DECLARATION

I, Laura Nugent, based on my personal knowledge and information, hereby declare as follows:

1. I am Director of Administration and Events of the IETF Administration LLC ("IETF"). IETF is the acronym for the Internet Engineering Task Force, which is an activity of the Internet Society.

2. One of my responsibilities with IETF is to act as the custodian of Internet-Drafts and records relating to Internet-Drafts. I am familiar with the record keeping practices relating to Internet-Drafts, including the creation and maintenance of such records.

3. I hereby declare that all statements made herein are of my own knowledge and information contained in the business records of IETF and are true, and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements may be punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

4. Since 1998, it has been the regular practice of the IETF to publish Internet-Drafts and make them available to the public on its website at www.ietf.org (the IETF website). The IETF maintains copies of Internet-Drafts in the ordinary course of its regularly conducted activities.

5. Any Internet-Draft published on the IETF website was reasonably accessible to the public and was disseminated or otherwise available to the extent that persons interested and ordinarily skilled in the subject matter or art exercising reasonable diligence could have located it. In particular, the Internet-Drafts were indexed and searchable on the IETF website.

6. Internet-Drafts are posted to an IETF online directory. When an Internet-Draft is published, an announcement of its publication that describes the Internet-Draft is disseminated. Typically, that dated announcement is made within 24 hours of the publication of the Internet-Draft. The announcement is kept in the IETF email archive, and the date is affixed automatically.

7. The records of posting the Internet-Drafts in the IETF online repository are kept in the course of the IETF's regularly conducted activity and ordinary course of business. The records are made pursuant to established procedures and are relied upon by the IETF in the performance of its functions.

8. It is the regular practice of the IETF to make and keep the records in the online repository.

9. Exhibit 1 is a true and correct copy of draft-ietf-l2vpn-vpls-bgp-02, titled "Virtual Private LAN Service." I have determined that an announcement of the publication of this Internet-Draft was made May 18, 2004. Therefore, based on the normal practice of the IETF, that Internet-Draft was reasonably available to the public within 24 hours of that announcement. At that time, the Internet-Draft would have been disseminated or otherwise available to the extent that persons interested and ordinarily skilled in the subject matter or art, exercising reasonable diligence, could have located it.

10. Exhibit 2 is a true and correct copy of draft-ietf-l2vpn-vpls-bgp-08, titled "Virtual Private LAN Service (VPLS) Using BGP for Auto-discovery and Signaling." I have determined that an announcement of the publication of this Internet-Draft was made June 22, 2006. Therefore, based on the normal practice of the IETF, that Internet-Draft was reasonably available to the public within 24 hours of that announcement. At that time, the Internet-Draft would have

been disseminated or otherwise available to the extent that persons interested and ordinarily skilled in the subject matter or art, exercising reasonable diligence, could have located it.

11. Exhibit 3 is a true and correct copy of draft-ietf-l2vpn-vpls-ldp-03, titled “Virtual Private LAN Services over MPLS.” I have determined that an announcement of the publication of this Internet-Draft was made April 26, 2004. Therefore, based on the normal practice of the IETF, that Internet-Draft was reasonably available to the public within 24 hours of that announcement. At that time, the Internet-Draft would have been disseminated or otherwise available to the extent that persons interested and ordinarily skilled in the subject matter or art, exercising reasonable diligence, could have located it.

12. Exhibit 4 is a true and correct copy of draft-ietf-l2vpn-vpls-ldp-07, titled “Virtual Private LAN Services over MPLS.” I have determined that an announcement of the publication of this Internet-Draft was made July 18, 2005. Therefore, based on the normal practice of the IETF, that Internet-Draft was reasonably available to the public within 24 hours of that announcement. At that time, the Internet-Draft would have been disseminated or otherwise available to the extent that persons interested and ordinarily skilled in the subject matter or art, exercising reasonable diligence, could have located it.

13. Exhibit 5 is a true and correct copy of draft-ietf-l2vpn-vpls-ldp-09, titled “Virtual Private LAN Services Using LDP.” I have determined that an announcement of the publication of this Internet-Draft was made June 5, 2006. Therefore, based on the normal practice of the IETF, that Internet-Draft was reasonably available to the public within 24 hours of that announcement. At that time, the Internet-Draft would have been disseminated or otherwise available to the extent that persons interested and ordinarily skilled in the subject matter or art, exercising reasonable diligence, could have located it.

14. Exhibit 6 is a true and correct copy of draft-martini-ethernet-encap-mpls-01, titled "Encapsulation Methods for Transport of Ethernet Frames Over IP and MPLS Networks." I have determined that an announcement of the publication of this Internet-Draft was made July 5, 2002. Therefore, based on the normal practice of the IETF, that Internet-Draft was reasonably available to the public within 24 hours of that announcement. At that time, the Internet-Draft would have been disseminated or otherwise available to the extent that persons interested and ordinarily skilled in the subject matter or art, exercising reasonable diligence, could have located it.

Pursuant to Section 1746 of Title 28 of United States Code, I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct and that the foregoing is based upon personal knowledge and information and is believed to be true.

Date: 23 September 2024

By: ______________________
Laura Nugent

Network Working Group
Internet Draft
Category: Standards Track
Expires: November 2004
draft-ietf-l2vpn-vpls-bgp-02.txt

K. Kompella (Editor)
Y. Rekhter (Editor)
Juniper Networks
May 2004

Virtual Private LAN Service

Status of this Memo

Copyright Notice

Abstract

Virtual Private LAN Service (VPLS), also known as Transparent LAN Service, and Virtual Private Switched Network service, is a useful Service Provider offering. The service offered is a Layer 2 Virtual Private Network (VPN); however, in the case of VPLS, the customers in the VPN are connected by a multipoint network, in contrast to the usual Layer 2 VPNs, which are point-to-point in nature.

This document describes the functions required to offer VPLS, and describes a mechanism for signaling a VPLS, as well as for forwarding

VPLS frames across a packet switched network.

Conventions used in this document

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 [1].

1. Introduction

Virtual Private LAN Service (VPLS), also known as Transparent LAN Service, and Virtual Private Switched Network service, is a useful service offering. A Virtual Private LAN appears in (almost) all respects as a LAN to customers of a Service Provider. However, in a VPLS, the customers are not all connected to a single LAN; the customers may be spread across a metro or wide area. In essence, a VPLS glues several individual LANs across a packet-switched network to appear and function as a single LAN [2].

This document describes the functions needed to offer VPLS, and goes on to describe a mechanism for signaling a VPLS, as well as a mechanism for transport of VPLS frames over tunnels across a packet switched network. The signaling mechanism uses BGP as the control plane protocol. This document also briefly discusses deployment options, in particular, the notion of decoupling functions across devices.

Alternative approaches include: [3], which allows one to build a Layer 2 VPN with Ethernet as the interconnect; and [4], which allows one to set up an Ethernet connection across a packet-switched network. Both of these, however, offer point-to-point Ethernet services. What distinguishes VPLS from the above two is that a VPLS offers a multipoint service. A mechanism for setting up pseudowires for VPLS using the Label Distribution Protocol (LDP) is defined in [5].

1.1. Scope of this Document

This document has four major parts: defining a VPLS functional model; defining a control plane for setting up VPLS; defining the data plane for VPLS (encapsulation and forwarding of data); and defining various deployment options.

The functional model underlying VPLS is laid out in section 2. This describes the service being offered, the network components that interact to provide the service, and at a high level their interactions.

The control plane described in this document uses Multiprotocol BGP

[6] to establish VPLS service, i.e., for the autodiscovery of VPLS members and for the setup and teardown of the pseudowires that constitute a given VPLS. Section 3 also describes how a VPLS that spans Autonomous System boundaries is set up, as well as how multi-homing is handled. Using BGP as the control plane for VPNs is not new (see [3], [7] and [8]): what is described here is based on the mechanisms proposed in [7].

The forwarding plane and the actions that a participating PE must take is described in section 4.

In section 5, the notion of 'decoupled' operation is defined, and the interaction of decoupled and non-decoupled PEs is described. Decoupling allows for more flexible deployment of VPLS.

2. Functional Model

This will be described with reference to Figure 1.

Figure 1: Example of a VPLS

```
                                                       -----
                                                      /  A1 \
        ----                                     ____CE1     |
       /    \          --------       --------  /    |       |
      |  A2 CE2-      /        \     /        PE1     \     /
       \    /   \    /          \___/          | \     -----
        ----     ---PE2                        |  \
                    |                          |   \   -----
                    | Service Provider Network |    \ /     \
                    |                          |     CE5  A5 |
                    |            ___           |   /  \     /
```

```
             |----|  \          /   \         PE4_/    -----
             |u-PE|--PE3       /     \       /
             |----|    --------       -------
      ----  /   |    ----
     /    \/    \   /    \               CE = Customer Edge Device
    |  A3 CE3    --CE4 A4 |              PE = Provider Edge Router
     \    /         \    /               u-PE = Layer 2 Aggregation
      ----           ----                A<n> = Customer site n
```

2.1. Terminology

Terminology similar to that in [7] is used, with the addition of "u-PE", a Layer 2 PE device used for Layer 2 aggregation. A u-PE is owned and operated by the Service Provider (as is the PE). PE and u-PE devices are "VPLS-aware", which means that they know that a VPLS service is being offered. We will call these VPLS edge devices,

which could be either a PE or an u-PE, a VE.

In contrast, the CE device (which may be owned and operated by either the SP or the customer) is VPLS-unaware; as far as the CE is concerned, it is connected to the other CEs in the VPLS via a Layer 2 switched network. This means that there should be no changes to a CE device, either to the hardware or the software, in order to offer VPLS.

A CE device may be connected to a PE or a u-PE via Layer 2 switches that are VPLS-unaware. From a VPLS point of view, such Layer 2 switches are invisible, and hence will not be discussed further. Furthermore, a u-PE may be connected to a PE via Layer 2 and Layer 3 devices; this will be discussed further in a later section.

The term "demultiplexor" refers to an identifier in a data packet that identifies both the VPLS to which the packet belongs as well as the ingress PE. In this document, the demultiplexor is an MPLS label.

The term "VPLS" will refer to the service as well as a particular instantiation of the service (i.e., an emulated LAN); it should be clear from the context which usage is intended.

2.2. Assumptions

The Service Provider Network is a packet switched network. The PEs are assumed to be (logically) full-meshed with tunnels over which packets that belong to a service (such as VPLS) are encapsulated and forwarded. These tunnels can be IP tunnels, such as GRE, or MPLS tunnels, established by RSVP-TE or LDP. These tunnels are established independently of the services offered over them; the signaling and establishment of these tunnels are not discussed in this document.

"Flooding" and MAC address "learning" (see section 4) are an integral part of VPLS. However, these activities are private to an SP device, i.e., in the VPLS described below, no SP device requests another SP device to flood packets or learn MAC addresses on its behalf.

All the PEs participating in a VPLS are assumed to be fully meshed, i.e., every (ingress) PE can send a VPLS packet to the egress PE(s) directly, without the need for an intermediate PE (see the section below on "Split Horizon" Flooding). This assumption reduces (but does not eliminate) the need to run Spanning Tree Protocol among the PEs.

2.3. Interactions

VPLS is a successful "LAN Service" if CE devices that belong to VPLS V can interact through the SP network as if they were connected by a LAN. VPLS is "private" if CE devices that belong to different VPLSs cannot interact. VPLS is "virtual" if multiple VPLSs can be offered over a common packet switched network.

PE devices interact to "discover" all the other PEs participating in the same VPLS (i.e., that are attached to CE devices that belong to the same VPLS), and to exchange demultiplexors. These interactions are control-driven, not data-driven.

U-PEs interact with PEs to establish connections with remote PEs or u-PEs in the same VPLS. Again, this interaction is control-driven.

3. Control Plane

There are two primary functions of the VPLS control plane: autodiscovery, and setup and teardown of the pseudowires that constitute the VPLS, often called signaling. The first two subsections describe these functions. The next subsection describes the setting up of pseudowires that span Autonomous Systems. The last subsection details how multi-homing is handled.

3.1. Autodiscovery

Discovery refers to the process of finding all the PEs that participate in a given VPLS. A PE can either be configured with the identities of all the other PEs in a given VPLS, or the PE can use some protocol to discover the other PEs. The latter is called autodiscovery.

The former approach is fairly configuration-intensive, especially since it is required (in this and other VPLS approaches) that the PEs participating in a given VPLS are fully meshed (i.e., every pair of PEs in a given VPLS establish pseudowires to each other). Furthermore, when the topology of a VPLS changes (i.e., a PE is added to, or removed from the VPLS), the VPLS configuration on all PEs in that VPLS must be changed.

In the autodiscovery approach, each PE "discovers" which other PEs are part of a given VPLS by means of some protocol, in this case BGP. This allows each PE's configuration to consist only of the identity of the VPLS that each customer belongs to, not the identity of every other PE in that VPLS. Moreover, when the topology of a VPLS changes, only the affected PE's configuration changes; other PEs

automatically find out about the change and adapt.

3.1.1. Functions

A PE that participates in a given VPLS V must be able to tell all other PEs in VPLS V that it is also a member of V. A PE must also have a means of declaring that it no longer participates in a VPLS. To do both of these, the PE must have a means of identifying a VPLS and a means by which to communicate to all other PEs.

U-PE devices also need to know what constitutes a given VPLS;

however, they don't need the same level of detail. The PE (or PEs) to which a u-PE is connected gives the u-PE an abstraction of the VPLS; this is described in section 5.

3.1.2. Protocol Specification

The specific mechanism for autodiscovery described here is based on [3] and [7]; it uses BGP extended communities [9] to identify members of a VPLS. A more generic autodiscovery mechanism is described in [8]. The specific extended community used is the Route Target, whose format is described in [9]. The semantics of the use of Route Targets is described in [7]; their use in VPLS is identical.

As it has been assumed that VPLSs are fully meshed, a single Route Target RT suffices for a given VPLS V, and in effect that RT is the identifier for VPLS V.

A PE announces (typically via I-BGP) that it belongs to VPLS V by annotating its NLRIs for V (see next subsection) with Route Target RT, and acts on this by accepting NLRIs from other PEs that have Route Target RT. A PE announces that it no longer participates in V by withdrawing all NLRIs that it had advertised with Route Target RT.

3.2. Signaling

Once discovery is done, each pair of PEs in a VPLS must be able to establish (and tear down) pseudowires to each other, i.e., exchange (and withdraw) demultiplexors. This process is known as signaling. Signaling is also used to initiate "relearning", and to transmit certain characteristics of the PE regarding a given VPLS.

Recall that a demultiplexor is used to distinguish among several different streams of traffic carried over a tunnel, each stream possibly representing a different service. In the case of VPLS, the demultiplexor not only says to which specific VPLS a packet belongs, but also identifies the ingress PE. The former information is used for forwarding the packet; the latter information is used for

learning MAC addresses. The demultiplexor described here is an MPLS label, even though the PE-to-PE tunnels may not be MPLS tunnels.

3.2.1. Setup and Teardown

The VPLS BGP NLRI described below, with a new AFI and SAFI (see [6]) is used to exchange demultiplexors.

A PE advertises a VPLS NLRI for each VPLS that it participates in. If the PE is doing learning and flooding, i.e., it is the VE, it announces a single set of VPLS NLRIs for each VPLS that it is in. If the PE is connected to several u-PEs, it announces one set of VPLS NLRIs for each u-PE. A hybrid scheme is also possible, where the PE learns MAC addresses on some interfaces (over which it is directly connected to CEs) and delegates learning on other interfaces (over which it is connected to u-PEs). In this case, the PE would announce one set of VPLS NLRIs for each u-PE that has customer ports in a given VPLS, and one set for itself, if it has customer ports in that VPLS.

Each set of NLRIs defines the demultiplexors for a range of other PEs in the VPLS. Ideally, a single NLRI suffices to cover all PEs in a VPLS; however, there are cases (such as a newly added PE) where the pre-existing NLRI does not have enough labels. In such cases, advertising an additional NLRI for the same VPLS serves to add labels for the new PEs without disrupting service to the pre-existing PEs. If service disruption is acceptable (or when the PE restarts its BGP process), a PE MAY consider coalescing all NLRIs for a VPLS into a single NLRI.

If a PE X is part of VPLS V, and X receives a VPLS NLRI for V from PE Y that includes a demultiplexor that X can use, X sets up its ends of a pair of pseudowires between X and Y. X may also have to advertise a new NLRI for V that includes a demultiplexor that Y can use, if its pre-existing NLRI for V did not include a demultiplexor for Y.

If Y's configuration is changed to remove it from VPLS V, then Y MUST withdraw all its NLRIs for V. If all Y's links to CEs in V go down, then Y SHOULD either withdraw all its NLRIs for V, or let other PEs in the VPLS V know in some way that Y is no longer connected to its CEs.

If Y withdraws an NLRI for V that X was using, then X MUST tear down its ends of the pseudowires between X and Y.

The format of the VPLS NLRI is given below. The AFI and SAFI are the same as for the L2 VPN NLRI [3].

Figure 2: BGP NLRI for VPLS Information

```
   +------------------------------------+
   |  Length (2 octets)                 |
   +------------------------------------+
   |  Route Distinguisher  (8 octets)   |
   +------------------------------------+
   |  VE ID (2 octets)                  |
   +------------------------------------+
   |  VE Block Offset (2 octets)        |
   +------------------------------------+
   |  VE Block Size (2 octets)          |
   +------------------------------------+
   |  Label Base (3 octets)             |
   +------------------------------------+
```

3.2.2. Signaling PE Capabilities

The Encaps Type and Control Flags are encoded in an extended attribute. The community type also is used in L2 VPNs [3].

The Encaps Type for VPLS is 19.

Figure 3: layer2-info extended community

```
   +------------------------------------+
   | Extended community type (2 octets) |
   +------------------------------------+
   |  Encaps Type (1 octet)             |
   +------------------------------------+
   |  Control Flags (1 octet)           |
   +------------------------------------+
   |  Layer-2 MTU (2 octet)             |
   +------------------------------------+
   |  Reserved (2 octets)               |
   +------------------------------------+
```

Figure 4: Control Flags Bit Vector

```
    0 1 2 3 4 5 6 7
   +-+-+-+-+-+-+-+-+
   | MBZ |P|Q|F|C|S|      (MBZ = MUST Be Zero)
```

```
   +-+-+-+-+-+-+-+-+
```

With reference to Figure 4, the following bits are defined; the MBZ bits MUST be set to zero.

Name	Meaning
P	If set to 1, then the PE will strip the outermost VLAN tag from the customer frame on ingress, and push a VLAN tag on egress. If set to 0, the customer frame is left unchanged.
Q	Reserved.
F	If set to 1 (0), the PE is (not) capable of flooding.
C	If set to 1 (0), Control word is (not) required when encapsulating Layer 2 frames [10].
S	If set to 1 (0), Sequenced delivery of frames is (not) required.

3.3. Multi-AS VPLS

As in [3] and [7], the above autodiscovery and signaling functions are typically announced via I-BGP. This assumes that all the sites in a VPLS are connected to PEs in a single Autonomous System (AS).

However, sites in a VPLS may connect to PEs in different ASes. This leads to two issues: 1) there would not be an I-BGP connection between those PEs, so some means of signaling across ASes may be needed; and 2) there may not be PE-to-PE tunnels between the ASes.

A similar problem is solved in [7], Section 10. Three methods are suggested to address issue (1); all these methods have analogs in multi-AS VPLS.

Here is a diagram for reference:

```
     __________       ____________       ____________       __________
    /          \     /            \     /            \     /          \
                \___/        AS 1  \   /  AS 2        \___/
                                    \ /
      +-----+           +-------+    |    +-------+           +-----+
      | PE1 | ---...--- | ASBR1 | ======= | ASBR2 | ---...--- | PE2 |
      +-----+           +-------+    |    +-------+           +-----+
```

```
                 ___                / \                ___
                /   \              /   \              /   \
     \_________/     \____________/     \____________/     \_________/
```

a) VPLS-to-VPLS connections at the AS border routers.

In this method, an AS Border Router (ASBR1) acts as a PE for all VPLSs that span AS1 and an AS to which ASBR1 is connected, such as AS2 here. The ASBR on the neighboring AS (ASBR2) is viewed by

ASBR1 as a CE for the VPLSs that span AS1 and AS2; similarly, ASBR2 acts as a PE for this VPLS from AS2's point of view, and views ASBR1 as a CE.

This method does not require MPLS on the ASBR1-ASBR2 link, but does require that this link carry Ethernet traffic, and that there be a separate VLAN sub-interface for each VPLS traversing this link. It further requires that ASBR1 does the PE operations (discovery, signaling, MAC address learning, flooding, encapsulation, etc.) for all VPLSs that traverse ASBR1. This imposes a significant burden on ASBR1, both on the control plane and the data plane, which limits the number of multi-AS VPLSs.

Note that in general, there will be multiple connections between a pair of ASes, for redundancy. In this case, the Spanning Tree Protocol must be run on each VPLS that spans these ASes, so that a loop-free topology can be constructed in each VPLS. This imposes a further burden on the ASBRs and PEs participating in those VPLSs, as these devices would need to run the Spanning Tree Protocol for each such VPLS..

b) EBGP redistribution of VPLS information between ASBRs.

This method requires I-BGP peerings between the PEs in AS1 and ASBR1 in AS1 (perhaps via route reflectors), an E-BGP peering between ASBR1 and ASBR2 in AS2, and I-BGP peerings between ASBR2 and the PEs in AS2. In the above example, PE1 sends a VPLS NLRI to ASBR1 with a label block and itself as the BGP nexthop; ASBR1 sends the NLRI to ASBR2 with new labels and itself as the BGP nexthop; and ASBR2 sends the NLRI to PE2 with new labels and itself as the nexthop.

The VPLS NLRI that ASBR1 sends to ASBR2 (and the NLRI that ASBR2 sends to PE2) is identical to the VPLS NLRI that PE1 sends to ASBR1, except for the label block. To be precise, the Length, the Route Distinguisher, the VE ID, the VE Block Offset, and the VE Block Size MUST be the same; the Label Base may be different. Furthermore, ASBR1 must also update its forwarding path as follows: if the Label Base sent by PE1 is L1, the Label-block Size is N, the Label Base sent by ASBR1 is L2, and the tunnel label from ASBR1 to PE1 is T, then ASBR1 must install the following in the forwarding path:

```
    swap L2     with L1     and push T,
    swap L2+1   with L1+1   and push T,
    ...
    swap L2+N-1 with L1+N-1 and push T.
```

ASBR2 must act similarly, except that it may not need a tunnel label if it is directly connected with ASBR1.

When PE2 wants to send a VPLS packet to PE1, PE2 uses its VE ID to get the right VPLS label from ASBR2's label block for PE1, and uses a tunnel label to reach ASBR2. ASBR2 swaps the VPLS label with the label from ASBR1; ASBR1 then swaps the VPLS label with the label from PE1, and pushes a tunnel label to reach PE1.

In this method, one needs MPLS on the ASBR1-ASBR2 interface, but there is no requirement that the link layer be Ethernet. Furthermore, the ASBRs take part in distributing VPLS information. However, the data plane requirements of the ASBRs is much simpler than in method (a), being limited to label operations. Finally, the construction of loop-free VPLS topologies is done by routing decisions, viz. BGP path and nexthop selection, so there is no need to run the Spanning Tree Protocol on a per-VPLS basis. Thus, this method is considerably more scalable than method (a).

c) Multi-hop EBGP redistribution of VPLS information between ASes.

In this method, there is a multi-hop E-BGP peering between the PEs (or preferably, a Route Reflector) in AS1 and the PEs (or Route Reflector) in AS2. PE1 sends a VPLS NLRI with labels and nexthop self to PE2; if this is via route reflectors, the BGP nexthop is not changed. This requires that there be a tunnel LSP from PE1 to PE2. This tunnel LSP can be created exactly as in [7], section 10 (c), for example using E-BGP to exchange labeled IPv4 routes for the PE loopbacks.

When PE1 wants to send a VPLS packet to PE2, it pushes the VPLS label corresponding to its own VE ID onto the packet. It then pushes the tunnel label(s) to reach PE2.

This method requires no VPLS information (in either the control or the data plane) on the ASBRs. The ASBRs only need to set up PE-to-PE tunnel LSPs in the control plane, and do label operations in the data plane. Again, as in the case of method (b), the construction of loop-free VPLS topologies is done by routing decisions, i.e., BGP path and nexthop selection, so there is no need to run the Spanning Tree Protocol on a per-VPLS basis. This option is likely to be the most scalable of the three methods presented here.

In order to ease the allocation of VE IDs for a VPLS that spans multiple ASes, one can allocate ranges for each AS. For example, AS1 uses VE IDs in the range 1 to 100, AS2 from 101 to 200, etc. If there are 10 sites attached to AS1 and 20 to AS2, the allocated VE

IDs could be 1-10 and 101 to 120. This minimizes the number of VPLS NLRIs that are exchanged while ensuring that VE IDs are kept unique.

In the above example, if AS1 needed more than 100 sites, then another range can be allocated to AS1. The only caveat is that there is no overlap between VE ID ranges among ASes. The exception to this rule is multi-homing, which is dealt with below.

3.4. Multi-homing and Path Selection

It is often desired to multi-home a VPLS site, i.e., to connect it to multiple PEs, perhaps even in different ASes. In such a case, the PEs connected to the same site can either be configured with the same VE ID or with different VE IDs. In the latter case, it is mandatory to run STP on the CE device, and possibly on the PEs, to construct a loop-free VPLS topology.

In the case where the PEs connected to the same site are assigned the same VE ID, a loop-free topology is constructed by routing mechanisms, in particular, by BGP path selection. When a BGP speaker receives two equivalent NLRIs (see below for the definition), it applies standard path selection criteria such as Local Preference and AS Path Length to determine which NLRI to choose; it MUST pick only one. If the chosen NLRI is subsequently withdrawn, the BGP speaker applies path selection to the remaining equivalent VPLS NLRIs to pick another; if none remain, the forwarding information associated with that NLRI is removed.

Two VPLS NLRIs are considered equivalent from a path selection point of view if the Route Distinguisher, the VE ID and the VE Block Offset are the same. If two PEs are assigned the same VE ID in a given VPLS, they MUST use the same Route Distinguisher, and they MUST announce the same VE Block Size for a given VE Offset.

4. Data Plane

This section discusses two aspects of the data plane for PEs and u-PEs implementing VPLS: encapsulation and forwarding.

4.1. Encapsulation

Ethernet frames received from CE devices are encapsulated for transmission over the packet switched network connecting the PEs. The encapsulation is as in [10], with one change: a PE that sets the P bit in the Control Flags strips the outermost VLAN from an Ethernet frame received from a CE before encapsulating it, and pushes a VLAN onto a decapsulated frame before sending it to a CE.

4.2. Forwarding

Forwarding of VPLS packets is based on the interface over which the packet is received, which determines which VPLS the packet belongs to, and the destination MAC address. The former mapping is determined by configuration. The latter is the focus of this section.

4.2.1. MAC address learning

As was mentioned earlier, the key distinguishing feature of VPLS is that it is a multipoint service. This means that the entire Service Provider network should appear as a single logical learning bridge for each VPLS that the SP network supports. The logical ports for the SP "bridge" are the connections from the SP edge, be it a PE or a u-PE, to the CE. Just as a learning bridge learns MAC addresses on its ports, the SP bridge must learn MAC addresses at its VEs.

Learning consists of associating source MAC addresses of packets with the (logical) ports on which they arrive; this association is the Forwarding Information Base (FIB). The FIB is used for forwarding packets. For example, suppose the bridge receives a packet with source MAC address S on (logical) port P. If subsequently, the bridge receives a packet with destination MAC address S, it knows that it should send the packet out on port P.

There are two modes of learning: qualified and unqualified learning.

In qualified learning, the learning decisions at the VE are based on the customer ethernet packet's MAC address and VLAN tag, if one exists. This VLAN is often called the "service delimiting VLAN". Each VLAN on a given port is mapped to a different service (VPLS, IP VPN, point-to-point Layer 2 VPN, etc.); each VLAN that is mapped to a VPLS service has its own VPLS FIB.

In unqualified learning, learning is based on a customer ethernet packet's MAC address only. This is also called "port-mode VPLS".

4.2.2. Flooding

When a bridge receives a packet to a destination that is not in its FIB, it floods the packet on all the other ports. Similarly, a VE will flood packets to an unknown destination to all other VEs in the VPLS.

In Figure 1 above, if CE2 sent an Ethernet frame to PE2, and the destination MAC address on the frame was not in PE2's FIB (for that VPLS), then PE2 would be responsible for flooding that frame to every

other PE in the same VPLS. On receiving that frame, PE1 would be responsible for further flooding the frame to CE1 and CE5 (unless PE1 knew which CE "owned" that MAC address).

On the other hand, if PE3 received the frame, it could delegate further flooding of the frame to its u-PE. If PE3 was connected to 2 u-PEs, it would announce that it has two u-PEs. PE3 could either announce that it is incapable of flooding, in which case it would receive two frames, one for each u-PE, or it could announce that it is capable of flooding, in which case it would receive one copy of the frame, which it would then send to both u-PEs.

4.2.3. "Split Horizon" Flooding

When a PE capable of flooding receives a broadcast Ethernet frame, or one with an unknown destination MAC address, it must flood the frame. If the frame arrived from an attached CE, the PE must send a copy of the frame to every other attached CE, as well as to all PEs participating in the VPLS. If the frame arrived from another PE, however, the PE must only send a copy of the packet to attached CEs. The PE MUST NOT send the frame to other PEs. This notion has been termed "split horizon" flooding, and is a consequence of the PEs being logically full-meshed -- if a broadcast frame is received from PEx, then PEx would have sent a copy to all other PEs.

5. Deployment Options

In deploying a network that supports VPLS, the SP must decide whether the VPLS-aware device closest to the customer (the VE) is a u-PE or a PE. The default case described in this document is that the VE is a PE. However, there are a number of reasons that the VE might be a u-PE, i.e., a device that does layer 2 functions such as MAC address learning and flooding, and some limited layer 3 functions such as communicating to its PE, but doesn't do full-fledged discovery and PE-to-PE signaling.

As both of these cases have benefits, one would like to be able to "mix and match" these scenarios. The signaling mechanism presented here allows this. PE1 may be directly connected to CE devices; PE2 may be connected to u-PEs that are connected to CEs; and PE3 may be connected directly to a customer over some interfaces and to u-PEs over others. All these PEs do discovery and signaling in the same manner. How they do learning and forwarding depends on whether or not there is a u-PE; however, this is a local matter, and is not signaled.

6. Normative References

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

[6] Bates, T., Rekhter, Y., Chandra, R., and Katz, D., "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000

[9] Sangli, S., D. Tappan, and Y. Rekhter, "BGP Extended Communities Attribute", draft-ietf-idr-bgp-ext-communities-07.txt (work in progress)

[10] Martini, L., et al, "Encapsulation Methods for Transport of Ethernet Frames Over IP/MPLS Networks", draft-ietf-pwe3-ethernet-encap-06.txt (work in progress)

[11] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option," RFC 2385, August 1998

7. Informative References

[2] Andersson, L., and Rosen, E., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", draft-ietf-l2vpn-l2-framework-04.txt (work in progress)

[3] Kompella, K., (Editor), "Layer 2 VPNs Over Tunnels", draft-kompella-l2vpn-l2vpn-00.txt (work in progress)

[4] Martini, L., et al, "Pseudowire Setup and Maintenance using LDP" draft-ietf-pwe3-control-protocol-06.txt (work in progress)

[5] Kompella, V., et al, "Virtual Private LAN Services over MPLS", draft-ietf-ppvpn-vpls-ldp-03.txt (work in progress)

[7] Rosen, E., and Rekhter, Y., Editors, "BGP/MPLS VPNs", draft-ietf-l3vpn-rfc2547bis-01.txt (work in progress)

[8] Ould-Brahim, H., Rosen, E., and Rekhter, Y., "Using BGP as an Auto-Discovery Mechanism for Layer-3 and Layer-2 VPNs", draft-ietf-l3vpn-bgpvpn-auto-04.txt (work in progress)

Security Considerations

The focus in Virtual Private LAN Service is the privacy of data, i.e., that data in a VPLS is only distributed to other nodes in that VPLS and not to any external agent or other VPLS. Note that VPLS does not offer security or authentication: VPLS packets are sent in the clear in the packet-switched network, and a man-in-the-middle can eavesdrop, and may be able to inject packets into the data stream. If security is desired, the PE-to-PE tunnels can be IPsec tunnels. For more security, the end systems in the VPLS sites can use appropriate means of encryption to secure their data even before it enters the Service Provider network.

There are two aspects to achieving data privacy in a VPLS: securing the control plane, and protecting the forwarding path. Compromise of the control plane could result in a PE sending data belonging to some VPLS to another VPLS, or blackholing VPLS data, or even sending it to an eavesdropper, none of which are acceptable from a data privacy point of view. Since all control plane exchanges are via BGP, techniques such as in [11] help authenticate BGP messages, making it harder to spoof updates (which can be used to divert VPLS traffic to the wrong VPLS), or withdraws (denial of service attacks). In the multi-AS options (b) and (c), this also means protecting the inter-AS BGP sessions, between the ASBRs, the PEs or the Route Reflectors. Note that [11] will not help in keeping VPLS labels private -- knowing the labels, one can eavesdrop on VPLS traffic. However, this requires access to the data path within a Service Provider network.

Protecting the data plane requires ensuring that PE-to-PE tunnels are well-behaved (this is outside the scope of this document), and that VPLS labels are accepted only from valid interfaces. For a PE, valid interfaces comprise links from P routers. For an ASBR, a valid interface is a link from an ASBR in an AS that is part of a given VPLS. It is especially important in the case of multi-AS VPLSs that one accept VPLS packets only from valid interfaces.

IANA Considerations

IANA is asked to allocate an AFI for Layer 2 information (suggested value: 25).

Contributors

The following contributed to this document:

Javier Achirica, Telefonica
Loa Andersson, TLA
Chaitanya Kodeboyina, Juniper
Giles Heron, Consultant
Sunil Khandekar, Alcatel
Vach Kompella, Alcatel
Marc Lasserre, Riverstone
Pierre Lin, Yipes
Pascal Menezes, Terabeam
Ashwin Moranganti, Appian
Hamid Ould-Brahim, Nortel
Seo Yeong-il, Korea Tel

Acknowledgments

Thanks to Joe Regan and Alfred Nothaft for their contributions.

Authors' Addresses

Kireeti Kompella

Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
kireeti@juniper.net

Yakov Rekhter
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
yakov@juniper.net

IPR Notice

The IETF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on the IETF's procedures with respect to rights in standards-track and standards-related documentation can be found in BCP-11. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementors or users of this specification can be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

Full Copyright Notice

Copyright (C) The Internet Society (2004). 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 assignees.

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."

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.

Network Working Group
Internet-Draft
Expires: December 23, 2006

K. Kompella, Ed.
Y. Rekhter, Ed.
Juniper Networks

June 21, 2006

Virtual Private LAN Service (VPLS) Using BGP for Auto-discovery and Signaling

draft-ietf-l2vpn-vpls-bgp-08

Status of this Memo

This Internet-Draft will expire on December 23, 2006.

Copyright Notice

Abstract

Virtual Private LAN (Local Area Network) Service (VPLS), also known as Transparent LAN Service, and Virtual Private Switched Network service, is a useful Service Provider offering. The service offers a Layer 2 Virtual Private Network (VPN); however, in the case of VPLS, the customers in the VPN are connected by a multipoint Ethernet LAN, in contrast to the usual Layer 2 VPNs, which are point-to-point in

nature.

This document describes the functions required to offer VPLS, a mechanism for signaling a VPLS, and rules for forwarding VPLS frames across a packet switched network.

Table of Contents

1. Introduction

Virtual Private LAN Service (VPLS), also known as Transparent LAN Service, and Virtual Private Switched Network service, is a useful service offering. A Virtual Private LAN appears in (almost) all respects as an Ethernet LAN to customers of a Service Provider. However, in a VPLS, the customers are not all connected to a single LAN; the customers may be spread across a metro or wide area. In essence, a VPLS glues together several individual LANs across a packet-switched network to appear and function as a single LAN ([9]). This is accomplished by incorporating MAC address learning, flooding and forwarding functions in the context of pseudowires that connect these individual LANs across the packet-switched network.

This document details the functions needed to offer VPLS, and then goes on to describe a mechanism for the autodiscovery of the endpoints of a VPLS as well as for signaling a VPLS. It also describes how VPLS frames are transported over tunnels across a packet switched network. The autodiscovery and signaling mechanism uses BGP as the control plane protocol. This document also briefly discusses deployment options, in particular, the notion of decoupling functions across devices.

Alternative approaches include: [14], which allows one to build a Layer 2 VPN with Ethernet as the interconnect; and [13]), which allows one to set up an Ethernet connection across a packet-switched network. Both of these, however, offer point-to-point Ethernet services. What distinguishes VPLS from the above two is that a VPLS offers a multipoint service. A mechanism for setting up pseudowires for VPLS using the Label Distribution Protocol (LDP) is defined in [10].

1.1. Scope of this Document

This document has four major parts: defining a VPLS functional model; defining a control plane for setting up VPLS; defining the data plane for VPLS (encapsulation and forwarding of data); and defining various deployment options.

The functional model underlying VPLS is laid out in Section 2. This describes the service being offered, the network components that interact to provide the service, and at a high level their interactions.

The control plane described in this document uses Multiprotocol BGP [4] to establish VPLS service, i.e., for the autodiscovery of VPLS members and for the setup and teardown of the pseudowires that constitute a given VPLS instance. Section 3 focuses on this, and

also describes how a VPLS that spans Autonomous System boundaries is set up, as well as how multi-homing is handled. Using BGP as the control plane for VPNs is not new (see [14], [6] and [11]): what is described here is based on the mechanisms proposed in [6].

The forwarding plane and the actions that a participating Provider Edge (PE) router offering the VPLS service must take is described in Section 4.

In Section 5, the notion of 'decoupled' operation is defined, and the interaction of decoupled and non-decoupled PEs is described. Decoupling allows for more flexible deployment of VPLS.

1.2. Conventions used in this document

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 ([1]).

1.3. Changes from version 06 to 07

[NOTE to RFC Editor: this section is to be removed before publication.]

Note: the DISCUSSes below are referred to by id; they can be accessed at https://datatracker.ietf.org/public/pidtracker.cgi?command=view_comment&id=[ID]

Updated title of doc to reflect use of BGP. (Fenner's DISCUSS id 44901).

Addressed Russ Housley's DISCUSSes on Figure 6 and Section 6 (ids 44778 and 44779).

Addressed Sam Hartman's DISCUSS on the Security Considerations (id 48432).

Resolution of Kessens' DISCUSS (id 44870):

1. Reference to RFC 4364 has been made normative. There is no normative text in ref draft-kompella-l2vpn-l2vpn -- any such text has long since been incorporated directly into this document.

2. Description and IANA section updated.

3. Expanded section (b) of Section 3.4 to clarify the data plane operation for option b.

4. Updated Section 3.5 to clarify that a VPLS customer can run STP independent of whether the SP uses multi-homing or not.

5. P bit text deleted (left over from an earlier edit.)

6. Addressed (hopefully) by Sam's DISCUSS.

7. Updated Security Considerations to incorporate the techniques described in RFC 4364 for inter-AS VPNs. Also, added a paragraph stating that misconfiguration could cause inter-VPLS connections, just as can happen with RFC 4364.

Updated references; added reference to RFC 4023.

1.4. Changes from version 05 to 06

[NOTE to RFC Editor: this section is to be removed before publication.]

Changes in response to GenART review.

Updated Abstract and Introduction to make it clear that VPLS is an Ethernet-based service.

Added sections on Aging, Broadcast and Multicast, Qualified and Unqualified learning and CoS. Also added a section on scaling the BGP control plane. These were requested for consistency between the BGP and LDP VPLS documents.

Added a section clarifying the concepts of label blocks, why they are necessary and how they are used.

For multi-AS operation, added a short introduction to the three options, comparing their usage.

Lots of clean-up: consistent usage of terms, expansion of acronyms before use, references.

1.5. Changes from version 04 to 05

[NOTE to RFC Editor: this section is to be removed before publication.]

Updated IANA section to reflect agreement with authors of [11] that the two docs should use the same AFI for L2VPN information.

Addressed comments received from Alex Zinin. No technical changes, but a more complete description to cover the issues that Alex raised:

1. encoding of BGP NEXT_HOP for the new AFI/SAFI is not described

2. VE ID, Block offset, Block size, Label base are not described anywhere

3. no information on how the receiving PE choose the PW label

4. section 3.2.2 talks about PE capabilities all of a sudden and introduces a L2 Info Community, whose fields and use are not described

Changes to address these:

1. Broke up section 3.2.1 into "Concepts" and "PW Setup".

2. Expanded section on "Signaling PE Capabilities".

3. Added a new section 3.3 "BGP VPLS Operation".

4. Minor tweaking, e.g. to fix section number references.

1.6. Changes from version 03 to 04

[NOTE to RFC Editor: this section is to be removed before publication.]

Incorporated IDR review comments from Eric Ji, Chaitanya Kodeboyina, and Mike Loomis. Most changes are clarifications and rewording for better readability. The substantive changes are to remove several flags from the control field.

2. Functional Model

This will be described with reference to the following figure.

```
                                                        -----
                                                       /  A1 \
         ----                                     ____CE1     |
        /    \          --------       --------  /    |       |
       |  A2 CE2-      /        \     /        PE1     \     /
        \    /   \    /          \___/          | \     -----
         ----     ---PE2                        |  \
                     |                          |   \   -----
                     | Service Provider Network |    \ /     \
                     |                          |     CE5  A5 |
                     |            ___           |   /  \     /
              |----|  \          /   \         PE4_/    -----
              |u-PE|--PE3       /     \       /
              |----|    --------       -------
       ----  /   |    ----
      /    \/    \   /    \               CE = Customer Edge Device
     |  A3 CE3    --CE4 A4 |              PE = Provider Edge Router
      \    /         \    /               u-PE = Layer 2 Aggregation
       ----           ----                A<n> = Customer site n
```

Figure 1: Example of a VPLS

2.1. Terminology

Terminology similar to that in [6] is used: a Service Provider (SP) network with P (Provider-only) and PE (Provider Edge) routers, and customers with CE (Customer Edge) devices. Here, however, there is an additional concept, that of a "u-PE", a Layer 2 PE device used for Layer 2 aggregation. The notion of u-PE is described further in Section 5. PE and u-PE devices are "VPLS-aware", which means that they know that a VPLS service is being offered. We will call these VPLS edge devices, which could be either a PE or an u-PE, a VE.

In contrast, the CE device (which may be owned and operated by either the SP or the customer) is VPLS-unaware; as far as the CE is concerned, it is connected to the other CEs in the VPLS via a Layer 2 switched network. This means that there should be no changes to a CE device, either to the hardware or the software, in order to offer VPLS.

A CE device may be connected to a PE or a u-PE via Layer 2 switches that are VPLS-unaware. From a VPLS point of view, such Layer 2 switches are invisible, and hence will not be discussed further.

Furthermore, a u-PE may be connected to a PE via Layer 2 and Layer 3

devices; this will be discussed further in a later section.

The term "demultiplexor" refers to an identifier in a data packet that identifies both the VPLS to which the packet belongs as well as the ingress PE. In this document, the demultiplexor is an MPLS label.

The term "VPLS" will refer to the service as well as a particular instantiation of the service (i.e., an emulated LAN); it should be clear from the context which usage is intended.

2.2. Assumptions

The Service Provider Network is a packet switched network. The PEs are assumed to be (logically) fully meshed with tunnels over which packets that belong to a service (such as VPLS) are encapsulated and forwarded. These tunnels can be IP tunnels, such as GRE, or MPLS tunnels, established by RSVP-TE or LDP. These tunnels are established independently of the services offered over them; the signaling and establishment of these tunnels are not discussed in this document.

"Flooding" and MAC address "learning" (see Section 4) are an integral part of VPLS. However, these activities are private to an SP device, i.e., in the VPLS described below, no SP device requests another SP device to flood packets or learn MAC addresses on its behalf.

All the PEs participating in a VPLS are assumed to be fully meshed in the data plane, i.e., there is a bidirectional pseudowire between every pair of PEs participating in that VPLS, and thus every (ingress) PE can send a VPLS packet to the egress PE(s) directly, without the need for an intermediate PE (see Section 4.2.5.) This requires that VPLS PEs are logically fully meshed in the control plane so that a PE can send a message to another PE to set up the necessary pseudowires. See Section 3.6 for a discussion on alternatives to achieve a logical full mesh in the control plane.

2.3. Interactions

VPLS is a "LAN Service" in that CE devices that belong to VPLS V can interact through the SP network as if they were connected by a LAN. VPLS is "private" in that CE devices that belong to different VPLSs cannot interact. VPLS is "virtual" in that multiple VPLSs can be offered over a common packet switched network.

PE devices interact to "discover" all the other PEs participating in the same VPLS, and to exchange demultiplexors. These interactions are control-driven, not data-driven.

u-PEs interact with PEs to establish connections with remote PEs or u-PEs in the same VPLS. This interaction is control-driven.

PE devices can participate simultaneously in both VPLS and IP VPNs ([6]). These are independent services, and the information exchanged for each type of service is kept separate as the Network Layer Reachability Information (NLRI) used for this exchange have different Address Family Identifiers (AFI) and Subsequent Address Family Identifiers (SAFI). Consequently, an implementation MUST maintain a separate routing storage for each service. However, multiple services can use the same underlying tunnels; the VPLS or VPN label is used to demultiplex the packets belonging to different services.

3. Control Plane

There are two primary functions of the VPLS control plane: autodiscovery, and setup and teardown of the pseudowires that constitute the VPLS, often called signaling. Section 3.1 and Section 3.2 describe these functions. Both of these functions are accomplished with a single BGP Update advertisement; Section 3.3 describes how this is done by detailing BGP protocol operation for VPLS. Section 3.4 describes the setting up of pseudowires that span Autonomous Systems. Section 3.5 describes how multi-homing is handled.

3.1. Autodiscovery

Discovery refers to the process of finding all the PEs that participate in a given VPLS instance. A PE can either be configured with the identities of all the other PEs in a given VPLS, or the PE can use some protocol to discover the other PEs. The latter is called autodiscovery.

The former approach is fairly configuration-intensive, especially since it is required that the PEs participating in a given VPLS are fully meshed (i.e., that every PE in a given VPLS establish pseudowires to every other PE in that VPLS). Furthermore, when the topology of a VPLS changes (i.e., a PE is added to, or removed from the VPLS), the VPLS configuration on all PEs in that VPLS must be changed.

In the autodiscovery approach, each PE "discovers" which other PEs are part of a given VPLS by means of some protocol, in this case BGP. This allows each PE's configuration to consist only of the identity of the VPLS instance established on this PE, not the identity of every other PE in that VPLS instance -- that is auto-discovered. Moreover, when the topology of a VPLS changes, only the affected PE's configuration changes; other PEs automatically find out about the change and adapt.

3.1.1. Functions

A PE that participates in a given VPLS instance V must be able to tell all other PEs in VPLS V that it is also a member of V. A PE must also have a means of declaring that it no longer participates in a VPLS. To do both of these, the PE must have a means of identifying a VPLS and a means by which to communicate to all other PEs.

U-PE devices also need to know what constitutes a given VPLS; however, they don't need the same level of detail. The PE (or PEs) to which a u-PE is connected gives the u-PE an abstraction of the

VPLS; this is described in section 5.

3.1.2. Protocol Specification

The specific mechanism for autodiscovery described here is based on [14] and [6]; it uses BGP extended communities [5] to identify members of a VPLS, in particular, the Route Target community, whose format is described in [5]. The semantics of the use of Route Targets is described in [6]; their use in VPLS is identical.

As it has been assumed that VPLSs are fully meshed, a single Route Target RT suffices for a given VPLS V, and in effect that RT is the identifier for VPLS V.

A PE announces (typically via I-BGP) that it belongs to VPLS V by annotating its NLRIs for V (see next subsection) with Route Target RT, and acts on this by accepting NLRIs from other PEs that have Route Target RT. A PE announces that it no longer participates in V by withdrawing all NLRIs that it had advertised with Route Target RT.

3.2. Signaling

Once discovery is done, each pair of PEs in a VPLS must be able to establish (and tear down) pseudowires to each other, i.e., exchange (and withdraw) demultiplexors. This process is known as signaling. Signaling is also used to transmit certain characteristics of the pseudowires that a PE sets up for a given VPLS.

Recall that a demultiplexor is used to distinguish among several different streams of traffic carried over a tunnel, each stream possibly representing a different service. In the case of VPLS, the demultiplexor not only says to which specific VPLS a packet belongs, but also identifies the ingress PE. The former information is used for forwarding the packet; the latter information is used for learning MAC addresses. The demultiplexor described here is an MPLS label. However, note that the PE-to-PE tunnels need not be MPLS tunnels.

Using a distinct BGP Update message to send a demultiplexor to each remote PE would require the originating PE to send N such messages for N remote PEs. The solution described in this document allows a PE to send a single (common) Update message that contains demultiplexors for all the remote PEs, instead of N individual messages. Doing this reduces the control plane load both on the originating PE as well as on the BGP Route Reflectors that may be involved in distributing this Update to other PEs.

3.2.1. Label Blocks

To accomplish this, we introduce the notion of "label blocks". A label block, defined by a label base LB and a VE block size VBS, is a contiguous set of labels {LB, LB+1, ..., LB+VBS-1}. Here's how label blocks work. All PEs within a given VPLS are assigned unique VE IDs as part of their configuration. A PE X wishing to send a VPLS update sends the same label block information to all other PEs. Each receiving PE infers the label intended for PE X by adding their (unique) VE ID to the label base. In this manner, each receiving PE gets a unique demultiplexor for PE X for that VPLS.

This simple notion is enhanced with the concept of a VE block offset VBO. A label block defined by <LB, VBO, VBS> is the set {LB+VBO, LB+VBO+1, ..., LB+VBO+VBS-1}. Thus, instead of a single large label block to cover all VE IDs in a VPLS, one can have several label blocks, each with a different label base. This makes label block management easier, and also allows PE X to cater gracefully to a PE joining a VPLS with a VE ID that is not covered by the set of label blocks that that PE X has already advertised.

When a PE starts up, or is configured with a new VPLS instance, the BGP process may wish to wait to receive several advertisements for that VPLS instance from other PEs to improve the efficiency of label block allocation.

3.2.2. VPLS BGP NLRI

The VPLS BGP NLRI described below, with a new AFI and SAFI (see [4]) is used to exchange VPLS membership and demultiplexors.

A VPLS BGP NLRI has the following information elements: a VE ID, a VE Block Offset, a VE Block Size and a label base. The format of the VPLS NLRI is given below. The AFI is the L2VPN AFI (to be assigned by IANA), and the SAFI is the VPLS SAFI (65). The Length field is in octets.

```
+------------------------------------+
|  Length (2 octets)                 |
+------------------------------------+
|  Route Distinguisher  (8 octets)   |
+------------------------------------+
```

```
|  VE ID (2 octets)                     |
+---------------------------------------+
|  VE Block Offset (2 octets)           |
+---------------------------------------+
|  VE Block Size (2 octets)             |
+---------------------------------------+
|  Label Base (3 octets)                |
+---------------------------------------+
```

Figure 2: BGP NLRI for VPLS Information

A PE participating in a VPLS must have at least one VE ID. If the PE is the VE, it typically has one VE ID. If the PE is connected to several u-PEs, it has a distinct VE ID for each u-PE. It may additionally have a VE ID for itself, if it itself acts as a VE for that VPLS. In what follows, we will call the PE announcing the VPLS NLRI PE-a, and we will assume that PE-a owns VE ID V (either belonging to PE-a itself, or to a u-PE connected to PE-a).

VE IDs are typically assigned by the network administrator. Their scope is local to a VPLS. A given VE ID should belong to only one PE, unless a CE is multi-homed (see Section 3.5).

A label block is a set of demultiplexor labels used to reach a given VE ID. A VPLS BGP NLRI with VE ID V, VE Block Offset VBO, VE Block Size VBS and label base LB communicates to its peers the following:

label block for V: labels from LB to (LB + VBS - 1), and

remote VE set for V: from VBO to (VBO + VBS - 1).

There is a one-to-one correspondence between the remote VE set and the label block: VE ID (VBO + n) corresponds to label (LB + n).

3.2.3. PW Setup and Teardown

Suppose PE-a is part of VPLS foo, and makes an announcement with VE ID V, VE Block Offset VBO, VE Block Size VBS and label base LB. If PE-b is also part of VPLS foo, and has VE ID W, PE-b does the following:

1. checks if W is part of PE-a's 'remote VE set': if VBO <= W < VBO + VBS, then W is part of PE-a's remote VE set. If not, PE-b

ignores this message, and skips the rest of this procedure.

2. sets up a PW to PE-a: the demultiplexor label to send traffic from PE-b to PE-a is computed as (LB + W - VBO).

3. checks if V is part of any 'remote VE set' that PE-b announced, i.e., PE-b checks if V belongs to some remote VE set that PE-b announced, say with VE Block Offset VBO', VE Block Size VBS' and label base LB'. If not, PE-b MUST make a new announcement as described in Section 3.3.

4. sets up a PW from PE-a: the demultiplexor label over which PE-b should expect traffic from PE-a is computed as: (LB' + V - VBO').

If Y withdraws an NLRI for V that X was using, then X MUST tear down its ends of the pseudowire between X and Y.

3.2.4. Signaling PE Capabilities

The following extended attribute, the "Layer2 Info Extended Community", is used to signal control information about the pseudowires to be setup for a given VPLS. The extended community value is to be allocated by IANA (currently used value is 0x800A). This information includes the Encaps Type (type of encapsulation on the pseudowires), Control Flags (control information regarding the pseudowires) and the Maximum Transmission Unit (MTU) to be used on the pseudowires.

The Encaps Type for VPLS is 19.

```
+------------------------------------+
| Extended community type (2 octets) |
+------------------------------------+
|  Encaps Type (1 octet)             |
+------------------------------------+
|  Control Flags (1 octet)           |
+------------------------------------+
|  Layer-2 MTU (2 octet)             |
+------------------------------------+
|  Reserved (2 octets)               |
+------------------------------------+
```

Figure 3: Layer2 Info Extended Community

```
     0 1 2 3 4 5 6 7
    +-+-+-+-+-+-+-+-+
    |   MBZ     |C|S|      (MBZ = MUST Be Zero)
    +-+-+-+-+-+-+-+-+
```

Figure 4: Control Flags Bit Vector

With reference to Figure 4, the following bits in the Control Flags are defined; the remaining bits, designated MBZ, MUST be set to zero when sending and MUST be ignored when receiving this community.

Name	Meaning
C	A Control word ([7]) MUST or MUST NOT be present when sending VPLS packets to this PE, depending on whether C is 1 or 0, respectively
S	Sequenced delivery of frames MUST or MUST NOT be used when sending VPLS packets to this PE. depending on whether S is 1 or 0, respectively

3.3. BGP VPLS Operation

To create a new VPLS, say VPLS foo, a network administrator must pick a RT for VPLS foo, say RT-foo. This will be used by all PEs that serve VPLS foo. To configure a given PE, say PE-a, to be part of VPLS foo, the network administrator only has to choose a VE ID V for PE-a. (If PE-a is connected to u-PEs, PE-a may be configured with more than one VE ID; in that case, the following is done for each VE ID). The PE may also be configured with a Route Distinguisher (RD); if not, it generates a unique RD for VPLS foo. Say the RD is RD-foo-a. PE-a then generates an initial label block and a remote VE set for V, defined by VE Block Offset VBO, VE Block Size VBS and label base LB. These may be empty.

PE-a then creates a VPLS BGP NLRI with RD RD-foo-a, VE ID V, VE Block Offset VBO, VE Block Size VBS and label base LB. To this, it attaches a Layer2 Info Extended Community and a RT, RT-foo. It sets the BGP Next Hop for this NLRI as itself, and announces this NLRI to its peers. The Network Layer protocol associated with the Network

Address of the Next Hop for the combination <AFI=L2VPN AFI, SAFI=VPLS SAFI> is IP; this association is required by [4], Section 5. If the value of the Length of the Next Hop field is 4, then the Next Hop contains an IPv4 address. If this value is 16, then the Next Hop contains an IPv6 address.

If PE-a hears from another PE, say PE-b, a VPLS BGP announcement with RT-foo and VE ID W, then PE-a knows that PE-b is a member of the same

VPLS (autodiscovery). PE-a then has to set up its part of a VPLS pseudowire between PE-a and PE-b, using the mechanisms in Section 3.2. Similarly, PE-b will have discovered that PE-a is in the same VPLS, and PE-b must set up its part of the VPLS pseudowire. Thus, signaling and pseudowire setup is also achieved with the same Update message.

If W is not in any remote VE set that PE-a announced for VE ID V in VPLS foo, PE-b will not be able to set up its part of the pseudowire to PE-a. To address this, PE-a can choose to withdraw the old announcement(s) it made for VPLS foo, and announce a new Update with a larger remote VE set and corresponding label block that covers all VE IDs that are in VPLS foo. This however, may cause some service disruption. An alternative for PE-a is to create a new remote VE set and corresponding label block, and announce them in a new Update, without withdrawing previous announcements.

If PE-a's configuration is changed to remove VE ID V from VPLS foo, then PE-a MUST withdraw all its announcements for VPLS foo that contain VE ID V. If all of PE-a's links to its CEs in VPLS foo go down, then PE-a SHOULD either withdraw all its NLRIs for VPLS foo, or let other PEs in the VPLS foo know in some way that PE-a is no longer connected to its CEs.

3.4. Multi-AS VPLS

As in [14] and [6], the above autodiscovery and signaling functions are typically announced via I-BGP. This assumes that all the sites in a VPLS are connected to PEs in a single Autonomous System (AS).

However, sites in a VPLS may connect to PEs in different ASes. This leads to two issues: 1) there would not be an I-BGP connection between those PEs, so some means of signaling across ASes is needed;

and 2) there may not be PE-to-PE tunnels between the ASes.

A similar problem is solved in [6], Section 10. Three methods are suggested to address issue (1); all these methods have analogs in multi-AS VPLS.

Here is a diagram for reference:

```
      __________       ____________       ____________       __________
     /          \     /            \     /            \     /          \
                 \___/        AS 1  \   /  AS 2        \___/
                                     \ /
       +-----+           +-------+    |    +-------+           +-----+
       | PE1 | ---...--- | ASBR1 | ======= | ASBR2 | ---...--- | PE2 |
       +-----+           +-------+    |    +-------+           +-----+
                  ___                / \                ___
                 /   \              /   \              /   \
     \__________/     \____________/     \____________/     \__________/
```

Figure 6: Inter-AS VPLS

As in the above reference, three methods for signaling inter-provider VPLS are given; these are presented in order of increasing scalability. Method (a) is the easiest to understand conceptually, and the easiest to deploy; however, it requires an Ethernet interconnect between the ASes, and both VPLS control and data plane state on the AS border routers (ASBRs). Method (b) requires VPLS control plane state on the ASBRs and MPLS on the AS-AS interconnect (which need not be Ethernet). Method (c) requires MPLS on the AS-AS interconnect, but no VPLS state of any kind on the ASBRs.

3.4.1. a) VPLS-to-VPLS connections at the ASBRs.

In this method, an AS Border Router (ASBR1) acts as a PE for all VPLSs that span AS1 and an AS to which ASBR1 is connected, such as AS2 here. The ASBR on the neighboring AS (ASBR2) is viewed by ASBR1 as a CE for the VPLSs that span AS1 and AS2; similarly, ASBR2 acts as a PE for this VPLS from AS2's point of view, and views ASBR1 as a CE.

This method does not require MPLS on the ASBR1-ASBR2 link, but does require that this link carry Ethernet traffic, and that there be a separate VLAN sub-interface for each VPLS traversing this link. It further requires that ASBR1 does the PE operations (discovery, signaling, MAC address learning, flooding, encapsulation, etc.) for all VPLSs that traverse ASBR1. This imposes a significant burden on ASBR1, both on the control plane and the data plane, which limits the number of multi-AS VPLSs.

Note that in general, there will be multiple connections between a pair of ASes, for redundancy. In this case, the Spanning Tree Protocol (STP) ([15]), or some other means of loop detection and prevention, must be run on each VPLS that spans these ASes, so that a loop-free topology can be constructed in each VPLS. This imposes a further burden on the ASBRs and PEs participating in those VPLSs, as these devices would need to run a loop detection algorithm for each such VPLS. How this may be achieved is outside the scope of this document.

3.4.2. b) EBGP redistribution of VPLS information between ASBRs.

This method requires I-BGP peerings between the PEs in AS1 and ASBR1 in AS1 (perhaps via route reflectors), an E-BGP peering between ASBR1 and ASBR2 in AS2, and I-BGP peerings between ASBR2 and the PEs in AS2. In the above example, PE1 sends a VPLS NLRI to ASBR1 with a label block and itself as the BGP nexthop; ASBR1 sends the NLRI to ASBR2 with new labels and itself as the BGP nexthop; and ASBR2 sends the NLRI to PE2 with new labels and itself as the nexthop. Correspondingly, there are three tunnels: T1 from PE1 to ASBR1, T2 from ASBR1 to ASBR2, and T3 from ASBR2 to PE2. Within each tunnel, the VPLS label to be used is determined by the receiving device; e.g., the VPLS label within T1 is a label from the label block that ASBR1 sent to PE1. The ASBRs are responsible for receiving VPLS packets encapsulated in a tunnel, and performing the appropriate label swap operations described next so that the next receiving device can correctly identify and forward the packet.

The VPLS NLRI that ASBR1 sends to ASBR2 (and the NLRI that ASBR2 sends to PE2) is identical to the VPLS NLRI that PE1 sends to ASBR1, except for the label block. To be precise, the Length, the Route Distinguisher, the VE ID, the VE Block Offset, and the VE Block Size MUST be the same; the Label Base may be different. Furthermore, ASBR1 must also update its forwarding path as follows: if the Label Base sent by PE1 is L1, the Label-block Size is N, the Label Base sent by ASBR1 is L2, and the tunnel label from ASBR1 to PE1 is T, then ASBR1 must install the following in the forwarding path:

swap L2 with L1 and push T,

swap L2+1 with L1+1 and push T, ...

swap L2+N-1 with L1+N-1 and push T.

ASBR2 must act similarly, except that it may not need a tunnel label if it is directly connected with ASBR1.

When PE2 wants to send a VPLS packet to PE1, PE2 uses its VE ID to get the right VPLS label from ASBR2's label block for PE1, and uses a tunnel label to reach ASBR2. ASBR2 swaps the VPLS label with the label from ASBR1; ASBR1 then swaps the VPLS label with the label from PE1, and pushes a tunnel label to reach PE1.

In this method, one needs MPLS on the ASBR1-ASBR2 interface, but

there is no requirement that the link layer be Ethernet. Furthermore, the ASBRs take part in distributing VPLS information. However, the data plane requirements of the ASBRs is much simpler than in method (a), being limited to label operations. Finally, the construction of loop-free VPLS topologies is done by routing decisions, viz. BGP path and nexthop selection, so there is no need to run the Spanning Tree Protocol on a per-VPLS basis. Thus, this method is considerably more scalable than method (a).

3.4.3. c) Multi-hop EBGP redistribution of VPLS information between ASes.

In this method, there is a multi-hop E-BGP peering between the PEs (or preferably, a Route Reflector) in AS1 and the PEs (or Route Reflector) in AS2. PE1 sends a VPLS NLRI with labels and nexthop self to PE2; if this is via route reflectors, the BGP nexthop is not changed. This requires that there be a tunnel LSP from PE1 to PE2. This tunnel LSP can be created exactly as in [6], section 10 (c), for example using E-BGP to exchange labeled IPv4 routes for the PE loopbacks.

When PE1 wants to send a VPLS packet to PE2, it pushes the VPLS label corresponding to its own VE ID onto the packet. It then pushes the tunnel label(s) to reach PE2.

This method requires no VPLS information (in either the control or the data plane) on the ASBRs. The ASBRs only need to set up PE-to-PE tunnel LSPs in the control plane, and do label operations in the data plane. Again, as in the case of method (b), the construction of loop-free VPLS topologies is done by routing decisions, i.e., BGP path and nexthop selection, so there is no need to run the Spanning Tree Protocol on a per-VPLS basis. This option is likely to be the most scalable of the three methods presented here.

3.4.4. Allocation of VE IDs Across Multiple ASes

In order to ease the allocation of VE IDs for a VPLS that spans multiple ASes, one can allocate ranges for each AS. For example, AS1 uses VE IDs in the range 1 to 100, AS2 from 101 to 200, etc. If there are 10 sites attached to AS1 and 20 to AS2, the allocated VE IDs could be 1-10 and 101 to 120. This minimizes the number of VPLS NLRIs that are exchanged while ensuring that VE IDs are kept unique.

In the above example, if AS1 needed more than 100 sites, then another range can be allocated to AS1. The only caveat is that there be no overlap between VE ID ranges among ASes. The exception to this rule is multi-homing, which is dealt with below.

3.5. Multi-homing and Path Selection

It is often desired to multi-home a VPLS site, i.e., to connect it to multiple PEs, perhaps even in different ASes. In such a case, the PEs connected to the same site can either be configured with the same

VE ID or with different VE IDs. In the latter case, it is mandatory to run STP on the CE device, and possibly on the PEs, to construct a loop-free VPLS topology. How this can be accomplished is outside the scope of this document; however, the rest of this section will describe in some detail the former case. Note that multi-homing by the SP and STP on the CEs can co-exist; thus it is recommended that the VPLS customer run STP if the CEs are able to.

In the case where the PEs connected to the same site are assigned the same VE ID, a loop-free topology is constructed by routing mechanisms, in particular, by BGP path selection. When a BGP speaker receives two equivalent NLRIs (see below for the definition), it applies standard path selection criteria such as Local Preference and AS Path Length to determine which NLRI to choose; it MUST pick only one. If the chosen NLRI is subsequently withdrawn, the BGP speaker applies path selection to the remaining equivalent VPLS NLRIs to pick another; if none remain, the forwarding information associated with that NLRI is removed.

Two VPLS NLRIs are considered equivalent from a path selection point of view if the Route Distinguisher, the VE ID and the VE Block Offset are the same. If two PEs are assigned the same VE ID in a given VPLS, they MUST use the same Route Distinguisher, and they SHOULD announce the same VE Block Size for a given VE Offset.

3.6. Hierarchical BGP VPLS

This section discusses how one can scale the VPLS control plane when using BGP. There are at least three aspects of scaling the control plane:

1. alleviating the full mesh connectivity requirement among VPLS BGP speakers;

2. limiting BGP VPLS message passing to just the interested speakers rather than all BGP speakers; and

3. simplifying the addition and deletion of BGP speakers, whether for VPLS or other applications.

Fortunately, the use of BGP for Internet routing as well as for IP VPNs has yielded several good solutions for all these problems. The basic technique is hierarchy, using BGP Route Reflectors (RRs) ([8]).

The idea is to designate a small set of Route Reflectors which are themselves fully meshed, and then establish a BGP session between each BGP speaker and one or more RRs. In this way, there is no need of direct full mesh connectivity among all the BGP speakers. If the particular scaling needs of a provider requires a large number of RRs, then this technique can be applied recursively: the full mesh connectivity among the RRs can be brokered by yet another level of RRs. The use of RRs solves problems 1 and 3 above.

It is important to note that RRs, as used for VPLS and VPNs, are purely a control plane technique. The use of RRs introduces no data plane state and no data plane forwarding requirements on the RRs, and does not in any way change the forwarding path of VPLS traffic. This is in contrast to the technique of Hierarchical VPLS defined in [10].

Another consequence of this approach is that it is not required that one set of RRs handles all BGP messages, or that a particular RR handle all messages from a given PE. One can define several sets of RRs, for example a set to handle VPLS, another to handle IP VPNs and another for Internet routing. Another partitioning could be to have some subset of VPLSs and IP VPNs handled by one set of RRs, and another subset of VPLSs and IP VPNs handled by another set of RRs; the use of Route Target Filtering (RTF), described in [12] can make this simpler and more effective.

Finally, problem 2 (that of limiting BGP VPLS message passing to just the interested BGP speakers) is addressed by the use of RTF. This technique is orthogonal to the use of RRs, but works well in conjunction with RRs. RTF is also very effective in inter-AS VPLS; more details on how RTF works and its benefits are provided in [12].

It is worth mentioning an aspect of the control plane that is often a source of confusion. No MAC addresses are exchanged via BGP. All MAC address learning and aging is done in the data plane individually by each PE. The only task of BGP VPLS message exchange is autodiscovery and label exchange.

Thus, BGP processing for VPLS occurs when

1. a PE joins or leaves a VPLS; or

2. a failure occurs in the network, bringing down a PE-PE tunnel or a PE-CE link.

These events are relatively rare, and typically, each such event causes one BGP update to be generated. Coupled with BGP's messaging efficiency when used for signaling VPLS, these observations lead to the conclusion that BGP as a control plane for VPLS will scale quite

well both in terms of processing and memory requirements.

4. Data Plane

This section discusses two aspects of the data plane for PEs and u-PEs implementing VPLS: encapsulation and forwarding.

4.1. Encapsulation

Ethernet frames received from CE devices are encapsulated for transmission over the packet switched network connecting the PEs. The encapsulation is as in [7].

4.2. Forwarding

VPLS packets are classified as belonging to a given service instance and associated forwarding table based on the interface over which the packet is received. Packets are forwarded in the context of the service instance based on the destination MAC address. The former mapping is determined by configuration. The latter is the focus of this section.

4.2.1. MAC address learning

As was mentioned earlier, the key distinguishing feature of VPLS is that it is a multipoint service. This means that the entire Service Provider network should appear as a single logical learning bridge for each VPLS that the SP network supports. The logical ports for the SP "bridge" are the customer ports as well as the pseudowires on a VE. Just as a learning bridge learns MAC addresses on its ports, the SP bridge must learn MAC addresses at its VEs.

Learning consists of associating source MAC addresses of packets with the (logical) ports on which they arrive; this association is the Forwarding Information Base (FIB). The FIB is used for forwarding packets. For example, suppose the bridge receives a packet with source MAC address S on (logical) port P. If subsequently, the bridge receives a packet with destination MAC address S, it knows that it should send the packet out on port P.

If a VE learns a source MAC address S on logical port P, then later sees S on a different port P', then the VE MUST update its FIB to reflect the new port P'. A VE MAY implement a mechanism to damp flapping of source ports for a given MAC address.

4.2.2. Aging

VPLS PEs SHOULD have an aging mechanism to remove a MAC address associated with a logical port, much the same as learning bridges do. This is required so that a MAC address can be relearned if it "moves" from a logical port to another logical port, either because the station to which that MAC address belongs really has moved, or because of a topology change in the LAN that causes this MAC address to arrive on a new port. In addition, aging reduces the size of a VPLS MAC table to just the active MAC addresses, rather than all MAC addresses in that VPLS.

The "age" of a source MAC address S on a logical port P is the time since it was last seen as a source MAC on port P. If the age exceeds the aging time T, S MUST be flushed from the FIB. This of course means that every time S is seen as a source MAC address on port P, S's age is reset.

An implementation SHOULD provide a configurable knob to set the aging time T on a per-VPLS basis. In addition, an implementation MAY accelerate aging of all MAC addresses in a VPLS if it detects certain situations, such as a Spanning Tree topology change in that VPLS.

4.2.3. Flooding

When a bridge receives a packet to a destination that is not in its FIB, it floods the packet on all the other ports. Similarly, a VE will flood packets to an unknown destination to all other VEs in the VPLS.

In Figure 1 above, if CE2 sent an Ethernet frame to PE2, and the destination MAC address on the frame was not in PE2's FIB (for that VPLS), then PE2 would be responsible for flooding that frame to every other PE in the same VPLS. On receiving that frame, PE1 would be responsible for further flooding the frame to CE1 and CE5 (unless PE1 knew which CE "owned" that MAC address).

On the other hand, if PE3 received the frame, it could delegate further flooding of the frame to its u-PE. If PE3 was connected to 2 u-PEs, it would announce that it has two u-PEs. PE3 could either announce that it is incapable of flooding, in which case it would receive two frames, one for each u-PE, or it could announce that it is capable of flooding, in which case it would receive one copy of the frame, which it would then send to both u-PEs.

4.2.4. Broadcast and Multicast

There is a well-known broadcast MAC address. An Ethernet frame whose destination MAC address is the broadcast MAC address must be sent to all stations in that VPLS. This can be accomplished by the same means that is used for flooding.

There is also an easily recognized set of "multicast" MAC addresses.

Ethernet frames with a destination multicast MAC address MAY be broadcast to all stations; a VE MAY also use certain techniques to restrict transmission of multicast frames to a smaller set of receivers, those that have indicated interest in the corresponding multicast group. Discussion of this is outside the scope of this document.

4.2.5. "Split Horizon" Forwarding

When a PE capable of flooding (say PEx) receives a broadcast Ethernet frame, or one with an unknown destination MAC address, it must flood the frame. If the frame arrived from an attached CE, PEx must send a copy of the frame to every other attached CE, as well as to all other PEs participating in the VPLS. If, on the other hand, the frame arrived from another PE (say PEy), PEx must send a copy of the packet only to attached CEs. PEx MUST NOT send the frame to other PEs, since PEy would have already done so. This notion has been termed "split horizon" forwarding, and is a consequence of the PEs being logically fully meshed for VPLS.

Split horizon forwarding rules apply to broadcast and multicast packets, as well as packets to an unknown MAC address.

4.2.6. Qualified and Unqualified Learning

The key for normal Ethernet MAC learning is usually just the (6-octet) MAC address. This is called "unqualified learning". However, it is also possible that the key for learning includes the VLAN tag when present; this is called "qualified learning".

In the case of VPLS, learning is done in the context of a VPLS instance, which typically corresponds to a customer. If the customer uses VLAN tags, one can make the same distinctions of qualified and unqualified learning. If the key for learning within a VPLS is just the MAC address, then this VPLS is operating under unqualified learning. If the key for learning is (customer VLAN tag + MAC address), then this VPLS is operating under qualified learning.

Choosing between qualified and unqualified learning involves several factors, the most important of which is whether one wants a single global broadcast domain (unqualified), or a broadcast domain per VLAN (qualified). The latter makes flooding and broadcasting more efficient, but requires larger MAC tables. These considerations apply equally to normal Ethernet forwarding and to VPLS.

4.2.7. Class of Service

In order to offer different Classes of Service within a VPLS, an implementation MAY choose to map 802.1p bits in a customer Ethernet frame with a VLAN tag to an appropriate setting of EXP bits in the pseudowire and/or tunnel label, allowing for differential treatment of VPLS frames in the packet-switched network.

To be useful, an implementation SHOULD allow this mapping function to be different for each VPLS, as each VPLS customer may have their own view of the required behavior for a given setting of 802.1p bits.

5. Deployment Options

In deploying a network that supports VPLS, the SP must decide what functions the VPLS-aware device closest to the customer (the VE) supports. The default case described in this document is that the VE is a PE. However, there are a number of reasons that the VE might be a device that does all the Layer 2 functions (such as MAC address learning and flooding), and a limited set of Layer 3 functions (such as communicating to its PE), but, for example, doesn't do full-fledged discovery and PE-to-PE signaling. Such a device is called a "u-PE".

As both of these cases have benefits, one would like to be able to "mix and match" these scenarios. The signaling mechanism presented here allows this. For example, in a given provider network, one PE may be directly connected to CE devices; another may be connected to u-PEs that are connected to CEs; and a third may be connected directly to a customer over some interfaces and to u-PEs over others. All these PEs perform discovery and signaling in the same manner. How they do learning and forwarding depends on whether or not there is a u-PE; however, this is a local matter, and is not signaled. However, the details of the operation of a u-PE and its interactions with PEs and other u-PEs is beyond the scope of this document.

6. Security Considerations

The focus in Virtual Private LAN Service is the privacy of data, i.e., that data in a VPLS is only distributed to other nodes in that VPLS and not to any external agent or other VPLS. Note that VPLS does not offer confidentiality, integrity, or authentication: VPLS packets are sent in the clear in the packet-switched network, and a man-in-the-middle can eavesdrop, and may be able to inject packets into the data stream. If security is desired, the PE-to-PE tunnels can be IPsec tunnels. For more security, the end systems in the VPLS sites can use appropriate means of encryption to secure their data even before it enters the Service Provider network.

There are two aspects to achieving data privacy in a VPLS: securing the control plane, and protecting the forwarding path. Compromise of the control plane could result in a PE sending data belonging to some VPLS to another VPLS, or blackholing VPLS data, or even sending it to an eavesdropper, none of which are acceptable from a data privacy point of view. Since all control plane exchanges are via BGP, techniques such as in [2] help authenticate BGP messages, making it harder to spoof updates (which can be used to divert VPLS traffic to the wrong VPLS), or withdraws (denial of service attacks). In the multi-AS options (b) and (c), this also means protecting the inter-AS BGP sessions, between the ASBRs, the PEs or the Route Reflectors. One can also use the techniques described in section 10 (b) and (c) of [6], both for the control plane and the data plane. Note that [2] will not help in keeping VPLS labels private -- knowing the labels, one can eavesdrop on VPLS traffic. However, this requires access to the data path within a Service Provider network.

There can also be misconfiguration leading to unintentional connection of CEs in different VPLSs. This can be caused, for example, by associating the wrong Route Target with a VPLS instance. This problem, shared by [6], is for further study.

Protecting the data plane requires ensuring that PE-to-PE tunnels are well-behaved (this is outside the scope of this document), and that VPLS labels are accepted only from valid interfaces. For a PE, valid interfaces comprise links from P routers. For an ASBR, a valid interface is a link from an ASBR in an AS that is part of a given VPLS. It is especially important in the case of multi-AS VPLSs that one accept VPLS packets only from valid interfaces.

MPLS-in-IP and MPLS-in-GRE tunneling are specified in [3]. If it is desired to use such tunnels to carry VPLS packets, then the security considerations described in Section 8 of that document must be fully understood. Any implementation of VPLS that allows VPLS packets to be tunneled as described in that document MUST contain an

implementation of IPsec that can be used as therein described. If the tunnel is not secured by IPsec, then the technique of IP address filtering at the border routers, described in Section 8.2 of that document, is the only means of ensuring that a packet that exits the tunnel at a particular egress PE was actually placed in the tunnel by the proper tunnel head node (i.e., that the packet does not have a spoofed source address). Since border routers frequently filter only source addresses, packet filtering may not be effective unless the egress PE can check the IP source address of any tunneled packet it receives, and compare it to a list of IP addresses that are valid tunnel head addresses. Any implementation that allows MPLS-in-IP and/or MPLS-in-GRE tunneling to be used without IPsec MUST allow the egress PE to validate in this manner the IP source address of any tunneled packet that it receives.

7. IANA Considerations

IANA is asked to allocate an AFI for L2VPN information (suggested value: 25). This should be the same as the AFI requested by [11].

IANA is asked to allocate an extended community value for the Layer2 Info Extended Community (suggested value: 0x800a).

8. References

8.1. Normative References

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

[2] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

[3] Worster, T., Rekhter, Y., and E. Rosen, "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.

[4] Bates, T., "Multiprotocol Extensions for BGP-4", draft-ietf-idr-rfc2858bis-10 (work in progress), March 2006.

[5] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

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

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

8.2. Informative References

[8] Bates, T., Chandra, R., and E. Chen, "BGP Route Reflection - An Alternative to Full Mesh IBGP", RFC 2796, April 2000.

[9] Andersson, L. and E. Rosen, "Framework for Layer 2 Virtual Private Networks (L2VPNs)", draft-ietf-l2vpn-l2-framework-05 (work in progress), June 2004.

[10] Lasserre, M. and V. Kompella, "Virtual Private LAN Services Using LDP", draft-ietf-l2vpn-vpls-ldp-09 (work in progress), June 2006.

[11] Ould-Brahim, H., "Using BGP as an Auto-Discovery Mechanism for VR-based Layer-3 VPNs", draft-ietf-l3vpn-bgpvpn-auto-07 (work in progress), April 2006.

[12] Marques, P., "Constrained VPN Route Distribution", draft-ietf-l3vpn-rt-constrain-02 (work in progress), June 2005.

[13] Martini, L., "Pseudowire Setup and Maintenance using the Label

Distribution Protocol", draft-ietf-pwe3-control-protocol-17 (work in progress), June 2005.

[14] Kompella, K., "Layer 2 VPNs Over Tunnels", draft-kompella-l2vpn-l2vpn-01 (work in progress), January 2006.

[15] Institute of Electrical and Electronics Engineers, "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.", IEEE Standard 802.1D, July 1998.

Appendix A. Contributors

The following contributed to this document:

Javier Achirica, Telefonica
Loa Andersson, Acreo
Chaitanya Kodeboyina, Juniper
Giles Heron, Tellabs
Sunil Khandekar, Alcatel
Vach Kompella, Alcatel
Marc Lasserre, Riverstone
Pierre Lin

Pascal Menezes
Ashwin Moranganti, Appian
Hamid Ould-Brahim, Nortel
Seo Yeong-il, Korea Tel

Appendix B. Acknowledgements

Thanks to Joe Regan and Alfred Nothaft for their contributions. Many thanks too to Eric Ji, Chaitanya Kodeboyina, Mike Loomis and Elwyn Davies for their detailed reviews.

Authors' Addresses

Kireeti Kompella (editor)
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti@juniper.net

Yakov Rekhter (editor)
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: yakov@juniper.net

Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM 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.

Copyright Statement

Copyright (C) The Internet Society (2006). This document is subject to the rights, licenses and restrictions contained in BCP 78, and

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.

Internet Draft Document
Provider Provisioned VPN Working Group
draft-ietf-l2vpn-vpls-ldp-03.txt
Expires: October 2004

Marc Lasserre
Vach Kompella
(Editors)

April 2004

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

1. Status of this Memo

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].

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
  - LDP security (authentication) methods as described in [RFC-3036] SHOULD be applied. This would prevent unauthorized participation by 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.

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

Internet Draft Document
L2VPN Working Group
draft-ietf-l2vpn-vpls-ldp-07.txt
Expires: January 2006

Marc Lasserre
Vach Kompella
(Editors)

July 2005

Virtual Private LAN Services over MPLS

Status of this Memo

By submitting this Internet-Draft, we certify that any applicable patent or other IPR claims of which we are aware have been disclosed, or will be disclosed, and any of which we become aware will be disclosed, in accordance with RFC 3668.

This document is an Internet-Draft and is in full conformance with Sections 5 and 6 of RFC3667 and Section 5 of RFC3668.

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.

IPR Disclosure Acknowledgement

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

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, i.e., it creates 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 pseudo-wire 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].

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.

2. Table of Contents

3. 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 pseudo-wire basis, packet replication across pseudo-wires for multicast/broadcast traffic and for flooding of unknown unicast destination traffic.

[PWE3-ETHERNET] defines how to carry Layer 2 (L2) frames over point-to-point pseudo-wires (PW). 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.

The resulting set of interconnected devices forms a private MPLS VPN.

4. Topological Model for VPLS

An interface participating in a VPLS must be able to flood, forward, and filter Ethernet frames. The set of PE devices interconnected via PWs appears as a single emulated LAN to customer C1. 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.

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

```
                           |            +-- Emulated LAN
                        +-----+
                        | C1  |
                        +-----+
                        Site C
```

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 [PWE-CTRL]), 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]. In other words, the access 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 can be used. The use of IGMP snooping and PIM snooping techniques should be used to improve multicast efficiency. A description of these techniques is beyond the scope of this document.

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 PEs to 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 [BGP-VPN], 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: a PE MUST NOT forward traffic from one PW to another in the same VPLS mesh.

Note that customers are allowed to run the Spanning Tree Protocol (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 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 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 8 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 [PWE3-CTRL], 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 PWid FEC Element has been moved to Appendix 1. 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

[PWE3-CTRL] 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 [PWE3-CTRL].

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

VC info length: As specified in [PWE3-CTRL].

AGI, Length, Value: The unique name of this VPLS. The AGI identifies a type of name, Length denotes the length of Value, which is the name of the VPLS. We 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. The use of non-null TAII and SAII is reserved for future enhancements.

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 tagged mode, this parameter may be used to signal the insertion of the appropriate VLAN ID as specified in section 6.1.

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 a MAC 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 that is used to specify a list of MAC addresses that can be removed or unlearned 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).

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-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 #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 (subject to IANA approval). This identifies the TLV type as MAC List 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 MAC 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. 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 to 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 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 [PWE3-ETHERNET], the NSP 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 [PWE3-ETHERNET]. 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 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 [PWE3-ETHERNET] 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.

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 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 [PWE3-ETHERNET]) 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.

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 [PWE3-ETHERNET], the NSP 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 [PWE3-ETHERNET]. 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 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.

```
                                                            -----
                                                           /  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
            ----           ----
```

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

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 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.

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 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 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 MTU (multi-tenant unit) domain. This can be accomplished by treating the MTU as a PE and provisioning PWs between it and every other edge, as a basic VPLS. An alternative is to utilize [PWE3-ETHERNET] PWs 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 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, 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.

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 devices for supporting the spoke connections. For rest of this discussion we refer to a bridging capable MTU as MTU-s and a non-bridging capable PE as PE-r. We refer to a routing and bridging capable device as PE-rs.

10.1.1. Spoke connectivity for bridging-capable devices

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

```
                                                             \ --   /
       MTU-s = Bridging capable MTU                           ------
       PE-rs = VPLS capable PE                                PE3-rs
       Agg = Layer-2 Aggregation
       --
      /  \
      \S / = Virtual Switch Instance
       --
```

In the figure above where an MTU-s has a single connection to a PE-rs placed in the CO. 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 [PWE3-CTRL]. 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.

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 on to 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 in a CO or a multi-tenant building, or a PE-r might already be 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. As shown in this figure, the PE-r creates a point-to-point tunnel LSP to a PE-rs.

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

Then for every access port that needs to participate in a VPLS service, the PE-r creates a point-to-point PW that terminates on the physical port at the PE-r and terminates on the VSI of the VPLS service at the PE-rs.

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 setup from the PE-r to the PE-rs. Once the PWs are setup, 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.

10.2. Redundant Spoke Connections

An obvious weakness of the hub and spoke approach described thus far is that the MTU has a single connection to the PE-rs. In case of failure of the connection or the PE-rs, the MTU 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. The mechanism described is identical for both, MTU-s and PE-r devices.

10.2.1. Dual-homed MTU

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, as shown in figure-3. The PE-rs devices must be part of the same VPLS service instance.

An MTU-s can set 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, while the second PW is designated as secondary and is held in a standby state. The MTU 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 of the scope of this document. For example, a spanning tree instance running between only the MTU 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
```

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 which 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 4.2), 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.

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 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.

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 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 as 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 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 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 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 PWs is used among these PE-rs devices for carrying the SP BPDU packets for that island. On a per P-VLAN basis, 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.

12. Significant Modifications

Between rev 06 and this one, these are the changes:

- Incorporated comments from technical review team
- Clarifications and edits
- Fixed id-nits

13. Contributors

Loa Andersson, TLA
Ron Haberman, Alcatel
Juha Heinanen, Independent
Giles Heron, Tellabs
Sunil Khandekar, Alcatel
Luca Martini, Cisco
Pascal Menezes, Independent
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

14. 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.

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.

15. 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

    - 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 [RFC-3036] 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

16. IANA Considerations

The type field in the MAC TLV is defined as 0x404 in section 4.2.1 and is subject to IANA approval.

17. References

17.1. Normative References

[PWE3-ETHERNET] "Encapsulation Methods for Transport of Ethernet Frames Over IP/MPLS Networks", draft-ietf-pwe3-ethernet-encap-10.txt, Work in progress, June 2005.

[PWE3-CTRL] "Transport of Layer 2 Frames over MPLS", draft-ietf-pwe3-control-protocol-17.txt, Work in progress, June 2005.

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

[IANA] "IANA Allocations for pseudo Wire Edge to Edge Emulation (PWE3)" Martini,Townsley, draft-ietf-pwe3-iana-allocation-08.txt, Work in progress, February 2005.

17.2. Informative References

[BGP-VPN] "BGP/MPLS VPNs", draft-ietf-l3vpn-rfc2547bis-03.txt, Work in Progress, October 2004.

[RADIUS-DISC] "Using Radius for PE-Based VPN Discovery", draft-ietf-l2vpn-radius-pe-discovery-01.txt, Work in Progress, February

2005.

[BGP-DISC] "Using BGP as an Auto-Discovery Mechanism for Network-based VPNs", draft-ietf-l3vpn-bgpvpn-auto-06.txt, Work in Progress, June 2005.

[L2FRAME] "Framework for Layer 2 Virtual Private Networks (L2VPNs)", draft-ietf-l2vpn-l2-framework-05, Work in Progress, June 2004.

[L2VPN-REQ] "Service Requirements for Layer-2 Provider Provisioned Virtual Private Networks", draft-ietf-l2vpn-requirements-04.txt, Work in Progress, October 2005.

[VPN-SEC] "Security Framework for Provider Provisioned Virtual Private Networks", draft-ietf-l3vpn-security-framework-03.txt, Work in Progress, November 2004.

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

18. Appendix: 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 VC FEC TLV.

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

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

```
 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|         PW Type             |PW info Length |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
   |                       Group ID                                |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                         VCID                                  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                        Interface parameters                   |
   ~                                                               ~
   |                                                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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 [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, so it has global significance across all PEs involved in the VPLS instance.

19. Authors' Addresses

Marc Lasserre
Riverstone Networks
Email: marc@riverstonenet.com

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

IPR Disclosure Acknowledgement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Copyright Notice

Disclaimer

Internet Draft Document
L2VPN Working Group
draft-ietf-l2vpn-vpls-ldp-09.txt
Expires: Dec 2006

Marc Lasserre
Vach Kompella
(Editors)

June 2006

Virtual Private LAN Services Using LDP

Status of this Memo

IPR Disclosure Acknowledgement

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, i.e., it creates 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 pseudo-wire labels using LDP [RFC3036], extending [RFC4447]. 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 [RFC4448].

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].

2. Table of Contents

3. 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 pseudo-wire basis, packet replication across pseudo-wires 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 pseudo-wires (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.

3.1. 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 delimitor	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

3.2. 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 secutity
L2TP	Layer Two Tunneling Protocol
LAN	Local Area Network
LDP	Label Distribution Protocol
MTU-s	Multi-Tenant Unit switch
PE	Provider Edge device
PW	Pseudo-wire
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, 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]. 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 PEs to 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 [BGP-VPN], 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: 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 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 8 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 PWid FEC Element has been moved to Appendix 1. 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 [IANA].

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, 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 terminate 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 [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 #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 (subject to IANA approval). 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 to 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 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 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.

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 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 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, 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 on to 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 setup from the PE-r to the PE-rs. Once the PWs are setup, 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, while 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 of 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 which 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 4.2), 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 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.

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 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 as 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 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 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 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 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

Juha Heinanen, Independent
Giles Heron, Tellabs
Sunil Khandekar, Alcatel
Luca Martini, Cisco
Pascal Menezes, Independent
Rob Nath, Lucent
Eric Puetz, SBC
Vasile Radoaca, Independent
Ali Sajassi, Cisco

Yetik Serbest, SBC
Nick Slabakov, Juniper
Andrew Smith, Consultant
Tom Soon, SBC
Nick Tingle, Alcatel

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 [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
    - 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 and is subject to IANA approval.

16. References

16.1. Normative References

[RFC4447] "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", L. Martini, et al., April 2006.

[RFC4448] "Encapsulation Methods for Transport of Ethernet over MPLS Networks", L. Martini, et al., 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] "LDP Specification", L. Andersson, et al., RFC 3036, January 2001.

[IANA] "IANA Allocations for pseudo Wire Edge to Edge Emulation (PWE3)" Martini,Townsley, draft-ietf-pwe3-iana-allocation-08.txt, Work in progress, February 2005.

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

16.2. Informative References

[BGP-VPN] "BGP/MPLS VPNs", draft-ietf-l3vpn-rfc2547bis-03.txt, Work in Progress, October 2004.

[RADIUS-DISC] "Using Radius for PE-Based VPN Discovery", draft-ietf-l2vpn-radius-pe-discovery-01.txt, Work in Progress, February 2005.

[BGP-DISC] "Using BGP as an Auto-Discovery Mechanism for Network-based VPNs", draft-ietf-l3vpn-bgpvpn-auto-06.txt, Work in Progress, June 2005.

[L2FRAME] "Framework for Layer 2 Virtual Private Networks (L2VPNs)", draft-ietf-l2vpn-l2-framework-05, Work in Progress, June 2004.

[L2VPN-REQ] "Service Requirements for Layer-2 Provider Provisioned Virtual Private Networks", draft-ietf-l2vpn-requirements-04.txt, Work in Progress, October 2005.

[VPN-SEC] "Security Framework for Provider Provisioned Virtual Private Networks", draft-ietf-l3vpn-security-framework-03.txt, Work in Progress, November 2004.

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

17. Appendix: 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, 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.

18. Authors' Addresses

Marc Lasserre
Lucent Technologies
Email: mlasserre@lucent.com

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

IPR Disclosure Acknowledgement

assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Copyright Notice

Copyright (C) The Internet Society (2006). This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

Disclaimer

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM 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.

Network Working Group Luca Martini
Internet Draft Nasser El-Aawar
Expiration Date: January 2003 Level 3 Communications, LLC.

Giles Heron Steve Vogelsang
PacketExchange Ltd. Laurel Networks, Inc.

Chris Liljenstolpe Vasile Radoaca
Cable & Wireless Nortel Networks

Daniel Tappan Kireeti Kompella
Eric C. Rosen Juniper Networks
Cisco Systems, Inc.

Andrew G. Malis Tricci So
Vinai Sirkay Chris Flores
Vivace Networks, Inc. Consultant

XiPeng Xiao David Zelig
Redback Networks Corrigent Systems

Raj Sharma Loa Andersson
Luminous Networks, Inc. Utfors

Nick Tingle
Sunil Khandekar
TiMetra Networks

July 2002

Encapsulation Methods for Transport of Ethernet Frames Over IP and MPLS Networks

draft-martini-ethernet-encap-mpls-01.txt

Status of this Memo

Abstract

An Ethernet PW allows Ethernet/802.3 Protocol Data Units (PDUs) to be carried over Packet Switched Networks (PSNs) using IP, L2TP or MPLS transport. This enables Service Providers to leverage their existing PSN to offer Ethernet services.

This document describes methods for encapsulating Ethernet/802.3 PDUs for transport over an MPLS or IP network.

Table of Contents

1. Specification of Requirements

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

2. Introduction

In an MPLS or IP network, it is possible to use control protocols such as those specified in [MARTINI-TRANS] to set up "emulated virtual circuits" that carry the the Protocol Data Units of layer 2 protocols across the network. A number of these emulated virtual circuits may be carried in a single tunnel. This requires of course that the layer 2 PDUs be encapsulated. We can distinguish three layers of this encapsulation:

- the "tunnel header", which contains the information needed to transport the PDU across the IP or MPLS network; this is header belongs to the tunneling protocol, e.g., MPLS, GRE, L2TP.

- the "demultiplexer field", which is used to distinguish individual emulated virtual circuits within a single tunnel; this field must be understood by the tunneling protocol as well; it may be, e.g., an MPLS label or a GRE key field.

- the "emulated VC encapsulation", which contains the information about the enclosed layer 2 PDU which is necessary in order to properly emulate the corresponding layer 2 protocol.

This document specifies the emulated Virtual Circuit (VC) encapsulation for the ethernet protocols. Although different layer 2 protocols require different information to be carried in this encapsulation, an attempt has been made to make the encapsulation as common as possible for all layer 2 protocols. Other layer 2 protocols are described in separate documents. [MARTINI-ATM] [MARTINI-FRAME] [MARTINI-PPP]

This document also specifies the way in which the demultiplexer field is added to the emulated VC encapsulation when an MPLS label is used as the demultiplexer field.

The scope of this document also includes:

- Pseudo-wire (PW) requirements for emulating Ethernet trunking and switching behavior.

- PE-bound and CE-bound packet processing of Ethernet PDUs

- QoS and security considerations

- Inter-domain transport considerations for Ethernet PE

The following two figures describe the reference models which are derived from [PWE3-FRAME] [PWE3-REQ] to support the Ethernet PW emulated services.

```
        Native    |<----- Pseudo Wire ---->|  Native
       Ethernet   |                        |  Ethernet
          or      |  |<-- PSN Tunnel -->|  |    or
        VLAN      V  V                  V  V   VLAN
       Service  +----+                  +----+ Service
+----+    |     | PE1|==================| PE2|     |    +----+
|    |----------|............PW1.............|----------|    |
| CE1|    |     |    |                  |    |     |    |CE2 |
```

```
|    |----------|............PW2.............|----------|    |
+----+    |     |    |==================|    |     |    +----+
     ^          +----+                  +----+     |    ^
     |      Provider Edge 1         Provider Edge 2     |
     |                                                  |
     |<-------------- Emulated Service ---------------->|
```

Figure 1: PWE3 Ethernet/VLAN Interface Reference Configuration

```
 +-------------+                                +-------------+
 |  Emulated   |                                |  Emulated   |
 |  Ethernet   |                                |  Ethernet   |
 | (including  |         Emulated Service       | (including  |
 |  VLAN)      |<==============================>|  VLAN)      |
 |  Services   |                                |  Services   |
 +-------------+           Pseudo Wire          +-------------+
 |Demultiplexer|<==============================>|Demultiplexor|
 +-------------+                                +-------------+
 |    PSN      |            PSN Tunnel          |    PSN      |
 | MPLS or IP  |<==============================>| MPLS or IP  |
```

```
    +--------------+                               +--------------+
    |  Physical    |                               |  Physical    |
    +-----+--------+                               +-----+--------+
          |                                              |
          |                MPLS or IP Network            |
          |             ____     ___       ____          |
          |           _/    \___/   \    _/    \__       |
          |          /               \__/         \_     |
          |         /                               \    |
          +========/                                 |===+
                   \                                 /
                    \                               /
                     \   ___      ___     __      _/
                      \_/   \____/   \___/  \____/
```

Figure 2: Ethernet PWE3 Protocol Stack Reference Model

For the purpose of this document R1 will be defined as the ingress router, and R2 as the egress router. A layer 2 PDU will be received at R1, encapsulated at R1, transported, decapsulated at R2, and transmitted out of R2.

3. Requirements for Ethernet Pseudo-Wire Emulation

An Ethernet PW emulates a single Ethernet link between exactly two endpoints. The following reference model describes the termination point of each end of the PW within the PE:

```
              +-----------------------------------+
              |                PE                 |
      +---+   +-+  +-----+  +------+  +------+  +-+
      |   |   |P|  |     |  |PW ter|  | PSN  |  |P|
```

```
   |   |<==|h|<=| NSP |<=|minati|<=|Tunnel|<=|h|<== From PSN
   |   |   |y|  |     |  |on    |  |      |  |y|
   | C |   +-+  +-----+  +------+  +------+  +-+
   | E |   |                                   |
   |   |   +-+  +-----+  +------+  +------+  +-+
   |   |   |P|  |     |  |PW ter|  | PSN  |  |P|
   |   |==>|h|=>| NSP |=>|minati|=>|Tunnel|=>|h|==> To PSN
   |   |   |y|  |     |  |on    |  |      |  |y|
   +---+   +-+  +-----+  +------+  +------+  +-+
           |                                   |
           +-----------------------------------+
                       ^        ^
                       |        |
                       A        B
```

Figure 3: PW reference diagram

The PW terminates at a logical port within the PE, defined at point A in the above diagram. This port provides an Ethernet MAC service that will deliver each Ethernet packet that is received at point A, unaltered, to the point A in the corresponding PE at the other end of the PW.

The "NSP" function includes packet processing needed to translate the Ethernet packets that arrive at the CE-PE interface to/from the Ethernet packets that are applied to the PW termination point. Such functions may include stripping, overwriting or adding VLAN tags, physical port multiplexing and demultiplexing, PW-PW bridging, L2 encapsulation, shaping, policing, etc.

The points to the left of A, including the physical layer between the CE and PE, and any adaptation (NSP) functions between it and the PW terminations, are outside of the scope of PWE3 and are not defined here.

"PW Termination", between A and B, represents the operations for setting up and maintaining the PW, and for encapsulating and decapsulating the Ethernet packets according to the PSN type in use. This document defines these operations, and the services offered and required at points A and B.

"PSN Tunnel" denotes the PSN tunneling technology that is being used: MPLS or GRE/IP.

A pseudo wire can be one of the two types: raw or tagged. This is a property of the emulated Ethernet link and indicates whether the pseudo wire MUST contain an 802.1Q VLAN tag (i.e. tagged mode) or MAY contain a tag (i.e. raw mode).

3.1. Packet Processing

3.1.1. Encapsulation

The entire Ethernet frame without any preamble or FCS is transported as a single packet. A VC label is prepended to this and the packet is forwarded through a PSN tunnel (either MPLS or GRE/IP).

3.1.2. MTU Management

Ingress and egress PWESs MUST agree on their maximum MTU size to be transported over the PSN.

3.1.3. Frame Ordering

In general, applications running over Ethernet do not require strict frame ordering. However the IEEE definition of 802.3 [802.3] requires that frames from the same conversation are delivered in sequence. Moreover, the PSN cannot (in the general case) be assumed to provide or to guarantee frame ordering. Therefore if strict frame ordering is required, the control word defined below MUST be utilized and its sequence number processing enabled.

3.1.4. Frame Error Processing

An encapsulated Ethernet frame traversing a psuedo-wire may be dropped, corrupted or delivered out-of-order. Per [PWE3-REQ], packet-loss, corruption, and out-of-order delivery is considered to be a "generalized bit error" of the psuedo-wire. Therefore, the native Ethernet frame error processing mechanisms MUST be extended to the corresponding psuedo-wire service. Therefore, if a PE device receives an Ethernet frame containing hardware level CRC errors, framing errors, or a runt condition, the frame MUST be discarded on input. Note that this processing is part of the NSP function and is outside the scope of this draft.

3.1.5. IEEE 802.3x Flow Control Interworking

In a standard Ethernet network, the flow control mechanism is optional and typically configured between the two nodes on a point-to-point link (e.g. between the CE and the PE). IEEE 802.3x PAUSE frames MUST NOT be carried across the PW. See Appendix A for notes on CE-PE flow control.

3.2. Maintenance

It is desirable to have a signaling mechanism for establishing Ethernet PWs and for detecting failure of an Ethernet PW. It is recommended that the procedures defined in [MARTINI-TRANS] be used for this purpose.

3.3. Management

The PW management model of Ethernet PW follows the general management guidelines for PW management as appear in [PW-MIB] and defined in [PWE3-REQ], [PWE3-FRAME]. It is composed of 3 components. [PW-MIB] defines the parameters common to all types of PW and PSNs, for example common counters, error handling, some maintenance protocol parameters etc. For each type of PSN there is a separate module that defines the association of the PW to the PSN tunnel, see example in [PW-MPLS-MIB] for the MPLS PSN. For Ethernet PW, an additional MIB module [PW-ENET-MIB] defines the Ethernet specific parameters required to be configured or monitored.

The above modules enable both manual configuration and the use of maintenance procedures to set up the Ethernet PW and monitor PW state where applicable.

As specified in [PWE3-REQ] and [PWE3-FRAME], an implementation SHOULD support the relevant PW MIB modules for PW set-up and monitoring. Other mechanisms for PW set up (command line interface for example) MAY be supported.

3.4. QoS Considerations

The ingress PE MAY consider the user priority (PRI) field [802.1Q] of the VLAN tag header when determining the value to be placed in the Quality of Service field of the encapsulating protocol (e.g., the EXP fields of the MPLS label stack). In a similar way, the egress PE MAY consider the Quality of Service field of the encapsulating protocol when queuing the packet for CE-bound.

A PE MUST support the ability to carry the Ethernet PW as a best effort service over the PSN. Transparency of PRI bits (if sent from CE to PE) between CE devices, regardless of the COS support of the PSN. Where the 802.1Q VLAN field is added at the PE, a default PRI setting of zero MUST be supported, a configured default value is recommended.

A PE may support additional QOS support by means of one or more of the following methods:

-i. One COS per PW End Service (PWES), mapped to a single COS PW at the PSN.
-ii. Multiple COS per PWES mapped to a single PW with multiple COS at the PSN.
-iii. Multiple COS per PWES mapped to multiple PWs at the PSN.

Examples of the cases above and details of the service mapping considerations are described in Appendix B.

The PW guaranteed rate at the PSN level is PW provider policy based on agreement with the customer, and may be different from the Ethernet physical port rate. Consideration of Ethernet flow control was discussed above.

3.5. Security Considerations

This document specifies the security consideration regarding the encapsulation for the PW. In terms of encapsulation, security of the encapsulated packets depends on the nature of the protocol that is carried by these packets, while the encapsulation itself shall not affect the related security issues.

Nevertheless, the security limitations of the PE and/or the PW MUST not restrict the security implementation choices of the user of the PWE3 (i.e. users should be able to implement IPSEC or any other appropriate security mechanism in addition to the security inherent in the PW)".

It is required that PEs will have user separation between different PW and different virtual ports that the PWs are connected to. For example: if two PWs are connected to the same physical port and associated to different virtual ports (i.e. VLANs), it is required that packets from one VC will not be forwarded to the VLAN that is associated to the second VCs.

A received packet is associated with a PW by means of the VC label. However this mechanism provides no guarantee that the packet was sent by the peer PE. Further checks may be useful to protect against misconfiguration and connection hijacking.

The PE must be able to be protected from malformed, or maliciously altered, customer traffic. This includes, but is not limited to, illegal VLAN use, short packets, long packets, etc.

Security achieved by access control of MAC addresses is out of scope of this document.

Additional security requirements related to the use of PW in a switching (virtual bridging) environment are not discussed here as they are not within the scope of this draft.

In the case of a PW crossing from one autonomous system to another, through a private interconnection, security considerations are much the same as in the intra-domain case. However in some cases the PW may travel through a third-party autonomous system, or across a public interconnection point. In these cases there may be a requirement to encrypt the user data using a method appropriate to the PSN tunneling mechanism.

4. General encapsulation method

4.1. The Control Word

When carrying Ethernet over an IP or MPLS backbone sequentiality may need to be preserved. The OPTIONAL control word defined here addresses this requirement. Implementations MUST support sending no control word, and MAY support sending a control word.

In all cases the egress router must be aware of whether the ingress router will send a control word over a specific virtual circuit. This may be achieved by configuration of the routers, or by signaling, for example as defined in [MARTINI-TRANS].

The control word is defined as follows:

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Reserved            |      Sequence Number          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

In the above diagram the first 16 bits are reserved for future use. They MUST be set to 0 when transmitting, and MUST be ignored upon receipt.

The next 16 bits provide a sequence number that can be used to guarantee ordered packet delivery. The processing of the sequence number field is OPTIONAL.

The sequence number space is a 16 bit, unsigned circular space. The sequence number value 0 is used to indicate an unsequenced packet.

4.1.1. Setting the sequence number

For a given emulated VC, and a pair of routers R1 and R2, if R1 supports packet sequencing then the following procedures should be used:

- the initial packet transmitted on the emulated VC MUST use sequence number 1
- subsequent packets MUST increment the sequence number by one for each packet
- when the transmit sequence number reaches the maximum 16 bit

value (65535) the sequence number MUST wrap to 1

If the transmitting router R1 does not support sequence number processing, then the sequence number field in the control word MUST be set to 0.

4.1.2. Processing the sequence number

If a router R2 supports receive sequence number processing, then the following procedures should be used:

When an emulated VC is initially set up, the "expected sequence number" associated with it MUST be initialized to 1.

When a packet is received on that emulated VC, the sequence number should be processed as follows:

- if the sequence number on the packet is 0, then the packet passes the sequence number check

- otherwise if the packet sequence number >= the expected sequence number and the packet sequence number - the expected sequence number < 32768, then the packet is in order.

- otherwise if the packet sequence number < the expected sequence number and the expected sequence number - the packet sequence number >= 32768, then the packet is in order.

- otherwise the packet is out of order.

If a packet passes the sequence number check, or is in order then, it can be delivered immediately. If the packet is in order, then the expected sequence number should be set using the algorithm:

```
expected_sequence_number := packet_sequence_number + 1 mod 2**16
if (expected_sequence_number = 0) then expected_sequence_number := 1;
```

Packets which are received out of order MAY be dropped or reordered at the discretion of the receiver.

If a router R2 does not support receive sequence number processing, then the sequence number field MAY be ignored.

4.2. MTU Requirements

The network MUST be configured with an MTU that is sufficient to transport the largest encapsulation frames. If MPLS is used as the tunneling protocol, for example, this is likely to be 8 or more bytes greater than the largest frame size. Other tunneling protocols may have longer headers and require larger MTUs. If the ingress router determines that an encapsulated layer 2 PDU exceeds the MTU of the tunnel through which it must be sent, the PDU MUST be dropped. If an egress router receives an encapsulated layer 2 PDU whose payload length (i.e., the length of the PDU itself without any of the encapsulation headers), exceeds the MTU of the destination layer 2 interface, the PDU MUST be dropped.

4.3. Tagged Mode

In this mode each frame MUST include an 802.1Q field. All frames in a PW MUST have the same 802.1Q tag value. Note that the tag may be overwritten by the NSP function at ingress or at egress.

Note that when using the signaling procedures defined in [MARTINI-TRANS], such a PW should be signaled as being of type "Ethernet VLAN".

4.4. Raw Mode

In this mode each frame MAY include an 802.1Q field. Multiple 802.1Q tag values MAY be transported over the same PW.

Note that when using the signaling procedures defined in [MARTINI-TRANS], such a PW should be signaled as being of type "Ethernet".

5. Using an MPLS Label as the Demultiplexer Field

To use an MPLS label as the demultiplexer field, a 32-bit label stack entry [MPLS-LABEL] is simply prepended to the emulated VC encapsulation, and hence will appear as the bottom label of an MPLS label stack. This label may be called the "VC label". The particular emulated VC identified by a particular label value must be agreed by the ingress and egress LSRs, either by signaling (e.g, via the methods of [MARTINI-TRANS]) or by configuration. Other fields of the label stack entry are set as follows.

5.1. MPLS Shim EXP Bit Values

If it is desired to carry Quality of Service information, the Quality of Service information SHOULD be represented in the EXP field of the VC label. If more than one MPLS label is imposed by the ingress LSR, the EXP field of any labels higher in the stack SHOULD also carry the same value.

5.2. MPLS Shim S Bit Value

The ingress LSR, R1, MUST set the S bit of the VC label to a value of 1 to denote that the VC label is at the bottom of the stack.

5.3. MPLS Shim TTL Values

The ingress LSR, R1, SHOULD set the TTL field of the VC label to a value of 255.

6. Security Considerations

This document specifies only encapsulations, and not the protocols used to carry the encapsulated packets across the network. Each such protocol may have its own set of security issues, but those issues are not affected by the encapsulations specified herein.

Specific security issues related to encapsulation are addressed in the requirements section above.

7. Intellectual Property Disclaimer

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

8. References

[MARTINI-TRANS] "Transport of Layer 2 Frames Over MPLS", Martini, L., et al., draft-martini-l2circuit-trans-mpls-09.txt, (work in progress), June 2002.

[MARTINI-ATM] "Encapsulation Methods for Transport of ATM Cells/Frame Over IP and MPLS Networks", Martini L., et al., draft-martini-atm-encap-mpls-00.txt, (work in progress), June 2002.

[MARTINI-FRAME] "Encapsulation Methods for Transport of Frame-Relay Over IP and MPLS Networks", Martini, L., et al., draft-martini-frame-encap-mpls-00.txt, (work in progress), June 2002.

[MARTINI-PPP] "Encapsulation Methods for Transport of PPP/HDLC Frames Over IP and MPLS Networks", Martini L., et al., draft-martini-ppp-hdlc-encap-mpls-00.txt, (work in progress), April 2002.

[PWE3-REQ] "Requirements for Pseudo Wire Emulation Edge-to-Edge (PWE3)", Xiao, X., McPherson, D., Pate, P., White, C., Kompella, K., Gill, V., Nadeau, T., draft-pwe3-requirements-03.txt, (work in progress), June 2002.

[PWE3-FRAME] "Framework for Pseudo Wire Emulation Edge-to-Edge (PWE3)", Pate, P., Xiao, X., So, T., Malis, A., Nadeau, T., White, C., Kompella, K., Johnson, T., Bryant, S.,

draft-pate-pwe3-framework-03.txt, (work in progress),
June 2002.

[PW-MIB] "Pseudo Wire (PW) Management Information Base using SMIv2", Zelig, D., Mantin, S., Nadeau, T., Danenberg, D., draft-zelig-pw-mib-02.txt, (work in progress), February 2002.

[PW-MPLS-MIB] "Pseudo Wire (PW) over MPLS PSN Management Information Base", Zelig D., Mantin, S., Nadeau, T., Danenberg, D., Malis, A., draft-zelig-pw-mpls-mib-01.txt, (work in progress), February 2002.

[PW-ENET-MIB] "Ethernet Pseudo Wire (PW) Management Information Base", Zelig, D., Nadeau, T., draft-zelig-pw-enet-mib-00.txt, (work in progress) February 2002.

[802.3] IEEE, ISO/IEC 8802-3: 2000 (E), "IEEE Standard for Information technology -- Telecommunications and information exchange between systems -- Local and metropolitan area networks -- Specific requirements -- Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications", 2000.

[802.1Q] ANSI/IEEE Standard 802.1Q, "IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks", 1998.

[MPLS-LABEL] "MPLS Label Stack Encoding", Rosen, E., Rekhter, E., Tappan, D., Fedorkow, G., Farinacci, D., Li, T., Conta, A., RFC 3032.

9. Author Information

Luca Martini
Level 3 Communications, LLC.
1025 Eldorado Blvd.
Broomfield, CO, 80021
e-mail: luca@level3.net

Nasser El-Aawar

Level 3 Communications, LLC.
1025 Eldorado Blvd.
Broomfield, CO, 80021
e-mail: nna@level3.net

Giles Heron
PacketExchange Ltd.
The Truman Brewery
91 Brick Lane
LONDON E1 6QL
United Kingdom
e-mail: giles@packetexchange.net

Dan Tappan
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824
e-mail: tappan@cisco.com

Eric Rosen
Cisco Systems, Inc.
250 Apollo Drive
Chelmsford, MA, 01824
e-mail: erosen@cisco.com

Steve Vogelsang
Laurel Networks, Inc.
Omega Corporate Center
1300 Omega Drive
Pittsburgh, PA 15205
e-mail: sjv@laurelnetworks.com

Andrew G. Malis

Vivace Networks, Inc.
2730 Orchard Parkway
San Jose, CA 95134
e-mail: Andy.Malis@vivacenetworks.com

Vinai Sirkay
Vivace Networks, Inc.
2730 Orchard Parkway
San Jose, CA 95134
e-mail: sirkay@technologist.com

Vasile Radoaca
Nortel Networks
600 Technology Park
Billerica MA 01821
e-mail: vasile@nortelnetworks.com

Chris Liljenstolpe
Cable & Wireless
11700 Plaza America Drive
Reston, VA 20190
e-mail: chris@cw.net

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
e-mail: kireeti@juniper.net

Tricci So
e-mail: tricciso@yahoo.ca

XiPeng Xiao
Redback Networks
300 Holger Way,
San Jose, CA 95134
e-mail: xipeng@redback.com

Chris Flores
Austin, Texas
e-mail: chris_flores@hotmail.com

David Zelig
Corrigent Systems
126, Yigal Alon St.
Tel Aviv, ISRAEL
e-mail: davidz@corrigent.com

Raj Sharma
Luminous Netwokrs, Inc.
10460 Bubb Road
Cupertino, CA 95014
e-mail: raj@luminous.com

Nick Tingle
TiMetra Networks
274 Ferguson Drive
Mountain View, CA 94043
e-mail: nick@timetra.com

Sunil Khandekar
TiMetra Networks
274 Ferguson Drive
Mountain View, CA 94043

email: sunil@timetra.com

Loa Andersson
Utfors
P.O. Box 525,
SE-169 29 Solna, Sweden
e-mail: loa.andersson@utfors.se

Appendix A - Interoperability Guidelines

Configuration Options

The following is a list of the configuration options for a point-to-point Ethernet PW based on the reference points of Figure 3:

Service and Encap on A	Encap on C	Operation at B ingress/egress	Remarks
1) Raw	Raw - Same as A		
2) Tag1	Tag2	Optional change of VLAN value	VLAN can be 0-4095 Change allowed in

			both directions
3) No Tag	Tag	Add/remove Tag field	Tag can be 0-4095 (note i)
4) Tag	No Tag	Remove/add Tag field	(note ii)

Figure 4: Configuration Options

Allowed combinations:

Raw and other services are not allowed on the same physical port (A). All other combinations are allowed, except that conflicting VLANs on (A) are not allowed.

Notes:

-i. Mode #3 MAY be limited to adding VLAN NULL only, since change of VLAN or association to specific VLAN can be done at the PW CE-bound side.

-ii. Mode #4 exists in layer 2 switches, but is not recommended when operating with PW since it may not preserve the user's PRI bits. If there is a need to remove the VLAN tag (for TLS at the other end of the PW) it is recommended to use mode #2 with tag2=0 (NULL VLAN) on the PW and use mode #3 at the other end of the PW.

IEEE 802.3x Flow Control Considerations

If the receiving node becomes congested, it can send a special frame, called the PAUSE frame, to the source node at the opposite end of the connection. The implementation MUST provide a mechanism for terminating PAUSE frames locally (i.e. at the local PE). It MUST operate as follows:

PAUSE frames received on a local Ethernet port SHOULD cause the PE device to buffer, or to discard, further Ethernet frames for that port until the PAUSE condition is cleared. Optionally the PE MAY simply discard PAUSE frames.

If the PE device wishes to pause data received on a local Ethernet port (perhaps because its own buffers are filling up or because it has received notification of congestion within the PSN) then it MAY issue a PAUSE frame on the local Ethernet port, but MUST clear this condition when willing to receive more data.

Appendix B - QoS Details

Section 3.7 describes various modes for supporting PW QOS over the PSN. Examples of the above for a point to point VLAN service are:

- The classification to the PW is based on VLAN field only, regardless of the user PRI bits. The PW is assigned a specific COS (marking, scheduling, etc.) at the tunnel level.

- The classification to the PW is based on VLAN field, but the PRI bits of the user is mapped to different COS marking (and network behavior) at the PW level. Examples are DiffServ coding in case of IP PSN, and E-LSP in MPLS PSN.

- The classification to the PW is based on VLAN field and the PRI bits, and packets with different PRI bits are mapped to different PWs. An example is to map a PWES to different L-LSPs in MPLS PSN in order to support multiple COS service over an L-LSP capable network.

  The specific value to be assigned at the PSN for various COS is not specified and is application specific.

Adaptation of 802.1Q COS to PSN COS

It is not required that the PSN will have the same COS definition of COS as defined in [802.1Q], and the mapping of 802.1Q COS to PSN QOS is application specific and depends on the agreement between the customer and the PW provider. However, the following principles adopted from 802.1Q table 8-2 MUST be met when applying set of PSN COS based on user's PRI bits.

	#of available classes of service							
User Priority	1	2	3	4	5	6	7	8
0 Best Effort (Default)	0	0	0	1	1	1	1	2
1 Background	0	0	0	0	0	0	0	0
2 Spare	0	0	0	0	0	0	0	1
3 Excellent Effort	0	0	0	1	1	2	2	3
4 Controlled Load	0	1	1	2	2	3	3	4
5 Interactive Multimedia	0	1	1	2	3	4	4	5
6 Interactive Voice	0	1	2	3	4	5	5	6
7 Network Control	0	1	2	3	4	5	6	7

Figure 5: IEEE 802.1Q COS Service Mapping

Drop precedence

The 802.1P standard does not support drop precedence, therefore from the PW PE-bound point of view there is no mapping required. It is however possible to mark different drop precedence for different PW packets based on the operator policy and required network behavior. This functionality is not discussed further here.

PSN COS labels interaction with VC label COS marking

Marking of COS bits at the VC level is not required if the PSN tunnel is PE to PE based, since only the PSN COS marking is visible to the PSN network. In cases where the VC multiplexing field is carried without an external tunnel (for example directly connected PEs with PHP, or PEs connected using GRE/IP), the rules stated above for tunnel COS marking apply also for the VC level.

In summary, the rules for COS marking shall be as follows:

- If there is only a VC label then, it shall contain the appropriate CoS value (e.g. MPLS between PEs which are directly adjacent to each other).

- If the VC label and PSN tunnel labels are both being used, then the CoS marking on the PSN header shall be marked with the correct CoS value.

- If the PSN marking is stripped at a node before the PE, the PSN marking MUST be copied to the VC label. An example is MPLS PSN with the use of PHP.

  PSN QOS support and signaling of QOS is out of scope of this document.