

Network Working Group
Internet Draft
Expiration Date: August 2003

Luca Martini
Nasser El-Aawar
Level 3 Communications, LLC.

Giles Heron
PacketExchange Ltd.

Eric C. Rosen
Cisco Systems, Inc.

February 2003

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

[draft-ietf-pwe3-ethernet-encap-02.txt](http://www.ietf.org/drafts/pwe3-ethernet-encap-02.txt)

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of [Section 10 of RFC2026](#).

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

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

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/1id-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

Abstract

An Ethernet Pseudowire (PW) is used to carry Ethernet/802.3 Protocol Data Units over an IP or MPLS network. This enables service providers to offer "emulated" ethernet services over existing IP or MPLS networks. This document specifies the encapsulation of Ethernet/802.3 PDUs within a pseudowire. It also specifies the procedures for using a PW to provide a "point-to-point ethernet" service.

Internet Draft [draft-ietf-pwe3-ethernet-encap-02.txt](#) February 2003

Table of Contents

1	Specification of Requirements	2
2	Introduction	2
3	Requirements for Ethernet Pseudo-Wires Emulating Point-to-Point Eth	
3.1	Frame Processing at the PW Endpoints	6
3.1.1	Encapsulation	6
3.1.2	Tagged Mode	6
3.1.3	MTU Management on the PE/CE Links	6
3.1.4	Frame Ordering	6
3.1.5	Frame Error Processing	7
3.1.6	IEEE 802.3x Flow Control Interworking	7
3.2	PW Setup and Maintenance	7
3.3	Management	7
3.4	The Control Word	8
3.4.1	Setting the sequence number	8
3.4.2	Processing the sequence number	9
3.5	QoS Considerations	9
3.6	Security Considerations	10
3.7	PSN MTU Requirements	10
4	Intellectual Property Disclaimer	11
5	References	11
6	Author Information	12
	Appendix A - Interoperability Guidelines	15
	Appendix B - QoS Details	16

[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

An Ethernet Pseudowire (PW) allows Ethernet/802.3 Protocol Data Units (PDUs) to be carried over an IP network or an MPLS network. In

addressing the issues associated with carrying an Ethernet PDU over a PSN, this document assumes that a Pseudowire (PW) has been set up by some means outside the scope of this document. This may be via manual configuration, or a signaling protocol such as that defined in [PWE3-CTRL] or [L2TPv3]. As described in [PWE3-FRAME], this PW may be

tunneled through an MPLS, IPv4 or IPv6 PSN.

In addition to the Ethernet PDU format used within the pseudowire, this document discusses:

- Procedures for using a PW in order to provide a pair of CEs with an emulated (point-to-point) ethernet service, including the procedures for the processing of PE-bound and CE-bound ethernet PDUs. [PWE3-FRAME]
- Ethernet-specific QoS and security considerations
- Inter-domain transport considerations for Ethernet PW

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

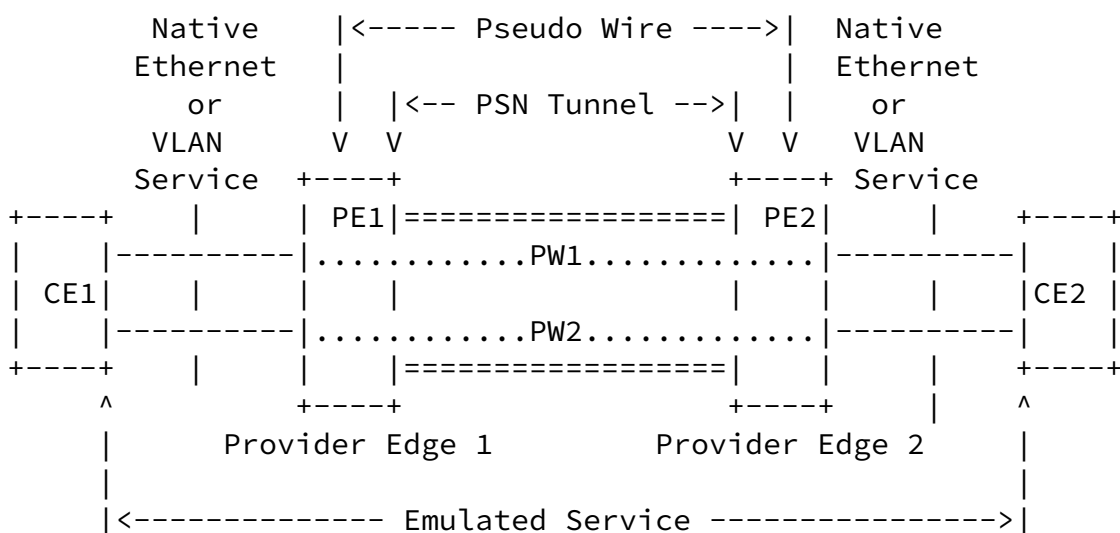


Figure 1: PWE3 Ethernet/VLAN Interface Reference Configuration

The "emulated service" shown in Figure 1 is, strictly speaking, a

bridged LAN; the PEs have MAC interfaces, consume MAC control frames, etc. However, the procedures specified herein only support the case in which there are two CEs on the "emulated LAN". Hence we refer to this service as "emulated point-to-point ethernet". Specification of the procedures for using pseudowires to emulate LANs with more than two CEs are out of scope of the current document.

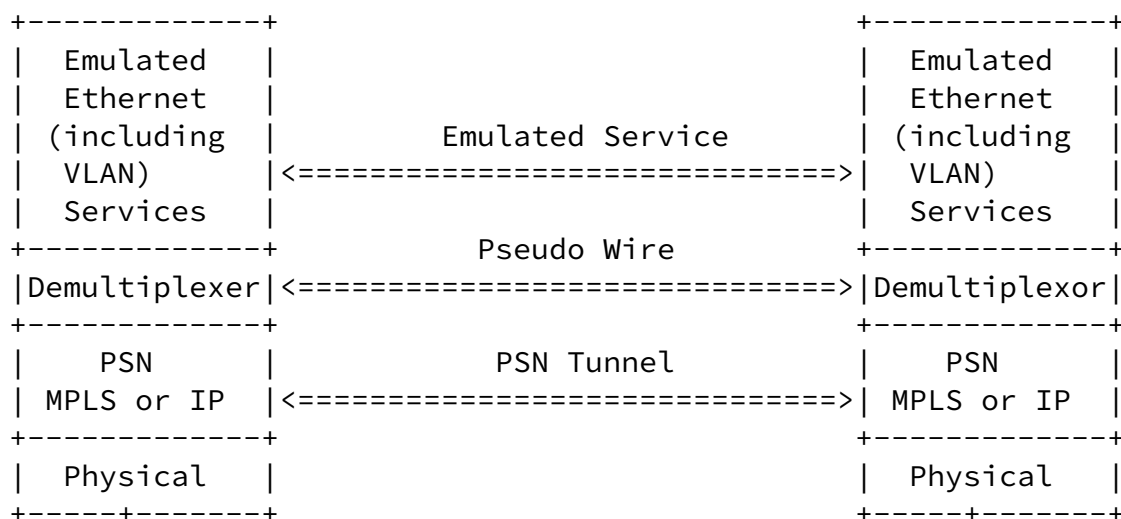


Figure 2: Ethernet PWE3 Protocol Stack Reference Model

For the purpose of this document, PE1 will be defined as the ingress router, and PE2 as the egress router. A layer 2 PDU will be received at PE1, encapsulated at PE1, transported, decapsulated at PE2, and transmitted out of PE2.

3. Requirements for Ethernet Pseudo-Wires Emulating Point-to-Point Ethernet Links

An Ethernet PW emulates a single Ethernet link between exactly two endpoints. The mechanisms described in this document are agnostic to that which is beneath the "Pseudo Wire" level in Figure 2, concerning

itself only with the "Emulated Service" portion of the stack.

The following reference model describes the termination point of each end of the PW within the PE:

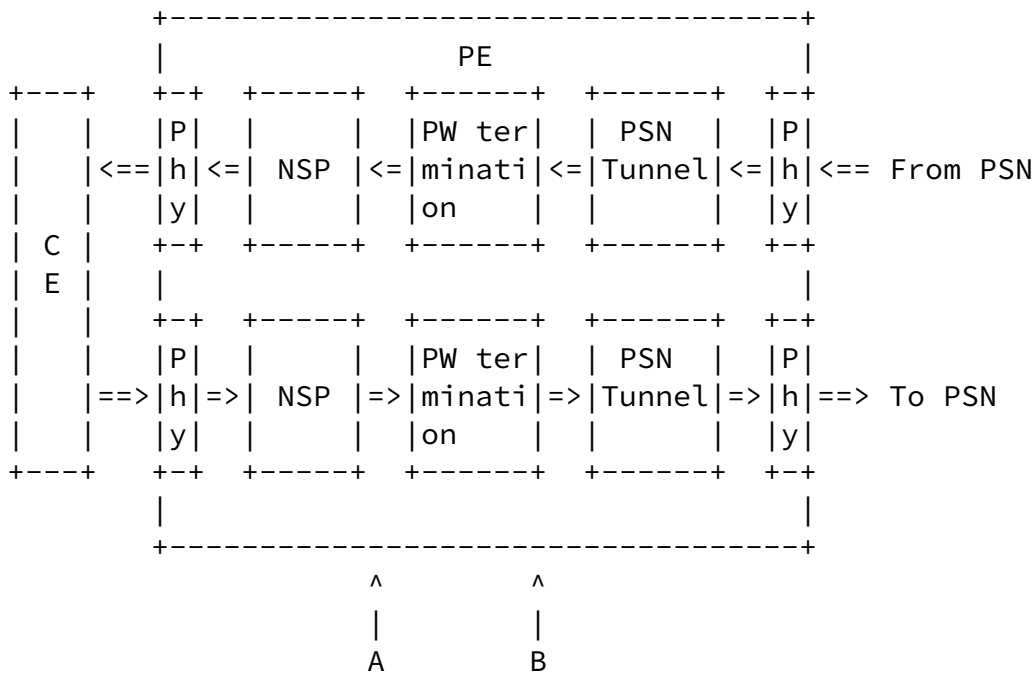


Figure 3: PW reference diagram

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 frame 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 frame processing that is required for the Ethernet frames that are forwarded 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 frames according to the PSN type in use.

An ethernet PW can operate in one of two modes: "raw mode" or "tagged mode". In tagged mode, each frame MUST contain an 802.1Q VLAN tag, and the tag value is meaningful to the NSPs at the two PW endpoints. That is, the two endpoints must have some agreement (signaled or manually configured) on how to process the tag. On a raw mode PW, a frame MAY contain an 802.1Q VLAN tag, but if it does, the tag is not

meaningful to the NSPs, and passes transparently through them.

[3.1](#). Frame Processing at the PW Endpoints

[3.1.1](#). Encapsulation

The entire Ethernet frame without any preamble or FCS is transported as a single frame over the Pseudowire. Note that when using the signaling procedures defined in [[PWE3-CRTL](#)] or [[L2TPv3](#)], a "Raw Mode" PW should be signaled as being of type "Ethernet".

[3.1.2](#). Tagged Mode

The ethernet frame may contain an 802.1Q tag, in this case the PE MUST signal that the PW will be transporting ethernet frames including 802.1Q tags. In this case 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 [[PWE3-CRTL](#)] or [[L2TPv3](#)], a "Tagged Mode" PW should be signaled as being of type "Ethernet VLAN".

[3.1.3.](#) MTU Management on the PE/CE Links

The Ethernet PW MUST NOT be enabled unless it is known that the MTUs of the CE-PE links are the same at both ends of the PW.

[3.1.4.](#) 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. An ethernet PW can, through use of the control word, provide strict frame ordering. If this option is enabled, any frames which get misordered by the PSN will be dropped by the receiving PW endpoint. If strict frame ordering is a requirement for a particular PW, this option MUST be enabled.

[3.1.5.](#) Frame Error Processing

An encapsulated Ethernet frame traversing a psuedo-wire may be dropped, corrupted or delivered out-of-order. As described in [[PWE3-REQ](#)], frame-loss, corruption, and out-of-order delivery is considered to be a "generalized bit error" of the psuedo-wire. PW frames that are corrupted will be detected at the PSN layer and dropped.

At the ingress of the PW the native Ethernet frame error processing mechanisms MUST be enabled. 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 defining this processing is part of the NSP function and is outside the scope of this draft.

[3.1.6. IEEE 802.3x Flow Control Interworking](#)

In a standard gigabit 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. PW Setup and Maintenance](#)

This document assumes that a mechanism exists to set up the ethernet PW. Maintenance of the PW (e.g. keepalives, status updates, etc) is generally tied closely to the PW Setup mechanisms. [PWE3-CTRL] and [L2TPv3] define two mechanisms for setup and maintenance of Ethernet PWs.

[3.3. Management](#)

The Ethernet PW management model follows the general management defined in [PWE3-FRAME] and [PWE3-MIB]. Many common PW management facilities are provided here, with no additional Ethernet specifics necessary. Ethernet-specific parameters are defined in an additional MIB module, [PW-ENET-MIB].

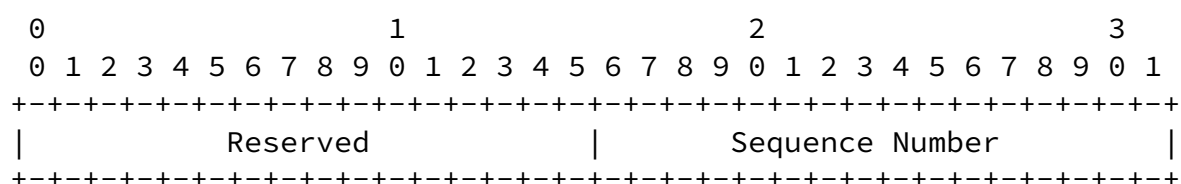
As specified in [PWE3-FRAME], an implementation SHOULD support the generic and specific PW MIB modules for PW set-up and monitoring. Other mechanisms for PW set up (command line interface for example) MAY be supported.

[3.4. 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 [PWE3-CRTL].

The control word is defined as follows:



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 frame 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 frame.

3.4.1. Setting the sequence number

For a given PW, and a pair of routers PE1 and PE2, if PE1 supports frame sequencing then the following procedures should be used:

- the initial frame transmitted on the PW MUST use sequence number 1
- subsequent frames MUST increment the sequence number by one for each frame
- when the transmit sequence number reaches the maximum 16 bit value (65535) the sequence number MUST wrap to 1

If the transmitting router PE1 does not support sequence number processing, then the sequence number field in the control word MUST be set to 0.

3.4.2. Processing the sequence number

If a router PE2 supports receive sequence number processing, then the following procedures should be used:

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

When a frame is received on that PW, the sequence number should be processed as follows:

- if the sequence number on the frame is 0, then the frame passes the sequence number check
- otherwise if the frame sequence number \geq the expected sequence number and the frame sequence number - the expected sequence number $<$ 32768, then the frame is in order.
- otherwise if the frame sequence number $<$ the expected sequence number and the expected sequence number - the frame sequence number \geq 32768, then the frame is in order.
- otherwise the frame is out of order.

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

```
expected_sequence_number := frame_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 PE2 does not support receive sequence number processing, then the sequence number field MAY be ignored.

3.5. 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 a QoS field of the encapsulating protocol (e.g., the EXP fields of the MPLS label stack or the DSCP of an IP packet). In a similar way, the egress PE MAY consider the QoS field of the PSN's encapsulating protocol when queuing the frame for CE-bound.

A PE MUST support the ability to carry the Ethernet PW as a best

effort service over the PSN. PRI bits are kept transparent between PE devices, regardless of the QoS support of the PSN.

If an 802.1Q VLAN field is added at the PE, a default PRI setting of zero MUST be supported, a configured default value is recommended, or the value may be mapped from the QoS field of the PSN, as referred to above.

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.

[3.6. Security Considerations](#)

The ethernet pseudowire type is subject to all of the general security considerations discussed in [\[PWE3-ARCH\]](#).

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.

[3.7. PSN MTU Requirements](#)

The PSN MUST be configured with an MTU that is large enough to

transport a maximum sized ethernet frame which has been encapsulated with a control word, a pseudowire demultiplexor, and a tunnel encapsulation. 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.](#) Intellectual Property Disclaimer

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

[5.](#) References

- [PWE3-CRTL] "Transport of Layer 2 Frames Over MPLS",
Martini, L., et al., [draft-ietf-pwe3-control-protocol-01.txt](#),
(work in progress), February 2003.
- [PWE3-ARCH] "Protocol Layering in PWE3"
Bryant, et al., [draft-ietf-pwe3-protocol-layer-00.txt](#)
(work in progress), November 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-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

Martini, et al.

[Page 11]

Internet Draft [draft-ietf-pwe3-ethernet-encap-02.txt](#) February 2003

Metropolitan Area Networks: Virtual Bridged Local Area Networks", 1998.

[L2TPv3] J. Lau, M. Townsley, A. Valencia, G. Zorn, I. Goyret, G. Pall, A. Rubens, B. Palter, Layer Two Tunneling Protocol (Version 3) "L2TPv3", work in progress, [draft-ietf-l2tpext-l2tp-base-03.txt](#), June 2002.

6. 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

Martini, et al.

[Page 12]

Internet Draft [draft-ietf-pwe3-ethernet-encap-02.txt](#)

February 2003

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

Martini, et al.

[Page 13]

Internet Draft [draft-ietf-pwe3-ethernet-encap-02.txt](#)

February 2003

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

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 frames 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 over an L-LSP capable network, or to multiple L2TPv3 sessions [[L2TPv3](#)].

The specific value to be assigned at the PSN for various COS is out of scope for this document.

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 COS 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

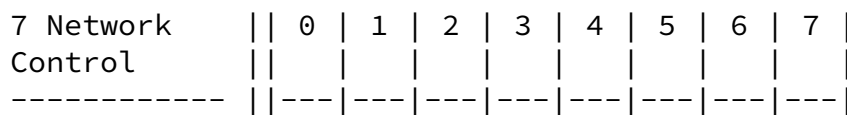


Figure 5: IEEE 802.1Q COS Service Mapping

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 frames based on the operator policy and required network behavior. This functionality is not discussed further here.

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

