PWE3 Working Group
Internet Draft
Expiration Date: March 2002

Tricci So
Caspian Networks
XiPeng Xiao
Loa Anderson
Utfors AB
Chris Flores
Nick Tingle
Sunil Khandeker
TiMetra Networks

David Zelig
Corrigent Systems
Giles Heron
PacketExchange Ltd.

Ethernet Pseudo Wire Emulation Edge-to-Edge (PWE3)

draft-so-pwe3-ethernet.txt

Status of this Memo

Abstract

This document describes the Psuedo Wire (PW) service specific implementation to support 802.3 and 802.1Q/VLAN Ethernet emulated services. An Ethernet PW allows Ethernet Protocol Data Units (PDUs) to be carried over Packet Switched Networks (PSNs) using IP, L2TP or MPLS transport. This will enable Service Providers to leverage their existing PSN to offer Ethernet services.

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.

Table Of Contents

1. Introduction

There is growing interest in using high speed and high performance IP and MPLS-enabled IP networks to transport legacy L2 technologies, such as Ethernet, Frame Relay and ATM as described in [Martini-encap], [Martini-trans], [Kompella] and [Rosen].

This draft defines encapsulation mechanisms to transport Ethernet traffic over the Packet Switched Networks (PSNs) using MPLS[MPLS], L2TP [L2TPv3] and GRE [GRE-encap] [GRE-IPv4] tunnels
This document defines the PDU processing, maintenance and encapsulation behaviors when emulating Ethernet services based on the PWE3 architecture[PWE3-frame] over a PSN. The scope of the document includes:

- Pseudo-wire (PW) requirements for emulating the Ethernet trunking and switching behavior.
- Setup of the Ethernet PW between PE devices over the PSN tunnel
- PE-bound and CE-bound packet processing of Ethernet PDUs
- Encapsulation of Ethernet PDUs over MPLS, IP/GRE and L2TP tunneling mechanisms
- Transport and delivery of encapsulated packets over Ethernet PW
- Maintenance function and interactions with the PSN tunnel for the Ethernet PW
- QoS and security considerations
- Inter-domain transport considerations for Ethernet PW

It is not within the scope of the document to specify how and when the PSN tunnel be set up. However, information needed to establish PWs over PSN is specified, as well as suggestions on how such PWs are maintained.

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

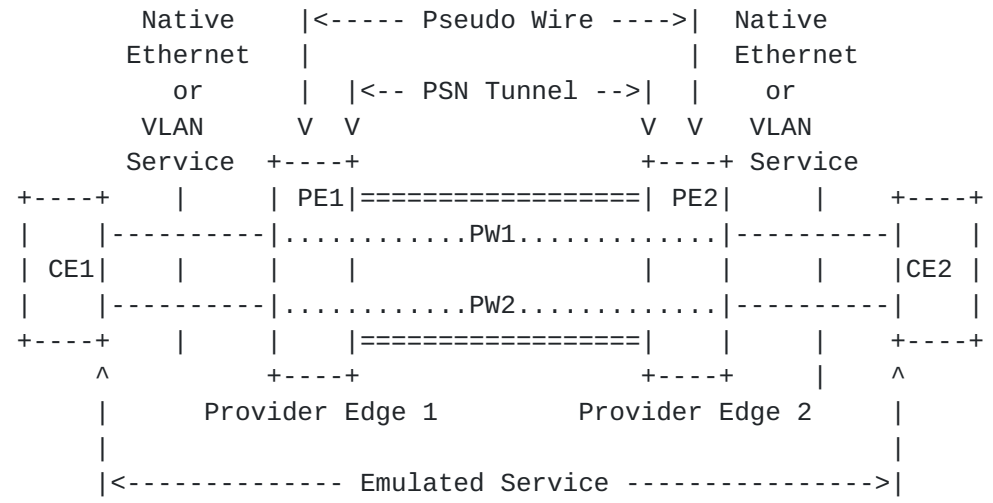


Figure 1: PWE3 Ethernet/VLAN Interface Reference Configuration

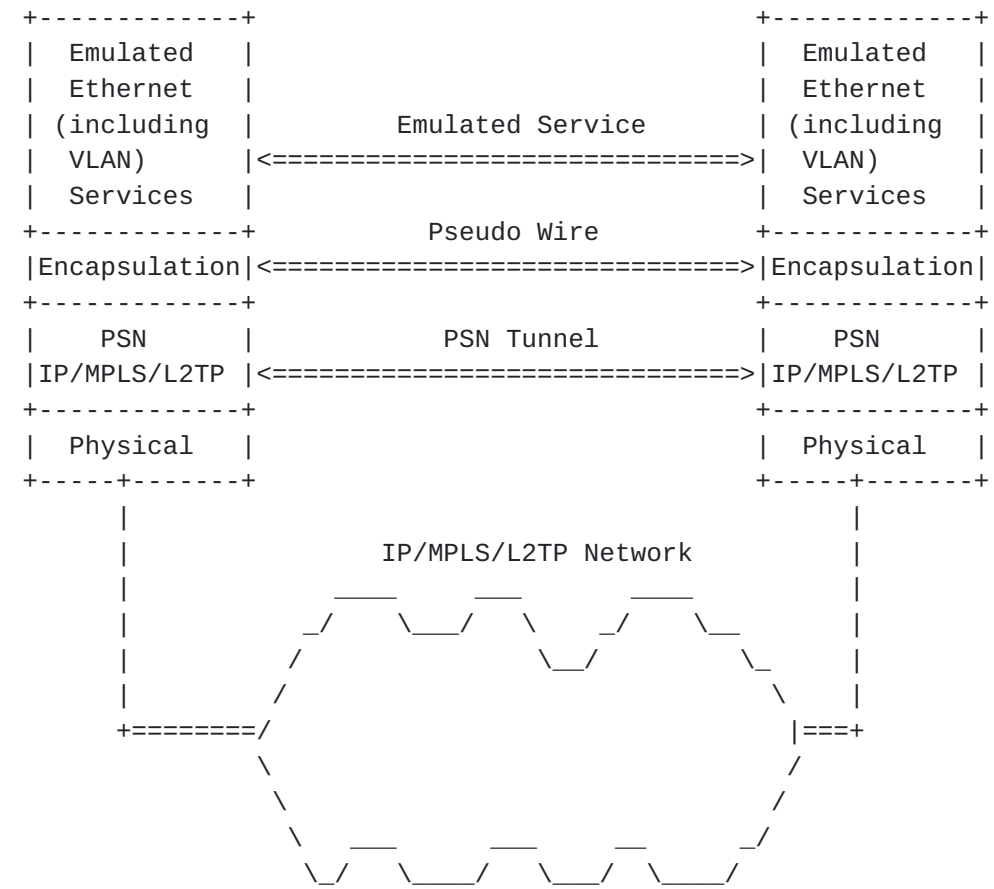


Figure 2: Ethernet PWE3 Protocol Stack Reference Model

2. Terminology

COS mark

The COS field marking at the PSN tunnel level or the VC level. For example: TOS field in IP, EXP bits in MPLS shim.

Multi-point Mode

A PW that contains an internal switching device to support multi-point services, for example TLS.

Point-to-Point Mode

A point-to-point Ethernet PW emulates a single Ethernet link between exactly two endpoints.

3. Requirements For Ethernet Pseudo-Wire Emulation

3.1. Point-to-Point Mode

A point-to-point 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:

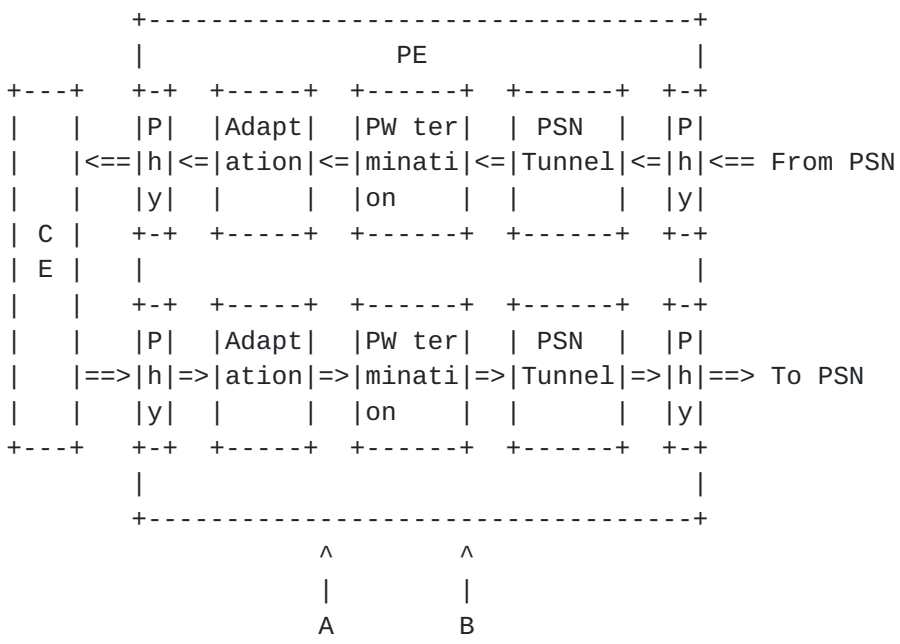


Figure 3: Point-to-point PW reference diagram

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 "Adaptation" 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 VLAN-tag stripping, overwriting or adding, physical port multiplexing/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 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: either IP/GRE, MPLS or L2TP.

A point-to-point pseudo wire can be one of the two types: raw or

tagged. This is a property of virtual Ethernet link and indicates whether the pseudo wire MUST contain an 802.1Q VLAN field (i.e. tagged mode) or may/may not contain a tag (i.e. raw mode).

The rest of this section describes the service at A and the PW Termination behavior that are common to all PSN types. Subsequent sections describe the specific mechanisms unique to each PSN type.

3.2. Multi-point Mode

A multi-point Ethernet PW would emulate a whole Ethernet segment. This segment could be broadcast or switched (like an 802.1D bridge [802.1D]). The reference diagram of section 3.1 still applies, with the following additions:

- There would be more PSN destinations to the right, to represent the additional endpoints of the PW.

- The PW termination function would have to include mechanisms for selecting the correct egress PE(s)(and hence PSN tunnel(s)) for each Ethernet packet presented to the PW. This could involve replication and/or MAC address learning.

As there are alternative mechanisms for providing virtual Ethernet segments using multiple point-to-point Ethernet PWs and a suitable Adaptation function (e.g. see [Vkompella]), the multipoint Ethernet PW is not addressed further in this document.

3.3. Packet Processing

3.3.1. Encapsulation

The entire Ethernet frame without the preamble or FCS is transported as a single packet. PSN-specific tunnel identifiers are prepended to this.

In the multi-point case where the egress PE needs to know which ingress PE forwarded the packet this information must be derived from the PW-specific tunnel identifiers. In the MPLS case this implies that a separate VC label be assigned to each ingress PE. With such consideration, the following implications shall be examined, i.e.

- It implies the use of the global VC label pool per node,

- It may limit the selection of the VC label distribution approach.

In the IP case this information can be derived from the source IP address of the packet. In the L2TP case this can be derived from

the Session ID in the received packet.

3.3.2. MTU Management

Ingress and egress PWESs MUST agree on their maximum MTU size to be transported over the PSN. The consideration of the MTU size management can be referred to Appendix-A. Each PSN-specific PWE

approach will determine if the Segmentation and Reassembly (SAR) will be supported, and if so, what the mechanism should be.

3.3.3. 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 frame ordering is required, a sequence number MUST be implemented and utilized.

The sequence number mechanism is PSN-specificand will be described in the PSN-specific section, if supported.

3.3.4. Frame Error Processing

An encapsulated Ethernet frame traversing a psuedo-wire can 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. Meaning, if a PE-bound device receives a standard Ethernet frame containing hardware level CRC errors, framing errors, or a runt condition, the frame MUST be discarded on input.

3.3.5. 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.

3.4. Maintenance

This section describes the PW link maintenance requirements in a point-to-point configuration.

For the requirements described below, if possible, it is desirable to have a common mechanism (e.g. signaling protocol) to meet those requirement objectives across various PSN types (e.g. MPLS, IP/IP, GRE, L2TP etc.) regardless of the type of maintenance functions such as link establishment or auto-discovery etc. One example is to use MP-BGP [MP-BGP] to auto-discover various PWESs at each PE and to distribute the PSN tunnel label; and to use LDP to distribute the PW's label across each PSN.

3.4.1. Pseudo-wire Establishment

An Ethernet PW can be established over a PSN either via configuration or via some sort of signaling mechanism, e.g. LDP, MP-BGP etc., whichever is applicable to the underlying PSN.

If available, an auto-discovery mechanism, which may be associated with some signaling mechanism (e.g. LDP, MP-BGP) or some server based solution (e.g. DNS), can be used to identify the Ethernet PW types (e.g. native Ethernet or VLAN type) among the PE peers across the PSN. It is expected that when a PW is established between the PEs, the Ethernet PW types are compatible at each end of the PW, i.e. tagged to tagged or raw to raw.

Multiple PWs can be set up within the same PSN tunnel and therefore, the PE/PWES is required to have an ability to support a mechanism to multiplex and de-multiplex the various PW instances. The VC label for an Ethernet PW instance shall be unique at least within the same PSN tunnel so that misrouting of the Ethernet packet can be avoided.

In the case when 802.1p [802.1p] support is required, a COS mapping mechanism (e.g. MPLS EXP field and IP Diffserv DSCP mapping) shall be configured at each PE. The policy of the service mapping between the PW and the PSN tunnel is outside the scope of this specification.

3.4.2. Link Maintenance

It is desirable to detect Ethernet PW failure at the PE which is caused either by the PW itself or by the PSN in a timely manner. Unlike some other transmission technologies, e.g. SONET/SDH, Ethernet does not have a specific standard performance requirement for fault detection and recovery. The requirement on an Ethernet PW is that its reliability performance is identical to standard Ethernet, the performance indicators for this is for further study.

The types of failures detection that have an impact on the Ethernet PW are:

1. PSN tunnel failures
2. VC tunnel failures
3. PWES failures

The detection and diagnostics of these failures requires a co-ordination between the PSN tunnel, the VC-tunnel and the PWES.

When triggering recovery mechanisms for tunnels that carries an Ethernet PW, one must consider the bi-directional nature of the Ethernet PW, and therefore, even if the PSN tunnel is uni-directional, it shall be transparent to the Ethernet PW.

The PW is considered to be active if all the following are true:

1. The local PWES is active.
2. The remote PWES is active.
3. The PSN tunnel used to transport the PW to the remote PE is up.
4. The PSN tunnel used to transport the PW from the remote PE is up.

In order to enable the remote PE to know the status of the local PWES, a PE which is using a maintenance mechanism to establish PWs MUST use its maintenance channel to the remote PE to gracefully withdraw the PW label prior to the local PWES goes down. In the case of manually configured PWs there is no such maintenance channel and thus the remote PE will be unaware of the local PWES status - and must assume it to be active.

The status of the PSN tunnels to and from the remote PE is not always known. PSN-specific considerations are detailed in the relevant sections below.

In the case when there is a high volume of Ethernet PWs across the PSN, a bundling approach can be used to enhance the scalability of the PW link state monitoring and error reporting.

Another aspect of the link state monitoring is to detect mis-routing, i.e. routing traffic from one PW to another PW, due to the corruption of forwarding table. This is especially essential when the PSN tunneling mechanism is connectionless based. Mechanism like Trace Route would be very useful for detecting failure condition.

It is possible that network may provide primary and secondary PSN tunnels to ensure fast recovery. In such case, the expected behavior shall be described of how to perform the failover for the Ethernet PW.

3.5. 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, additional MIB module defines the Ethernet specific parameters required to be configured or monitored. A MIB module for Ethernet service will be available soon.

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

As specified in [PWE3-req][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.

3.6. QoS Consideration

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 exist originally) between edges MUST be preserved, regardless of the COS support at the PSN. In case of adding VLAN field at the edges, a default PRI setting of zero MUST be supported, configured default value is recommended.

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

1. One COS per PW end service, mapped to a single COS PW at the PSN.

2. Multiple COS per PWES mapped to a single PW with multiple COS at the PSN.

3. Multiple COS per PWES mapped to multiple PWs at the PSN.

Examples of the cases above and details of the service mapping consideration 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 in 3.3.5. The mechanism to coordinate the transmission rate between the two PWESs will be discussed in more details in the PSN specific session, if supported.

3.7. Inter-provider PW Support Consideration

In the inter-provider case the requirements above all continue to apply.

3.7.1. PSN tunnel establishment

In the GRE and L2TP cases the PSN tunnel is implicitly formed from a (source, destination) IP address pair (as mentioned above.) For inter-provider operation both these IP addresses SHOULD be globally-unique (i.e. NIC assigned) addresses.

In the MPLS case the PSN tunnel is explicitly signaled, using a label distribution protocol. In order to support inter-domain operation the FEC for the PE SHOULD correspond to a globally-unique address. Furthermore a label distribution protocol suitable for inter-domain operation (e.g MP-BGP) should be used at edge of each autonomous system in the path. A method for establishing these PSN tunnels is given in [MMROZ].

3.7.2. PW establishment

If a signaling mechanism is used to establish the PW then the protocol chosen MUST be suitable for inter-domain operation. Furthermore, the identifier used for the PW SHOULD be globally unique.

3.8. Security Considerations

This document specifies the security consideration regarding the encapsulations and maintenance (signaling) for setting up 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. The signaling extensions as the result of the PW support shall not change nor introduce any security issue related to the existing protocols.

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 mis-configuration 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)is 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.

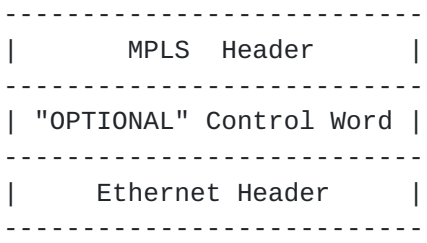
4. Ethernet PW Over MPLS

Ethernet PW packets are encapsulated over an MPLS network using the mechanisms defined in [RFC 3031].

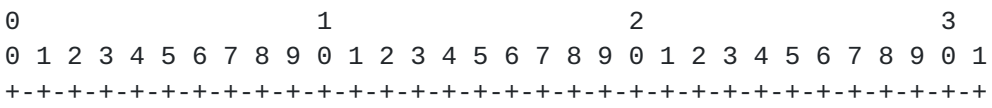
4.1. Packet Processing

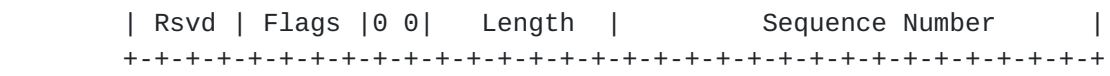
4.1.1. Encapsulation

An Ethernet packet is encapsulated within a MPLS header as shown below is similar to the approach defined in [Martini-encap]:



where: the "Control Word" is defined as follows:




| Rsvd | Flags |0 0| Length | Sequence Number |

The first 4-bit "Rsvd" field is reserved for future use. They MUST be set to 0 when transmitting, and MUST be ignored upon receipt.

The next 4-bit "Flags" field provides space for carrying protocol specific flags.

The next 2-bit MUST be set to 0 when transmitting.

The next 6-bit "Length" field is used as follows:
If the packet's length (defined as the length of the layer-2 payload plus the length of the control word) is less than 64 bytes, the field MUST be set to zero.
The value of the length field, if non-zero, can be used to remove any padding.
When the packet reaches the service providers' egress router, it may be desirable to remove the padding before forwarding the packet.

The next 16-bit provides the 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.

Figure 4: Ethernet PW over MPLS

There are two modes of encapsulation for the Ethernet packets:
Ethernet PW packets are encapsulated in the "Ethernet VLAN" mode for 802.1q VLAN tagged PWs, and
In the "Ethernet" mode for simple Ethernet port-to-port transport of raw PWs regardless the Ethernet packet is tagged or untagged.

4.1.2. Frame Ordering

If frame ordering must be preserved then the "control word" defined above is used. Since the minimum length of an Ethernet frame is 60 octets, and since the control word length field includes the length of the control word itself (4 octets), the length field of theh control word will always be set to zero (as will the reserved and flag bits.)

The default case for Ethernet PW is to operate without a control word.

4.1.2.1. Setting Sequence Number

Between the ingress and egress LERs - R1 and R2 respectively, if R1 supports packet sequencing then the following procedures should be used:

@

The initial packet transmitted on the emulated VC MUST use the 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.

4.1.2.2. Processing Sequence Number

If a router R2 support receive the 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 and the packet sequence number - the expected sequence number < 32768, then the packet is in ordered.
- 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 Management

The network MUST be configured with an MTU that is sufficient to transport the largest encapsulation frames. When using MPLS as the tunneling protocol, there is likely to be 12 or more bytes greater than the largest frame size.

As MPLS does not have the ability to support fragmentation, if an ingress LER determines an encapsulated Ethernet PDU whose payload length exceed the MTU of the MPLS network, the PDU MUST be dropped. If an egress LER receives an encapsulated Ethernet PDU whose payload length (i.e. the length of the Ethernet PDU itself without the MPLS header including the control word, if any), exceeds the MTU of the destination Ethernet interface, the PDU MUST be dropped.

4.3. Maintenance

If the ingress or egress LER detects a failure on the Ethernet logically or physically interface, or the interface is administratively disabled, it MUST withdraw the label mappings for all VCs associated with the interface.

4.3.1. VC Label Distribution

One way to support VC label distribution between the ingress and egress LERs is via the use of LDP extension as described in section 5 of the [Martini-trans]. Other alternative mechanism is for further study.

4.3.2. Link State Monitoring

The PSN tunnel to the remote PE is considered to be up if there is a valid label to reach the remote PE.

If LDP is being used to distribute the PSN tunnel label then there is no way to know if the PSN tunnel from the remote PE is up. However is RSVP-TE or CR-LDP is used to distribute the PSN tunnel label then the status of this tunnel is known.

4.4. Management

The management procedures are as defined above. Note that [PW-MPLS-MIB] defines the mapping from the PW to the MPLS tunnel.

4.5. Security

This draft does not affect the underlying security issues of MPLS (as specified in [MPLS]). Additional security measures MAY be used if the lookup process of the PW will include both PSN label and VC label in case of global VC labels. See [PW-MPLS-MIB] for more

details.

5. Ethernet PW Over IP/GRE

Ethernet PW packets are encapsulated over an IP network using the Generic Routing Encapsulation protocol as specified in [GRE-encap, GRE-IPv4][GRE-revised][GRE-KeyExt].

Note: an alternative method of encapsulating Ethernet PW packets over IP is to use the MPLS encapsulation (see sectioin 4) and an MPLS-in-IP protocol as described in [Rekhter][Worster].

5.1. Packet Processing

5.1.1. Encapsulation

An Ethernet packet is encapsulated into a single IP packet as shown below:

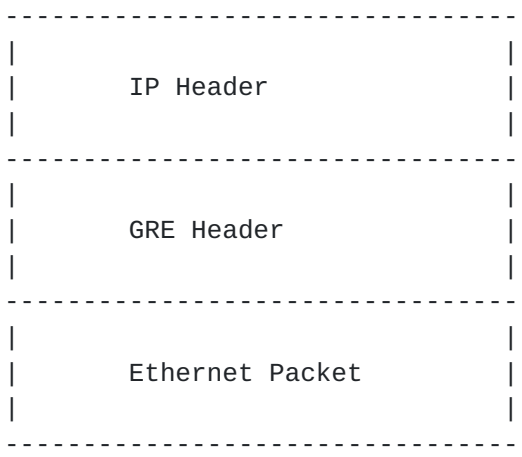


Figure 5: Ethernet Over IP/GRE

The IP header is constructed as follows:

IP Protocol	47h (GRE)
IP Source	Source PE
IP Dest	Dest PE
IP Flags (v4)	DF (don't fragment)
IP DSCP/CoS	Mapped from PW CoS

The Source Address in the IP header can be used to identify the sending PE, if necessary.

The GRE header MUST NOT have the optional Checksum field, MUST contain the optional Key field, and MAY have the optional Sequence Number field, as follows:

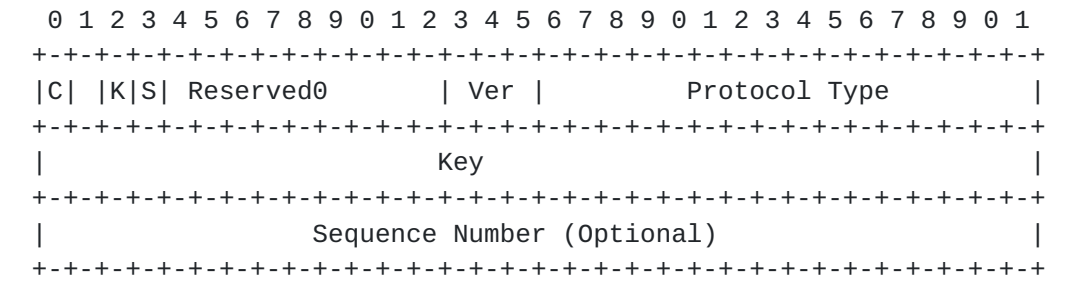


Figure 6: IP/GRE Header

Thus

C = 0
K = 1
S = 0 (sequence number not present)
1 (sequence number present)

The Protocol Type field is set to a number to be allocated by IEEE for this purpose.

The Key field contains the PW label that is assigned by the destination PE. The field is right-aligned and padded with zeros if necessary. The PW label is used as a demultiplexing field to allow multiple PWs between pairs of PEs. The egress PE should assign a PW label that will allow it to determine which PW an arriving packet belongs to.

The Sequence Number field (if present) is used as described in RFC2890 to maintain or to verify packet ordering within a particular PW.

5.1.2. Frame Ordering

The normative statements in RFC2890 apply as written to the sending and receiving PEs. In particular a receiving PE MUST correctly parse a GRE packet containing a sequence number, even if it is unable to provide sequencing.

5.1.3. MTU Management

Techniques such as Path MTU Discovery [MTU] may be used to determine the MTU of the IP/GRE Tunnel. Alternatively the MTU may be statically configured, or configured per destination PE.

The sending PE MUST NOT fragment an IPv4 packet containing an Ethernet PW PDU, and MUST set the DF bit in the IPv4 header.

5.2. Maintenance

Any Ethernet PW maintenance protocol that allows the distribution

of PW labels may be used. Other generic attributes that may be validated include MTU, sequencing preference, trunking mode, etc. The specific maintenance protocols and procedures are not defined

here.

IP protocols such as ICMP ping may be used to verify PW connectivity for all the PWs between a pair of PEs.

5.2.1. Link State Monitoring

The PSN tunnel to the remote PE is considered to be up if there is a valid route to reach the remote PE. There is, however, no way to determine if the PSN tunnel from the remote PE is up.

5.3. Management

Generic management procedures apply.

5.4. Security

The nature of the IP/GRE encapsulation means that it would be relatively easy for an external intruder to spoof packets that appeared to belong to a particular PW. The receiving PE SHOULD verify that the source PE IP address corresponds to the expected source IP address for the PW, and filtering of source-spoofed packets from outside a trusted domain may be necessary.

Another issue is the ability of IP/GRE PW packets to escape from a trusted domain due to transient routing changes/errors or an attack on the routing protocols themselves. To protect this it may be necessary to install filters to prevent IP/GRE Ethernet PW packets from leaving the domain.

Additionally or alternatively, IP security procedures such as IPSec may be used to further enhance the security of all PWs between a pair of PEs. This most likely be a requirement for inter-domain PWs.

5.5. QoS Consideration

The IP Diffserv model is used to provide differential class of service to different PWs, or to different packets within the PW. The mapping of Ethernet CoS markings to/from Diffserv codepoints is a local configuration matter, but must follow the requirements in section 3.7.

6. Ethernet PW Over L2TP

This section describes how to provide Ethernet PWE over L2TPv3.

[L2TP] was originally designed for tunneling PPP sessions. [L2TPv3] separates out mechanisms designed specifically for PPP and provides

new extensions for tunneling generic layer-2 protocols such as Ethernet, ATM and Frame Relay.

To provide Ethernet PWE between two PEs, an L2TPv3 control connection can be established first. Individual L2TPv3 sessions can then be established via signaling. Alternatively, L2TPv3 sessions can be manually configured without requiring a control connection. Each session can be used as a PW to connect two Ethernet ports or VLANs.

6.1. Packet Processing

6.1.1. Encapsulation

The entire Ethernet frame without the preamble or FCS is encapsulated in L2TPv3 and is sent as a single packet. This is done regardless of whether an 802.1Q tag is present in the Ethernet frame or not.

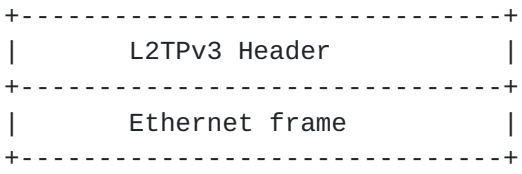


Figure 7: Ethernet over L2TPv3

An L2TPv3 data packet can be sent as over UDP or directly over IP. The selection of UDP vs. IP is beyond scope of this document.

L2TPv3 data channels do not provide reliable or in-order delivery. There is no sequence number field in an L2TPv3 header. If in-order delivery for Ethernet frames is desired, an optional 4-octet control word can be inserted between the L2TPv3 header and the encapsulated Ethernet frame. The format of the control word is identical to the control word defined in [Martini-encap]. The usage of the control word fields is identical to what is defined in [Martini-encap], except that the length field MUST be set to zero at the ingress PE and be ignored at the egress PE. The presence/non-presence of the control word for a particular PW session is signaled during setup of the PW. The signaling process is described in Section 6.2.

After the encapsulation, the whole packet is as follows:

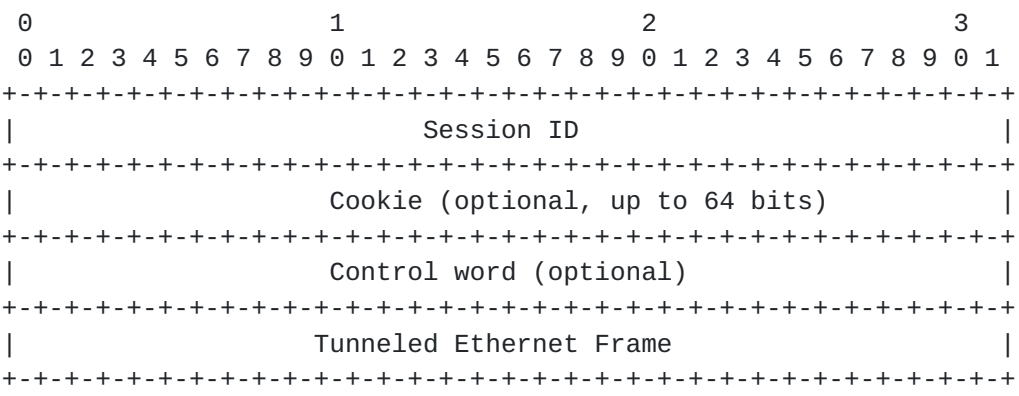


Figure 8: Encapsulation of Ethernet frames over L2TPv3

A session ID uniquely identifies a PW. It is used to multiplex and de-multiplex PWs between two PEs. Session IDs only have local significance. That is, the same PW will be given different Session IDs by each PE. The Session ID specified in each message is that of the intended recipient, not the sender [L2TPv3].

A cookie field is used to check the association of a received data packet with the PW identified by the Session ID. The cookie guards against the misrouting of data packets, which could result if the incorrect Session ID is specified in received packets (due to mis-configuration, header corruption, or otherwise) [L2TPv3].

6.1.2. Frame Ordering

In cases where in-order delivery of Ethernet frames is critical, the control word can be used. The sequence number field in the control word can be used to detect out-of-order delivery. The generation and processing of the sequence number at the ingress and egress PEs, respectively, are identical to what are defined in [Martini-encap]. The presence of a control word is signaled during setup of the L2TPv3 session for this PW. The signaling process is described in Section 6.2.

6.1.3. MTU Handling

With L2TPv3 as the tunneling protocol, the packet resulted from the encapsulation is N bytes longer than Ethernet frame without the preamble or FCS, where

N=8, without a control word and L2TPv3 data messages are over IP;
N=12, with a control word and L2TPv3 data messages are over IP;
N=16, without a control word and L2TPv3 data messages are over UDP;
N=20, with a control word and L2TPv3 data messages are over UDP;
(N does not include the IP header).

In order to avoid fragmentation, ideally the PSN should be configured with an MTU that is larger than or equal to the largest

Ethernet frame size (without the preamble or FCS) plus 20 bytes. If the PSN cannot support such a MTU, another option is to set the MTU size of the two Ethernet ports between the PEs and the CEs to (network_MTU - 20). This may imply that Ethernet jumbo frame cannot be used.

If the PSN cannot be configured with a sufficiently large MTU to avoid fragmentation, Ethernet PWE over L2TPv3 can rely on IP fragmentation.

6.2. Maintenance

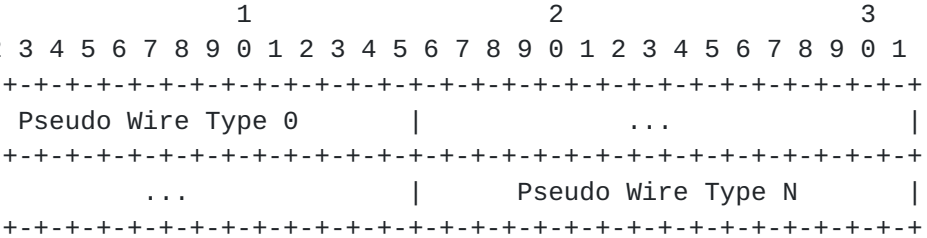
6.2.1. Pseudo-wire Establishment

With L2TPv3 as the tunneling protocol, Ethernet PWs are L2TPv3 sessions. There are two ways to set up L2TPv3 sessions:

(1) Manual configuration;
(2) Establishing an L2TPv3 control connection first and then establishing of individual sessions via signaling. The procedure is defined in [L2TPv3]. In order for an L2TPv3 control connection to support Ethernet PWs, it must be signaled to support Ethernet VLAN and Ethernet ports. This is done using the "Pseudo-wire capability list" Attribute-Value Pair (AVP).

6.2.1.1. Control Connection Establishment

If a control connection is to be established, the possible types of PW sessions associated with this control connection MUST be negotiated first. This is done using the Pseudo Wire Capabilities List AVP that indicates the L2 payload types that will be accepted by the PE that originates this control message. The Attribute Value field for this AVP has the following format:



Defined Pseudo Wire Types that may be included in the Pseudo Wire Capabilities List are as follows (pending IANA approval):

Legal "Pseudo Wire Types" that may be included in the Pseudo Wire Capabilities List are defined below (pending IANA approval):

0x0004 - Sessions without control word for connecting Ethernet

VLANs are allowed

0x0005 - Sessions without control word for connecting Ethernet ports are allowed

0x8004 - Sessions with control word for connecting Ethernet VLANs are allowed
0x8005 - Sessions with control word for connecting Ethernet ports are allowed

Note that the most significant bit of the "Pseudo-Wire Type" field is used to indicate the presence/non-presence of a control word in a PW session. If the bit is set, a control word is present. Otherwise, it is not.

6.2.1.2. PW session establishment

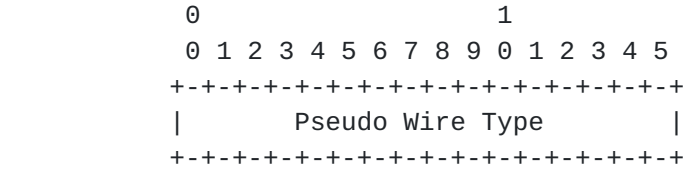
Pieces of information needed for each PW session are described below. Such information is either manually configured at the ingress and egress PEs, or dynamically signaled with L2TPv3 AVPs.

- Pseudo Wire Type

The type of a PW can be either "Ethernet port" or "Ethernet VLAN".

If signaling is used, the "Pseudo Wire Type" AVP, Attribute Type TBA, indicates the payload type for a PW.

The Attribute Value field for this AVP has the following format:



"Pseudo wire type" values are defined in Section 6.2.1.1.

A PE MUST NOT request to set up a PW with a "Pseudo wire type" AVP specifying a value not advertised in the "Pseudo Wire Capabilities List" AVP it received during control connection establishment. Attempts to do so will result in the failure of PW setup.

- Presence/non-presence of control word

If the presence/non-presence of control word for a PW session is to be signaled, then:

- If the two Pseudo-Wire End Services (PWES's) are Ethernet VLANs, the presence/non-presence of control word can be signaled by using the value 0x0004 or 0x8004, respectively.

- If the two PWES's are Ethernet ports, the presence/non-

presence of control word can be signaled by using the value 0x0005 or 0x8005, respectively.

That is, the most significant bit, i.e. the C bit, of the "Pseudo-Wire Type" field is used to indicate the presence/non-presence of

the control word in a PW. If the bit is set, the control word is present. Otherwise, it is not.

- PW ID

Each PW is associated with a PW ID. The two PEs of a PW have the same PW ID for it. Together with the Pseudo-Wire Type, a PW ID uniquely identifies a PW session at every PE. A new L2TPv3 AVP will be defined for signaling the PW ID.

- Group ID

A Group ID is used for referring to a group of PWs so that they can be signaled down collectively. The L2TPv3 "Private Group ID" AVP can be used for signaling the Group ID.

- PWES parameters

Three parameters are defined for each Ethernet PWES. They can be used for detecting mismatch between the two PWES's of a PW.

+ Description string

This is an informational description string for a PWES. For example, if the local PWES is VLAN 100 of interface GE 1/1 on PE1, then the description string of the PW can be "PE1-GE1/1-VLAN100")

A new L2TPv3 AVP will be defined for signaling this description string.

+ MTU

This parameter specifies the MTU size of the local PWES. It can be used for detecting MTU mismatch of the two PWES's of a PW. MTU mismatch SHOULD be logged if identified.

A new L2TPv3 AVP will be defined for signaling the MTU of a PWES.

+ Speed

This parameter specifies the Send and Receive speed of the PWES. That is, speed of an Ethernet PW is assumed to be symmetric. The speed of a PWES should not be higher than the speed of the physical port. The speed specification is mainly for informational purpose, e.g., for detecting speed mismatch of the two PWES's. An example of speed mismatch is: one PWES, a VLAN, is specified to have speed 20Mbps (possibly via rate-limiting) and the other PWES, another VLAN, is specified to

have speed 40Mbps. Speed mismatch SHOULD be logged if identified.

The L2TPv3 "Connect Speed" AVP can be used for signaling the speed of a PWES.

6.2.2. PW Status Monitoring

The working status of a PW is reflected by the state of the L2TPv3 session. If the corresponding L2TPv3 session is down, both PWES's associated with it MUST be shut down.

If a control connection is used, and the control channel and the data channels operate in-band, the keep-alive mechanism of L2TPv3 can serve as a link status monitoring mechanism for the PWs (i.e. sessions) associated with that control connection. If the control channel and the data channels operate out-of-band, an L2TPv3 data session may be dedicated for sending keep-alive information [Editor's note: details will be provided in the next version of the draft].

If one of the PWES is down, Ethernet PWE MUST treat it as an L2TPv3 "local close request" and tear down the PW associated with that PWES. When the remote PE cleans the state for the PW, it MUST shut down the PWES associated with it.

6.2.3. Fault Detection & Recovery

An Ethernet PW can incur loss, corruption, and out-of-order delivery of data packets. Packet loss, corruption, and out-of-order delivery can be considered as "generalized packet error" of an Ethernet PW. If the "generalized packet error" rate is higher than a configurable threshold, the PW MUST be signaled down with this reason explained in the "Result Code" AVP. The two PWES's MUST also be shut down.

6.3. Management

An Ethernet pseudo-wire emulation MIB will be defined in a companion draft.

6.4. Security

Ethernet pseudo-wire emulation does not affect the underlying security issues of L2TPv3 [Section 9, L2TPv3].

6.5. QoS Consideration

L2TPv3 provides reliable delivery for control messages. This reliable delivery mechanism is provided at the two L2TPv3 endpoints (i.e., L2TPv3 Control Connection Endpoint, or LCCE). More specifically, this is done by:

1.

Each L2TPv3 will acknowledge receipt of a control message;

2. If the sender of a control message does not receive an acknowledgement, it will retransmit.

This reliable delivery mechanism does not rely on any QoS mechanism of the PSN.

There is no reliable delivery mechanism for data messages.

By default, the control and data messages will receive best effort service inside the PSN. It is possible to use DiffServ and other traffic management mechanisms to provide better service quality to these messages. It is also possible to provide better service quality to the control messages than to the data message. What service quality to provide inside the PSN for the control messages and data messages depends on the domain policy of the PSN and is outside scope of this document.

7. Security Considerations

To Be Completed

8. Conclusion

To Be Completed

9. IANA Consideration

This section defined four Pseudo-Wire Types. The specific values used for these types are pending IANA approval. The PWE3 WG needs to work with the L2TP WG to agree on these numbers as well.

0x0004 - Ethernet VLAN, without a control word
0x0005 - Ethernet port, without a control word
0x8004 - Ethernet VLAN, with a control word
0x8005 - Ethernet port, with a control word

10. References

IETF RFC

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

[GRE-encap] Hanks, S., Li, T., Farinacci, D., "Generic Routing Encapsulation (GRE)", RFC 1701, October 1994.

[GRE-IPv4] Hanks, S., Li, T., Farinacci, D., Traina, P., "Generic Routing Encapsulation over IPv4 networks", RFC 1702, October 1994.

[GRE-KeyExt] Dommety, G., "Key and Sequence Number Extensions to GRE", RFC 2890, September 2000.

[GRE-Revised] Farinacci, D., Li, T., Hanks, S., Meyer, D., Traina, P., "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

[L2TP] Townsley, W., Valencia, A., Rubens, A., Singh Pall, G., Zorn, G., Palter, B., "Layer Two Tunneling Protocol (L2TP)", RFC 2661 August 1999

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

[MPLS] Rosen, E., Viswanathan, A., Callon, R., "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[MTU] Mogul, J., Deering, S., "Path MTU Discovery", RFC 1191, November 1990.

IETF Drafts

[ATM] T.B.D.

[CEM] Pate, P., Cohen, R., Zelig, D., "TDM Service Specification for Pseudo-Wire Emulation Edge-to-edge (PWE3)", (draft-pate-pwe3-tdm-00.txt), work in progress, March 2002.

[FR] Kawa, C., Malis, A., Pate, P., Bhat, R., Vasavada, N., "Frame relay over Pseudo-Wire Emulation Edge-to-Edge", (draft-kamapabhava-fr-pwe3-00.txt), work in progress, March 2002.

[Heron] Heron, G., Wilder, R., Heinanen, J., Soon, T., Martini, L., Kompella, V., Regan, J., Khandekar, S., "Requirements for Virtual Private Switched Networks", (draft-heron-ppvpn-vpsn-reqmts-00.txt), work in progress, July 2001.

[Kompella] Kompella, K., Leelanivas, M., Vohra, Q., Bonica, R., Metz, E., Ould-Brahim, H., Achirica, J., Liljenstolpe, C., Sargor, C., Srinivasan, V., Zhang, Z., "MPLS-based Layer 2 VPNs", (draft-kompella-ppvpn-l2vpn-00.txt), work in progress, July 2001.

[L2TPv3] Lau, J., Townsley, M., Valencia, A., Zorn, G., Goyret, I., Pall, G., Rubens, A., Palter, B., "Layer

Two Tunneling Protocol "L2TP"", (draft-ietf-l2tpext-l2tp-base-01.txt), work in progress, July 2001.

[Martini-encap]

Martini, L., El-Aawar, N., Tappan, D., Rosen, E., Jayakumar, J., Vlachos, D., Liljenstolpe, C., Heron, G., Kompella, K., Vogelsang, S., Shirron, J., Smith, T., Radoaca, V., Malis, A., Sirkay, V., Cooper, D., "Encapsulation Methods for Transport of Layer 2 Frames Over IP and MPLS Networks", (draft-martini-l2circuit-encap-mpls-03.txt), work in progress, July 2001.

[Martini-trans]

Martini, L., El-Aawar, N., Tappen, D., Rosen, E., Hamilton, A., Jayakumar, J., Vlachos, D., Liljenstolpe, C., Heron, G., Kompella, K., Vogelsang, S., Shirron, J., Smith, T., Radoaca, V., Malis, A., Sirkay, V., Cooper, D., "Transport of Layer 2 Frames Over MPLS", (draft-martini-l2circuit-trans-mpls-08.txt), work in progress, July 2001.

[MMROZ] M. Mroz, O. Stokes, V. Kanagasabapthy, V. Bhagavath, G. Heron, P. Lin, Y. Serbest, "Tunnel LSPs Extended Across Autonomous System Boundaries", (draft-mroz-ppvpn-inter-as-lsps-00.txt), work in progress, February 2002.

[PWE3-frame] Pate, P., Xiao, X., So, T., Malis, A., Nadeau, T., White, C., Kompella, K., Johnson, T., "Framework for Pseudo Wire Emulation Edge-to-Edge (PWE3)" (draft-pate-pwe3-framework-02.txt), work in progress, July 2001.

[PW-MIB] Zelig, D., Mantin, S., Nadeau, T., Danenbert, D., Malis, A., "Pseudo Wire (PW) Management Information Base Using SMIv2", (draft-zelig-pw-mib-00.txt), work in progress, July 2001.

[PW-MPLS-MIB] Danenberg, D., Park, S., Nadeau, T., Zelig, D., Malis, A., , "SONET/SDH Circuit Emulation Service Over MPLS (CEM) Management Information Base Using SMIv2", (draft-danenberg-pw-cem-mib-00.txt), work in progress, July 2001.

[Rekhter] Rekhter, Y., Tappen, D., Rosen, E., " MPLS Label Stack Encapsulation in GRE", (draft-rekhter-mpls-over-gre-03.txt), work in progress, February 2002.

[Rosen] Rosen, E., Filsfils, C., Malis, A., Vogelsang, S., Heron, G., Martini, L., "An Architecture for L2VPNs", (draft-ietf-ppvpn-l2vpn-00.txt), work in progress, July 2001.

[Vkompella] Kompella, V., Khandekar, S., Heron, G., Heinanen, J., Soon, T., Wilder, R., Martini, L., "Requirements for Virtual Private Switched Networks", (draft-heron-ppvpn-vpsn-reqmts-00.txt), work in progress, July 2001.

[Worster] Worster et al, "MPLS Label Stack Encapsulation in IP", (draft-worster-mpls-in-ip-05.txt), work in progress, February 2002.

[PWE3-req] Xiao, X., McPherson, D., Pate, P., White, C., Kompella, K., Gill, V., Nadeau, T., "Requirements for Pseudo Wire Emulation Edge-to-Edge (PWE3)" (draft-pwe3-requirements-01.txt), work in progress, July 2001.

IEEE

[802.1D] IEEE, "ISO/IEC 15802-3:1998,(802.1D, 1998 Edition), Information technology --Telecommunications and information exchange between systems --IEEE standard for local and metropolitan area networks --Common specifications-Media access control (MAC) Bridges", June, 1998.

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

[802.3] IEEE, "ISO/IEC 8802-3: 2000 (E), 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.

11. Authors' Addresses

Tricci So
Caspian Networks
170 Baytech Drive
San Jose, CA, USA 95134
Email:
tso@caspiannetworks.com

XiPeng Xiao
Email: xiaoxipe@cse.msu.edu

Giles Heron
PacketExchange Ltd.

Chris Flores
Austin, Texas

The Truman Brewery
91 Brick Lane
LONDON E1 6QL,

Email:
chris_flores@hotmail.com

United Kingdom
Email:
giles@packetexchange.net

David Zelig
Corrigent Systems
126, Yigal Alon st.
Tel Aviv, ISRAEL
Email: davidz@corrigent.com

Raj Sharma
Luminous Networks, Inc.,
10460 Bubb Road
Cupertino, CA 95014
Email: raj@luminous.com

Nick Tingle
TiMetra Networks
274 Ferguson Drive
Mountain View, CA, USA 94043
Email: nick@timetra.com

Sunil Khandekar
TiMetra Networks
274 Ferguson Drive
Mountain View, CA, USA
94043
Email: sunil@timetra.com

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

Appendix A - Interoperability Guidelines

Point to point services.

The following is a list of the configuration options for a point to point service, 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		(note 1)
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 5)
4) Tag	No Tag	Remove/add Tag field	(note 4)
5) Tag1-Tag2	Tag1-Tag2		VLAN can be 0-4095 Change of VLAN is not allowed

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:

1) This mode is equivalent to port mode in [Martini-trans] since any packet on the physical port is transmitted as is on the PW and vice versa.

2) The VLAN mode in [Martini-trans] is an example of service #2.

According to the default specification, it does not change the VLAN field.

3) In draft-martini any change of the VLAN tag is done at the PW CE-bound,in order to support equipment that cannot change the VLAN tag at the PW PE-bound. However, where possible, it is recommended to change the VLAN tag at the PW PE-bound, for compatibility with VPLS service requirements(see further details below).

4) Mode #4 exists in layer 2 switches, but is not allowed when operating with PW since it does not preserve the user's PRI bit, and in order to save configuration of additional service that can be achieved by other set of configuration. 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.

5) Mode #3 can 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.

The use of PW for a TLS service is shown in the following diagram:

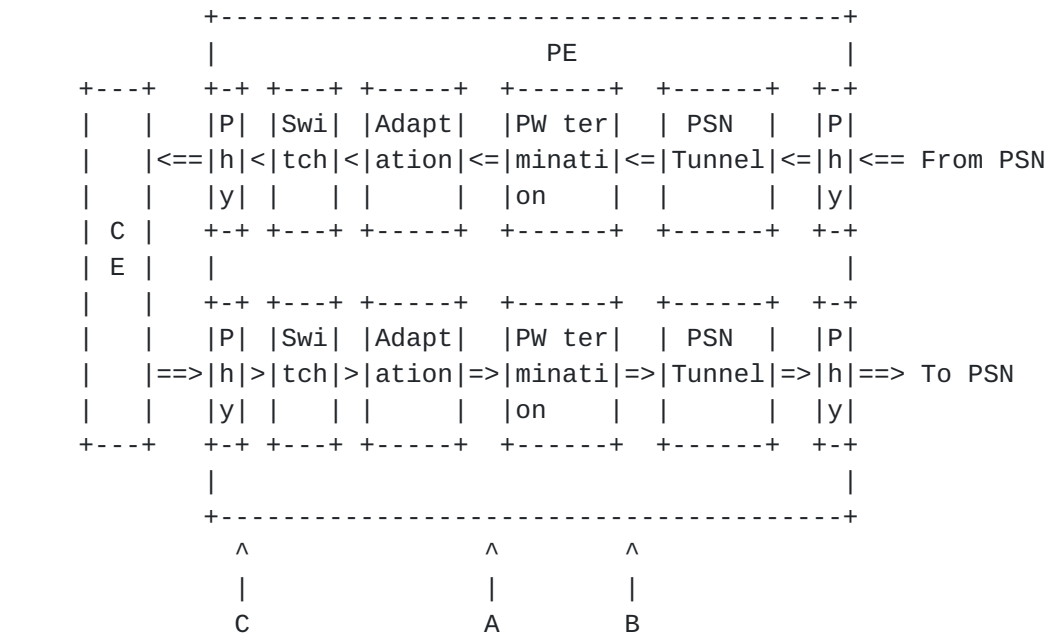


Figure 9: Point-to-point PW reference diagram

Switching (TLS) service (i.e. VPSN) and the allowed relations to the point to point encapsulations format:

It is assumed that the switching operation requires that the switch ports (see figure 2) will conform to the requirement of 802.1D, i.e. switching and learning is based on VLAN field on the interfaces. Packets without VLAN field or with VLAN NULL may be associated to a VLAN # on a per port/PW virtual interface basis.

Not all virtual interfaces of the same TLS instance may have the same VLAN values supported on them, i.e. the forwarding table shall be based on VLAN.

In order to support HUB and Spoke topology where the PE at the spoke cannot change the VLAN field in order to comply to 802.1D rules, a change of VLAN value may be needed at the adaptation process before the switching operation.

In most cases, port (A) can be defined as "RAW" and destination of packets may be selected based on the VLAN configuration on the PWs. However, if more than one switching instance is required for the same port, (A) MUST NOT be defined as "RAW" service, and conflicting VLAN ranges between the switching instances cannot be configured in this case.

Remarks:

1) Each PW may have different set of VLANs associated with. This enables to view the TLS network exactly the same as Enterprise switch.

2) Mode #2 with change of VLAN value is allowed for one VLAN only per PW. Recommended new value is 0 (NULL VLAN) on 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.

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.

MTU Coordination Considerations

All nodes comprising the PSN shall be configured such that their MTU is greater-than or equal-to the largest Ethernet frame plus PSN tunnel header. If MPLS is utilized as the tunneling mechanism, for example, assuming that there is no label stacking, 8 octets will be typically be added to the largest Ethernet frame size (4 octets for the tunnel label and 4 for the VC label) - creating the encapsulated Ethernet frame size. However, other tunneling mechanisms (i.e. L2TP, IP/GRE) may have longer headers and require

larger MTUs.

Appendix B - QOS details.

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

1) 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.

2) 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.

3) 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 o different L-LSPs in MPLS PSN in order to support multiple COS service over L-LSP capable network.

See the PSN specific sections for supported functionality for different PSN technologies.

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

1. 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 principals 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	0	0	0	1	1	2	2	3

Effort								
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 10: IEEE 802.1Q COS Service Mapping

2. 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.

3. 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 in MPLS), the rules stated above for tunnel COS marking apply also for 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.