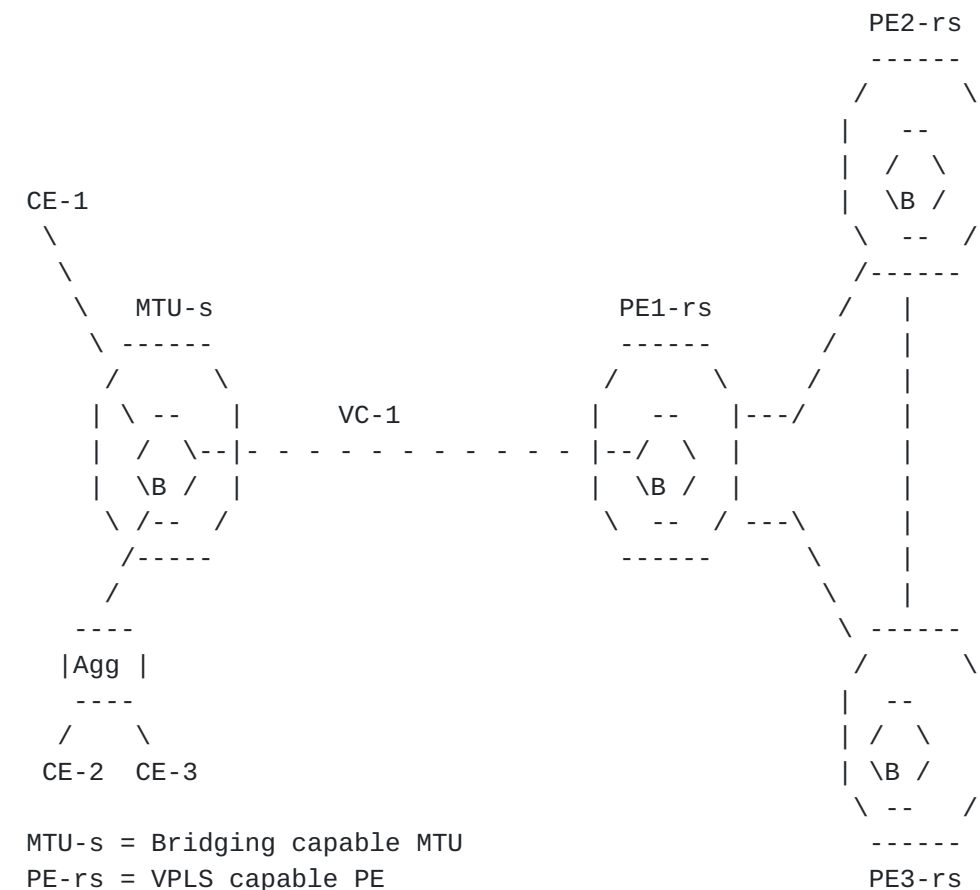
```
                                                      PE2-rs
                                                      ------
                                                     /      \
                                                    |   --   |
                                                    |  / \   |
    CE-1                                            |  \B /  |
      \                                              \  --  /
       \                                             /------
        \    MTU-s                     PE1-rs       /   |
         \ ------                      ------      /    |
         /      \                     /      \    /     |
        | \ --   |      VC-1         |   --   |---/      |
        |  /  \--|- - - - - - - - - -|--/  \  |         |
        |  \B / |                    |  \B / |          |
         \ /--  /                     \  --  / ---\      |
          /-----                       ------      \     |
         /                                          \    |
       ----                                          \ ------
      |Agg |                                         /      \
       ----                                         |   --   |
      /    \                                        | / \    |
   CE-2   CE-3                                      | \B /   |
                                                     \ --   /
    MTU-s = Bridging capable MTU                      ------
    PE-rs = VPLS capable PE                           PE3-rs


    --
   /  \
   \B / = Virtual VPLS(Bridge)Instance
    --
    Agg = Layer-2 Aggregation
```

The MTU-s device and the PE-rs device treat each spoke connection
like an access port of the VPLS service. On access ports, the
combination of the physical port and/or the VLAN tag is used to
associate the traffic to a VPLS instance while the pseudowire tag
(e.g., VC label) is used to associate the traffic from the virtual
spoke port with a VPLS instance, followed by a standard L2 lookup to
identify which customer port the frame needs to be sent to.

11.1.1.1.   MTU-s Operation

MTU-s device 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 virtual spoke port.
Packets to unknown destination are replicated to all ports in the
service including the virtual spoke port. Once the MAC address is
learned, traffic between CE1 and CE2 will be switched locally by the

MTU-s device saving the link capacity of the connection to the PE-
rs. Similarly traffic between CE1 or CE2 and any remote destination
is switched directly on to the spoke connection and sent to the PE-
rs over the point-to-point pseudowire.

Since the MTU-s is bridging capable, only a single pseudowire 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 connection
pseudowire.

11.1.1.2.   PE-rs Operation

The PE-rs device 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 pseudowire. Thus, the spoke pseudowire from
the PE-r is treated as a virtual port and the PE-rs device will
switch traffic between the spoke pseudowire, hub pseudowires, and
access ports once it has learned the MAC addresses.

11.1.2.   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
   pseudowires per service between all devices participating in the
   VPLS service.
- Minimizes signaling overhead since fewer pseudowires 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 device and all of it s
   locally connected MTU-s and PE-r.
- Addition of other sites requires configuration of the new MTU-s
   device 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.
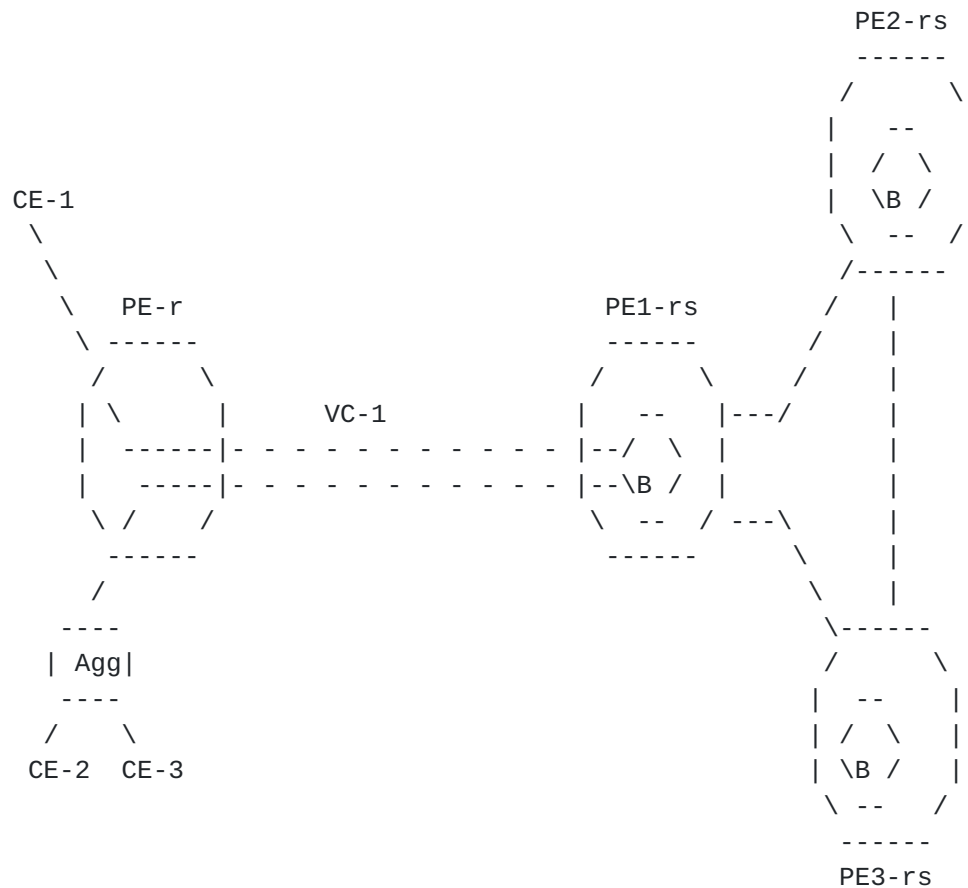
11.1.3.   Spoke connectivity for non-bridging devices

In some cases, a bridging PE-rs device may not be deployed in a CO

or a multi-tenant building while a PE-r might already be deployed.
If there is a need to provide VPLS service from the CO where the PE-
rs device is not available, the service provider may prefer to use

the PE-r device in the interim. In this section, we explain how a
PE-r device that does not support any of the VPLS bridging
functionality can participate in the VPLS service.

As shown in this figure, the PE-r device creates a point-to-point
tunnel LSP to a PE-rs device. Then for every access port that needs

```
                                               PE2-rs
                                               ------
                                              /      \
                                             |   --   |
                                             |  / \   |
  CE-1                                       |  \B /  |
    \                                         \  --  /
     \                                         /------
      \    PE-r                 PE1-rs        /   |
       \ ------                 ------      /     |
        /      \               /      \   /       |
       | \      |    VC-1      |   --   |---/      |
       |  ------|- - - - - - - - - - - |--/  \   |        |
       |   -----|- - - - - - - - - - - |--\B /   |        |
        \ /    /                \  --  / ---\     |
         ------                 ------       \    |
        /                               \    |
     ----                                \------
    | Agg|                               /    \
     ----                               |  --   |
    /    \                              | / \   |
  CE-2  CE-3                            | \B /  |
                                        \  --  /
                                         ------
                                         PE3-rs
```

to participate in a VPLS service, the PE-r device creates a point-
to-point [PWE3-ETHERNET] pseudowire that terminates on the physical
port at the PE-r and terminates on the virtual bridge instance of
the VPLS service at the PE-rs.

11.1.3.1.   PE-r Operation

The PE-r device is defined as a device that supports routing but
does not support any bridging functions. However, it is capable of
setting up [PWE3-ETHERNET] pseudowires between itself and the PE-rs.
For every port that is supported in the VPLS service, a [PWE3-
ETHERNET] pseudowire is setup from the PE-r to the PE-rs. Once the
pseudowires are setup, there is no learning or replication function

required on part of the PE-r. All traffic received on any of the
access ports is transmitted on the pseudowire. Similarly all
traffic received on a pseudowire is transmitted to the access port

where the pseudowire terminates. Thus traffic from CE1 destined for
CE2 is switched at PE-rs and not at PE-r.

This approach adds more overhead than the bridging capable (MTU-s)
spoke approach since a pseudowire is required for every access port
that participates in the service versus a single pseudowire required
per service (regardless of access ports) when a MTU-s type device 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 device.

11.2.   Redundant Spoke Connections

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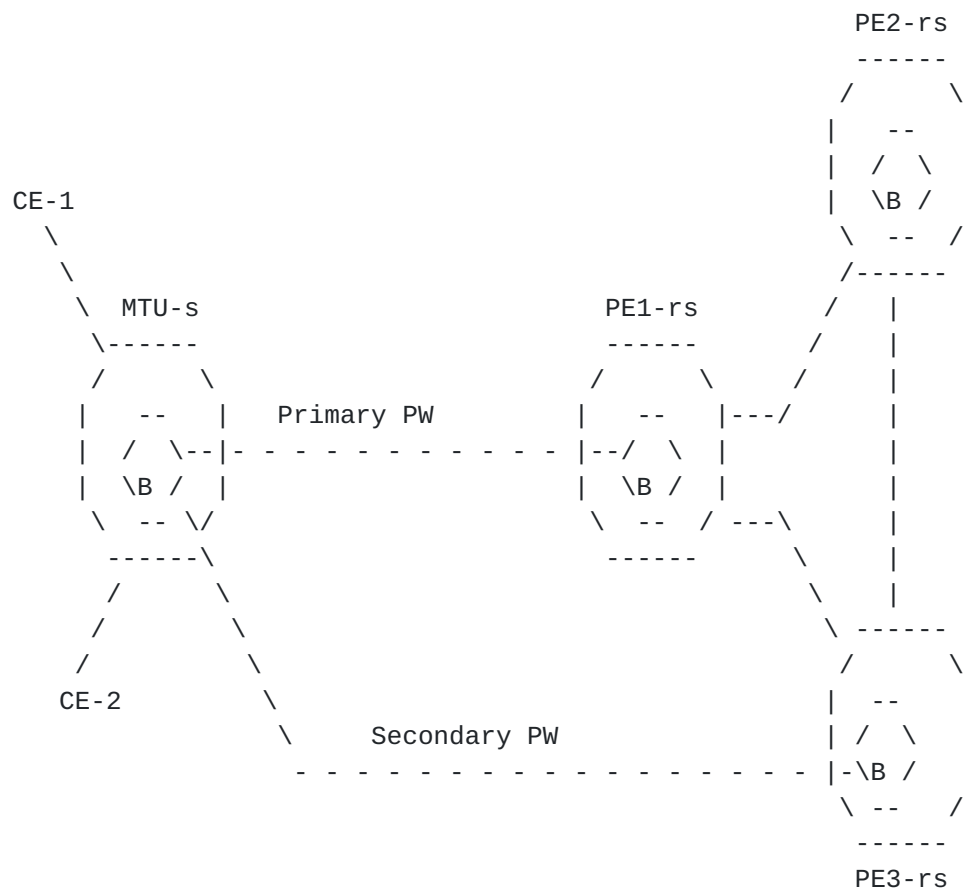device suffers total loss of connectivity.

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

11.2.1.   Dual-homed MTU device

To protect from connection failure of the pseudowire or the failure
of the PE-rs device, the MTU-s device or the PE-r is dual-homed into
two PE-rs devices, as shown in figure-3. The PE-rs devices must be
part of the same VPLS service instance.

An MTU-s device will setup two [PWE3-ETHERNET] pseudowires (one each
to PE-rs1 and PE-rs2) for each VPLS instance. One of the two
pseudowires is designated as primary and is the one that is actively
used under normal conditions, while the second pseudowire is
designated as secondary and is held in a standby state. The MTU
device negotiates the pseudowire labels for both the primary and
secondary pseudowires, but does not use the secondary pseudowire
unless the primary pseudowire fails. Since only one link is active
at a given time, a loop does not exist and hence 802.1D spanning
tree is not required.

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

11.2.2.   Failure detection and recovery

The MTU-s device controls the usage of the pseudowires to the PE-rs
nodes. Since LDP signaling is used to negotiate the pseudowire
labels, the hello messages used for the LDP session can be used to
detect failure of the primary pseudowire.

Upon failure of the primary pseudowire, MTU-s device immediately
switches to the secondary pseudowire. At this point the PE3-rs
device that terminates the secondary pseudowire starts learning MAC
addresses on the spoke pseudowire. 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 relearning 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 device where the
secondary pseudowire got activated may send out a flush message,
using the MAC 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.

11.3.   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 pseudowire 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 5.

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.

12.   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 are 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 device needs to have the capability of
adding the additional P-VLAN tag for non-multiplexed customer UNI
port where customer VLANs are not used as service delimiter. If
customer VLANs need to be treated as service delimiter (e.g.,
customer UNI port 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 in order to resolve overlapping VLAN-ids used by

different customers. Therefore, the MTU-s device in this model can be
considered as a typical bridge with this additional UNI capability.

The PE-rs device needs to be able to perform bridging functionality over the standard Ethernet ports toward the access network as well as over the pseudowires toward the network core. The set of pseudowires that corresponds to a VPLS instance would look just like a P-VLAN to the bridge portion of the PE-rs and that is why sometimes it is referred to as Emulated VLAN. In this model the PE-rs may need to run STP protocol in addition to split-horizon. Split horizon is run over MPLS-core; whereas, STP is run over the access network to accommodate any arbitrary access topology. In this model, the PE-rs needs to map a P-VLAN to a VPLS-instance and its associated pseudowires and vise versa.

The details regarding bridge operation for MTU-s and PE-rs (e.g., encapsulation format for QinQ messages, customer s Ethernet control protocol handling, etc.) are outside of the scope of this document and they are covered in [802.1ad]. However, the relevant part is the interaction between the bridge module and the MPLS/IP pseudowires in the PE-rs device.

12.1.   Scalability

Given that each P-VLAN corresponds to a VPLS instance, one may think that the total number of VPLS instances supported is limited to 4K. However, the 4K limit applies only to each Ethernet access network (Ethernet island) and not to the entire network. The SP network, in this model, 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). Each island may consist of many MTU-s devices, several aggregators, and one or more PE-rs devices. The PE-rs devices enable a P-VLAN to be extended from one island to others using a set of pseudowires (associated with that VPLS instance) and providing a loop free mechanism across the core network through split-horizon. Since a P-VLAN serves as a service delimiter within the provider's network, it does not get carried over the pseudowires and furthermore the mapping between P-VLAN and the pseudowires is a local matter. This means a VPLS instance can be represented by different P-VLAN in different Ethernet islands and furthermore each island can support 4K VPLS instances independent from one another.

12.2.   Dual Homing and Failure Recovery

In this model, an MTU-s can be dual or triple homed to different devices (aggregators and/or PE-rs devices). The failure protection for access network nodes and links can be provided through running MSTP in each island. The MSTP of each island is independent from

other islands and do not interact with each other. If an island has
more than one PE-rs, then a dedicated full-mesh of pseudowires is
used among these PE-rs devices for carrying the SP BPDU packets for
that island. On a per P-VLAN basis, the MSTP 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.

13.   Significant Modifications

Between rev 02 and this one, these are the changes:
    o Introduction of the Generalized PWid FEC in the signaling of
       a VPLS
    o Description of the use of Ethernet VLAN pseudowires

14.   Contributors

Loa Andersson, TLA
Ron Haberman, Masergy
Juha Heinanen, Independent
Giles Heron, Tellabs
Sunil Khandekar, Alcatel
Luca Martini, Cisco
Pascal Menezes, Terabeam
Rob Nath, Riverstone
Eric Puetz, SBC
Vasile Radoaca, Nortel
Ali Sajassi, Cisco
Yetik Serbest, SBC
Nick Slabakov, Riverstone
Andrew Smith, Consultant
Tom Soon, SBC
Nick Tingle, Alcatel

15.   Acknowledgments

We wish to thank Joe Regan, Kireeti Kompella, Anoop Ghanwani, Joel
Halpern, Rick Wilder, Jim Guichard, Steve Phillips, Norm Finn, Matt
Squire, Muneyoshi Suzuki, Waldemar Augustyn, Eric Rosen, Yakov
Rekhter, and Sasha Vainshtein for their valuable feedback. In
addition, we would like to thank Rajiv Papneja (ISOCORE), Winston
Liu (ISOCORE), and Charlie Hundall (Extreme) for identifying issues
with the draft in the course of the interoperability tests.

16.   Security Considerations

A more comprehensive description of the security issues involved in
L2VPNs is covered in [VPN-SEC]. An unguarded VPLS service is
vulnerable to some security issues which 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

        o Traffic isolation between VPLS domains is guaranteed by
           the use of per VPLS L2 FIB table and the use of per VPLS
           pseudowires
        o 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
        o 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
        o LDP security (authentication) methods as described in
           [[RFC-3036](RFC-3036)] SHOULD be applied. This would prevent
           unauthorized participation by a PE in a VPLS.
     . Denial of service attacks
        o Some means to limit the number of MAC addresses (per site
           per VPLS) that a PE can learn SHOULD be implemented.

17.   Intellectual Property Considerations

This document is being submitted for use in IETF standards
discussions.

18.   Full Copyright Statement

   Copyright (C) The Internet Society (2001). All Rights Reserved.
This document and translations of it may be copied and furnished to
others, and derivative works that comment on or otherwise explain it
or assist in its implementation may be prepared, copied, published
and distributed, in whole or in part, without restriction of any
kind, provided that the above copyright notice and this paragraph
are included on all such copies and derivative works. However, this
document itself may not be modified in any way, such as by removing
the copyright notice or references to the Internet Society or other
Internet organizations, except as needed for the purpose of
developing Internet standards in which case the procedures for
copyrights defined in the Internet Standards process must be
followed, or as required to translate it into languages other than
English.

The limited permissions granted above are perpetual and will not be
revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an
"AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING
TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING
BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION
HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF

MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

19.   References

[PWE3-ETHERNET] "Encapsulation Methods for Transport of Ethernet
Frames Over IP/MPLS Networks", draft-ietf-pwe3-ethernet-encap-
06.txt, Work in progress, April 2004.

[PWE3-CTRL] "Transport of Layer 2 Frames over MPLS", draft-ietf-
pwe3-control-protocol-06.txt, Work in progress, March 2004.

[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.

[BGP-VPN] "BGP/MPLS VPNs". draft-ietf-l3vpn-rfc2547bis-01.txt, Work
in Progress, September 2003.

[RFC3036] "LDP Specification", L. Andersson, et al. RFC 3036.
January 2001.

[RADIUS-DISC] "Using Radius for PE-Based VPN Discovery", draft-ietf-
l2vpn-radius-pe-discovery-00.txt, Work in Progress, February 2004.

[BGP-DISC] "Using BGP as an Auto-Discovery Mechanism for Network-
based VPNs", draft-ietf-l3vpn-bgpvpn-auto-02.txt, Work in Progress,
April 2004.

[LDP-DISC] "Discovering Nodes and Services in a VPLS Network",
draft-stokes-ppvpn-vpls-discover-00.txt, Work in Progress, June
2002.

[L2FRAME] "Framework for Layer 2 Virtual Private Networks (L2VPNs)",
draft-ietf-l2vpn-l2-framework-04, Work in Progress, March 2004.

[L2VPN-REQ] "Service Requirements for Layer-2 Provider Provisioned
Virtual Private Networks", draft-ietf-l2vpn-requirements-01.txt,
Work in Progress, February 2004.

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

[VPN-SEC] "Security Framework for Provider Provisioned Virtual
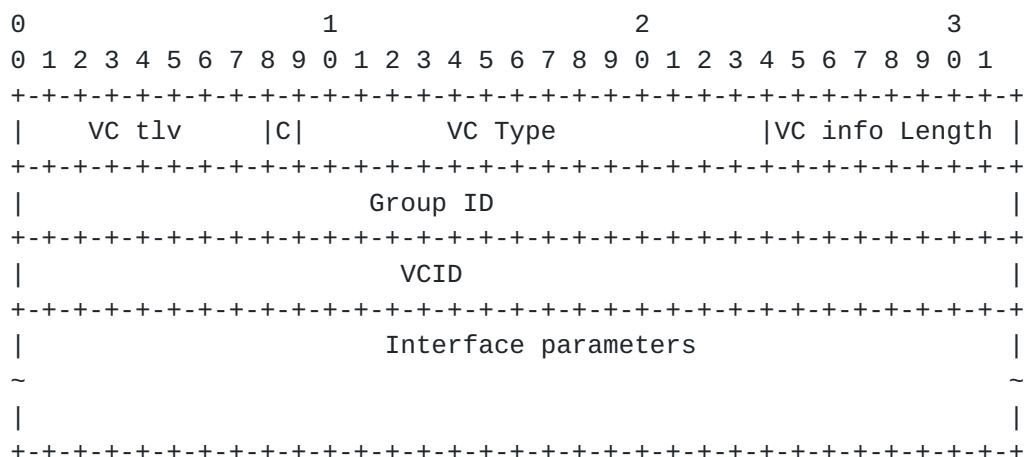
Private Networks", draft-ietf-l3vpn-security-framework-01.txt, Work in
Progress, February 2004.

Appendix 1. Signaling a VPLS 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 VC
FEC TLV.

VC, C, VC Info Length, Group ID, Interface parameters are as defined
in [PWE3-CTRL].

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    VC tlv     |C|         VC Type           |VC info Length |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Group ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          VCID                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Interface parameters                      |
~                                                             ~
|                                                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

We use the Ethernet pseudowire type to identify pseudowires that
carry Ethernet traffic for multipoint connectivity.

In a VPLS, we use a VCID (which has been substituted with a more
general identifier, to address extending the scope of a VPLS) to
identify an emulated LAN segment. Note that the VCID as specified
in [PWE3-CTRL] is a service identifier, identifying a service
emulating a point-to-point virtual circuit. In a VPLS, the VCID is
a single service identifier.

20.   Authors' Addresses

Marc Lasserre
Riverstone Networks
Email: marc@riverstonenet.com

Vach Kompella
Alcatel
Email: vach.kompella@alcatel.com